Gradient Projection Anti-windup Scheme

by

Chun Sang Justin Teo

B.Eng. Electrical and Electronic Engineering
Nanyang Technological University, 1999
M.Sc. Electrical Engineering
National University of Singapore, 2003

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Doctor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2011

Author . . . .
Department of Aeronautics and Astronautics
December 29, 2010

Certified by . . . .
Jonathan Patrick How
Richard C. Maclaurin Professor of Aeronautics and Astronautics
Thesis Supervisor

Certified by . . . .
Emilio Frazzoli
Associate Professor of Aeronautics and Astronautics

Certified by . . . .
Steven Ray Hall
Professor of Aeronautics and Astronautics
MacVicar Faculty Fellow

Certified by . . . .
Eugene Lavretsky
Senior Technical Fellow, The Boeing Company

Accepted by . . . .
Eytan Modiano
Associate Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Gradient Projection Anti-windup Scheme
by
Chun Sang Justin Teo

Submitted to the Department of Aeronautics and Astronautics
on December 29, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Science

Abstract

It is a well-recognized fact that control saturation affects virtually all practical control systems. It leads to controller *windup*, which degrades/limits the system's closed-loop performance, and may cause catastrophic failures if it induces instability. Anti-windup compensation is one of two main approaches to mitigate the effects of windup, and is conceptually and practically attractive. For the idealized case of constrained linear time invariant (LTI) plants driven by LTI controllers, numerous anti-windup schemes exist. However, most practical control systems are inherently nonlinear, and anti-windup compensation for nonlinear systems remains largely an open problem.

To this end, we propose the *gradient projection anti-windup* (GPAW) scheme, which is an extension of the conditional integration method to multi-input-multi-output (MIMO) nonlinear systems, using Rosen's gradient projection method for nonlinear programming. It achieves controller state-output consistency by projecting the controller state onto the unsaturated region induced by the control saturation constraints. The GPAW-compensated controller is a hybrid controller defined by the online solution to either a combinatorial optimization subproblem, a convex quadratic program, or a projection onto a convex polyhedral cone problem. We show that the GPAW-compensated system is obtained by modifying the uncompensated system with a passive operator.

Qualitative weaknesses of some existing anti-windup results are established, which motivated a new paradigm to address the anti-windup problem. It is shown that for a constrained first order LTI plant driven by a first order LTI controller, GPAW compensation can only maintain/enlarge its region of attraction (ROA). In this new paradigm, we derived some ROA comparison and stability results for MIMO nonlinear as well as MIMO LTI systems.

The thesis is *not* that the GPAW scheme solves a centuries-old open problem of immense practical importance, but rather, that it provides a potential path to a solution. We invite the reader to join us in this quest at the confluence of nonlinear systems, hybrid systems, projected dynamical systems, differential equations with discontinuous right-hand sides, combinatorial optimization, convex analysis and optimization, and passive systems.

Thesis Supervisor: Jonathan Patrick How
Title: Richard C. Maclaurin Professor of Aeronautics and Astronautics

To my three lovely, sweet, mischievous ladies,
Linee, Sonya, Clara
and not forgetting my confidante,
Kakuni

Preface

I surmise that readers of doctoral dissertations are motivated by a desire to find a more coherent treatment of the subject matter that may have been scattered across numerous publications, and that readers of *prefaces* of doctoral dissertations are interested in how the subject matter developed or evolved. Here, I provide a candid reflection of how the *gradient projection anti-windup* (GPAW) scheme came about, almost by accident.

I started the doctoral program at MIT's Department of Aeronautics and Astronautics in the fall of 2005, working with my advisor Prof. Jonathan P. How (or Jon) ever since. Initial tinkering with parallel optimization based trajectory planning leads to nowhere. Then came the Defense Advanced Research Projects Agency (DARPA) Urban Challenge[1] where I spent a very memorable period (Jun. 2006 to Dec. 2007) with a highly energized team developing the controller and Rapidly-Exploring Random Tree (RRT) based trajectory planner[2] for MIT's entry. However, nothing resembling a doctoral dissertation that truly interests me can be made out of it. I started looking into adaptive control, which occupied me for the next year or so. Although resulting in a journal technical note[3] (that has nothing to do with adaptive control), that line of research also fell somewhat short of a doctoral dissertation. By the fall of 2008, I was in a somewhat desperate state, being without a concrete thesis topic, and with funding for my doctoral studies[4] scheduled to run out in a year.

Then, as part of the process of catching up on coursework (after the neglect owing in no small part to involvement in the DARPA Urban Challenge), I took the class "Nonlinear Control System Design" by Prof. Jean-Jacques E. Slotine in the fall of 2008. The class requires completion of a project, for which I chose to study how prior knowledge of bounds on system parameters can be incorporated in an adaptive controller (that requires no such knowledge). Intuitively, incorporating the knowledge of such bounds can improve the system's performance, which I showed in the project report. Not surprisingly, it turned out that Prof. Slotine had proved similar and more general results two decades ago. The bounds on the system parameters I studied are interval bounds, so that the bounded region containing the unknown system parameters are cuboids. Considering more general bounding conditions, e.g. bounds imposed by a set of *nonlinear* inequalities, proved elusive, at least for the class project. A class project is nice in the sense that, unlike a research paper, reporting on *negative* results, like my unsuccessful attempt to generalize the results to regions defined by nonlinear inequalities, is acceptable. After my project presentation, numerous discussions with Prof. Slotine led me to projection methods for adaptive control[5] that have been used to bound estimates of system parameters in some a priori known region. However, such projection methods are limited to projection with respect to a *single* nonlinear inequality, in contrast to a *system* of nonlinear inequalities. To summarize, we

[1] The DARPA Urban Challenge (see `http://www.darpa.mil/grandchallenge` and `http://grandchallenge.mit.edu`) is a competition where autonomous robots are required to complete a mock mission in a suburban environment while obeying some basic traffic rules.

[2] Y. Kuwata, J. Teo, G. Fiore, S. Karaman, E. Frazzoli, and J. P. How, "Real-time motion planning with applications to autonomous urban driving," *IEEE Trans. Control Syst. Technol.*, vol. 17, no. 5, pp. 1105 – 1118, Sep. 2009.

[3] J. Teo, J. P. How, and E. Lavretsky, "Proportional-Integral Controllers for Minimum-Phase Nonaffine-in-Control Systems," *IEEE Trans. Autom. Control*, vol. 55, no. 6, pp. 1477 – 1483, Jun. 2010.

[4] By my employer and sponsor, DSO National Laboratories, Singapore, for which I am deeply grateful. In May 2009, 4 months to the original schedule where my funding was to run out, DSO extended funding for doctoral studies to 5 years (for me, ending in Sep. 2010), at a very much needed and opportune time.

[5] In P. A. Ioannou and J. Sun, *Robust Adaptive Control.* Upper Saddle River, NJ: Prentice Hall, 1996. [Online]. Available: `http://www-rcf.usc.edu/~ioannou/Robust_Adaptive_Control.htm`.

can project onto cuboid type regions or regions defined by a *single* nonlinear inequality, but not regions defined by *multiple* nonlinear inequalities.

The state of affairs being thus unsatisfactory, the natural progression is to attempt to extend the projection method of Ioannou and Sun to handle multiple nonlinear inequalities. This led me to the venerable gradient projection method for nonlinear programming[6] proposed by Rosen in 1960. In analogy to Ioannou and Sun's projection method, what was required is the extension of Rosen's gradient projection method to continuous time. Then the resulting projection mechanism (less the optimization part) allows projection of the adaptive control parameter estimates onto any region defined by a *system* of nonlinear inequalities. And so I went ahead and extended Rosen's method to continuous time.[7]

At that time, my prior experience as a control engineer with DSO National Laboratories made me realize that the same projection mechanism that can be used to bound parameter estimates in adaptive control, can be used in an anti-windup setting to achieve *controller state-output consistency*, i.e. to modify the controller state such that its output inherently satisfy the saturation constraints. This is easily seen by observing that the projection mechanism allows the projection of the controller state onto any region defined by a system of nonlinear inequalities, and that controller state-output consistency simply requires the controller state to remain in some region defined by a system of saturation constraint inequalities. The obvious thing to do next is to confirm this intuition in simulation, which is easily done. Simulations on an input-constrained two-link robot driven by an adaptive sliding mode controller (made familiar in Prof. Slotine's class) yields very promising results with little to no tweaking. On reflection, all I had done was simply used an extension of Rosen's (1960) method to extend the conditional integration anti-windup scheme that has long been adopted by practitioners for proportional-integral-derivative type controllers.

At this point in time (probably around Jan. 2009), I was clueless as to the magnitude of what I had stumbled on. I was aware that anti-windup compensation is a well researched topic. But hard as I try, I could not find any literature that treats general nonlinear systems.[8] It was only a few weeks later when I chanced on a survey paper on anti-windup compensation[9] that I realized that anti-windup compensation for nonlinear systems is still considered an open problem. This got me really excited, as I have *finally* found a topic worthy of a doctoral dissertation.

The next period up to the submission date for the IEEE Conference on Decision and Control (around Mar. 2009) was spent writing up these initial results. When the conference paper[10] was submitted, reality set in. Euphoria gave way to helplessness, when I realized I had created a beast that I need to tame in order to graduate, without the slightest clue on how to go about it. I felt like Victor Frankenstein[11] after successfully giving life to his

[6] J. B. Rosen, "The gradient projection method for nonlinear programming. part I. linear constraints," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 1, pp. 181 – 217, Mar. 1960.

[7] Without bothering much about whether it actually solves the underlying optimization problem. Numerical results suggest that it does. Since my interest lies entirely on only the projection mechanism, I leave the proofs to the continuous optimization community if it manages to generate any interest.

[8] I have only found literature applicable to feedback linearizable nonlinear systems and Euler-Lagrange systems.

[9] S. Tarbouriech and M. Tuner, "Anti-windup design: an overview of some recent advances and open problems," *IET Control Theory Appl.*, vol. 3, no. 1, pp. 1 – 19, Jan. 2009.

[10] J. Teo and J. P. How, "Anti-windup compensation for nonlinear systems via gradient projection: Application to adaptive control," in *Proc. 48th IEEE Conf. Decision and Control & 28th Chinese Control Conf.*, Shanghai, China, Dec. 2009, pp. 6910 – 6916.

[11] M. W. Shelley, *Frankenstein: or, The Modern Prometheus.* London, England: Thomas Davison, 1823. [Online]. Available: `http://books.google.com/books?id=5twBAAAAQAAJ`.

monster. Just to reflect on the magnitude of the task at hand, recall that the problem of control saturation existed since the dawn of control theory, since James Watt's governor was invented. For linear time invariant (LTI) systems, rigorous stability results for anti-windup schemes were obtained only from the late 1990s onwards. We have here a centuries-old open problem of immense practical importance, compounded by the complexities of nonlinear systems! Even for the controls gurus, (I suspect) this would have been daunting, and significantly more so for a panicky struggling graduate student with 6 months left before funding runs out.

In the face of a seemingly insurmountable task, the wisdom of Jon (my advisor) is revealed when he *insisted* that I look at the simplest possible feedback system, i.e. an input-constrained first order LTI plant driven by a first order LTI controller. Naturally, I was skeptical. How much can the simplest system reveal, when perhaps all there is to know is already known, and our objective is on the opposite end of the spectrum? Contrary to my initial pessimism, it turns out that a whole lot can be revealed, and the practitioner's ad hoc conditional integration method[12] can only maintain/enlarge the system's region of attraction (ROA), and is likely an optimal anti-windup scheme in the sense that no other anti-windup scheme can achieve a larger ROA. The culmination of this exercise is a technical report[13] and a (derivative) conference paper.[14] It surprised me that practitioners had developed a (likely) optimal anti-windup scheme perhaps without knowing the full power of what they had developed! Even more astounding (to me) is the amount of insights that this simple system (that I had formerly "despised") had revealed, which led indirectly to the rest of the development presented in this dissertation. The single-minded pursuit of this generalization of a decades-old method led me into areas as diverse as hybrid systems, projected dynamical systems,[15] differential equations with discontinuous right-hand sides, combinatorial optimization, convex analysis and optimization, passivity, and more.

This is the story of how the GPAW scheme arose from a class project, in perhaps some of the most unlikely circumstances brought about by a great deal of luck and coincidences (and of course, hard work). Through these, I learned that it pays to pursue any train of thought wholeheartedly without compromise, and that no system is too simple to be unworthy of study.

Obviously, my first thanks goes to my advisor Prof. Jonathan P. How. His ability to see through a complex sequence of operations and condense it into a one-liner is amazing. From Jon, I learned to ask the *right* questions, and to pursue a train of thought even if it is controversial and politically incorrect. He showed me how to build up an argument, and most of all, taught me what research is all about.[16]

My thesis committee members, Prof. Emilio Frazzoli, Prof. Steven R. Hall, and Dr. Eugene Lavretsky, have been instrumental in providing constructive criticism and feedback,

[12] The GPAW scheme, being the extension of the conditional integration method, naturally reduces to the conditional integration method for this simple system.

[13] J. Teo and J. P. How, "Gradient projection anti-windup scheme on constrained planar LTI systems," MIT, Cambridge, MA, Tech. Rep. ACL10-01, Mar. 2010, Aerosp. Controls Lab. [Online]. Available: `http://hdl.handle.net/1721.1/52600`.

[14] J. Teo and J. P. How, "Analysis of gradient projection anti-windup scheme," in *Proc. American Control Conf.*, Baltimore, MD, Jun./Jul. 2010, pp. 5966 – 5972.

[15] Projected dynamical systems is a significant line of independent research that has attracted the attention of mathematicians, physicists, and economists, among others.

[16] It may seem obvious to many people, but on hindsight, I was not doing *independent* research until recently. There came a point in time when I realized that I was constantly formulating non-obvious conjectures and then proving or disproving them. Only *then* do I consider myself doing research.

and steering my research to completion. It has been an honor to have them on my committee. From Emilio, I learned a great deal about RRTs during the DARPA Urban Challenge phase. Collaboration with Dr. Lavretsky on *approximate dynamic inversion*[17] accustomed me to singular perturbation theory, a key element needed to extend the GPAW results to systems driven by nonlinear controllers of general structure.

I am indebted to Prof. Jean-Jacques E. Slotine for seeding the crucial initial thoughts that led to the development of the GPAW scheme. Without his leads, this dissertation would be non-existent. His lectures are always insightful, as reflected in his book, "Applied Nonlinear Control," one of the few that I have read cover-to-cover. I also thank Dr. Louis Breger and Dr. Luca Bertucelli for being my dissertation readers.

Generous financial support from my employer and sponsor DSO National Laboratories, Singapore, has enabled my doctoral studies to progress without me having to worry too much on the means to provide for my family. The freedom to choose a thesis topic to my liking is a luxury not afforded to all doctoral candidates, but enabled by the DSO Postgraduate Scholarship. For these, I am deeply grateful, especially the extension of funding for doctoral studies at the most opportune time.

Involvement in the DARPA Urban Challenge has been one of the most memorable and intense phases of my MIT experience. Many days (and nights), some filled with exhilaration, many more filled with disappointment, were spent in deserted air bases with Dr. Yoshiaki Kuwata, Gaston Fiore, Stefan Campbell, Sertac Karaman, Andrew Patrikalakis, Dr. Luke Fletcher, Prof. Edwin Olson, David Moore, and Dr. Albert Huang, working on a Land Rover that in totality costs more than a Ferrari. The principal investigators, Prof. John Leonard, Prof. Jonathan P. How, Prof. Seth Teller, and Prof. David Barrett (of Olin College) have done a wonderful job which enabled our 4th place finishing, something I will continue to boast about for quite some time.

Kathryn Fischer has been most helpful throughout my 5 years at MIT, always ensuring that Aerospace Controls Laboratory (ACL) members have what they need, and always with a smile. When she works her magic, what would have taken weeks to accomplish reduces to mere hours. Her resourceful assistance will be remembered with much gratitude.

Members of ACL past and present have contributed positively to my MIT experience. I mention Dr. Yoshiaki Kuwata, Gaston Fiore, Prof. Han-Lim Choi, Dr. Louis Breger, Dr. Luca Bertucelli, Prof. Emily Craparo, Dr. Mehdi Alighanbari, Georges Aoude, Dr. Geoffrey Huntington, and Henry Jacques Lefebvre de Plinval-Salgues, all of whom once shared an office with me in the "Laboratory for Random Graduate Students".

Last, but certainly not least, my thanks goes to my family. My parents Johny Teo and Chan Gian Hoe have always been supportive, through my rebellious teenage years to the present. Fatherhood made me aware of the sacrifices they must have made and the trials they must have gone through. My heartfelt thanks goes to them for molding me into who I am. My wife Linee Yeo sacrificed a good career with great prospects and a job she loved that is better-paid than mine, to allow me the luxury of pursuing doctoral studies at MIT. She is the caregiver, disciplinarian, comforter, entertainer, activity planner, cook, driver, art and craft teacher, story teller, and more, to our two daughters Sonya and Clara. She shouldered much of the parenting task so that I can focus on my research. For these and many more untold sacrifices, as well as the love and companionship, I am infinitely grateful.

Cambridge, MA, November 2010 Justin Teo

[17] N. Hovakimyan, E. Lavretsky, and A. Sasane, "Dynamic inversion for nonaffine-in-control systems via time-scale separation. Part I," *J. Dyn. Control Syst.*, vol. 13, no. 4, pp. 451 – 465, Oct. 2007.

Contents

List of Figures

List of Tables

List of Symbols

We will adopt the following conventions. For vectors $x = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$ and $y = [y_1, y_2, \ldots, y_n]^{\mathrm{T}}$, the vector inequality $x \leq y$ (or $x < y$, $x \geq y$, $x > y$) is to be interpreted element-wise, i.e. $x_i \leq y_i$ (respectively, $x_i < y_i$, $x_i \geq y_i$, $x_i > y_i$) for all $i \in \{1, 2, \ldots, n\}$. For a square symmetric matrix $A = A^{\mathrm{T}}$, $A > 0$ ($A \geq 0$) means A is positive definite (respectively, positive semidefinite). Other symbols that will be encountered are listed below.

$\exists$	there exists, page 56
$\forall$	for all, page 23
$x := y$	x is defined to be equal to y, page 22
$x \equiv y$	x is identically equal to y, or x is equivalent to y, page 22
$x \to y$	x approaches or tends to y, page 54
$x \downarrow y$	scalar x approaches scalar y from above, page 78
$\neg A$	NOT A, for *logical* statement A that evaluates to *true* or *false*, page 102
$A \wedge B$	A AND B, for *logical* statements A, B, page 35
$A \vee B$	A OR B, for *logical* statements A, B, page 35
$\Leftrightarrow$	logical equivalence, or if and only if, page 102
$\Rightarrow$	imply/implies, page 83
$\emptyset$	empty set, page 38
$x \in X$	x belongs to the set X, page 23
$\bar{X}$	closure of set X, page 75
∂X	boundary of set X, page 76
$X \cap Y$	intersection of sets X and Y, page 81
$X \cup Y$	union of sets X and Y, page 39
$X \subset Y$	X is a (possibly non-proper) subset of Y, page 23
$X \supset Y$	X is a (possibly non-proper) superset of Y, page 92
$X \setminus Y$	set of elements in X but not in Y, i.e. relative complement of Y in X, page 39
$\ker(X)$	kernel of set X, page 118
$(A_1, A_2, \ldots, A_i)$	collection of matrices $A_1, A_2, \ldots, A_i$, page 22
$(x_1, x_2, \ldots, x_i)$	means $[x_1^{\mathrm{T}}, x_2^{\mathrm{T}}, \ldots, x_i^{\mathrm{T}}]^{\mathrm{T}}$ or the collection of vectors $x_1, x_2, \ldots, x_i$, the meaning of which should be clear from context, page 22
0	used for scalar zero, as well as the vector/matrix of zero elements, page 22
$[A_{ij}]$	matrix with A_{ij} as its i-th row j-th column element, page 61
$\mathrm{diag}(A_1, \ldots, A_i)$	block diagonal matrix with diagonal block entries $A_1, \ldots, A_i$, page 60
$\kappa(A)$	condition number of matrix A, page 58
$\lambda(A)$	eigenvalue of square matrix A, page 55
$\langle x, y \rangle$	dot product of vectors x and y, $\langle x, y \rangle = x^{\mathrm{T}} y = y^{\mathrm{T}} x$, page 80
$\lVert x \rVert, \lVert A \rVert$	Euclidean norm of vector x or spectral norm of matrix A, page 40

$\mathbb{R}$	field of real numbers, page 23
$\mathbb{R}^n$	set of n-tuples of elements belonging to $\mathbb{R}$, page 23
$\mathbb{R}^{n\times m}$	set of $n \times m$ arrays with entries in $\mathbb{R}$, page 36
$\mathrm{rank}(A)$	rank of matrix A, page 39
$A^{-\mathrm{T}}$	transpose of inverse of matrix A, $A^{-\mathrm{T}} := (A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}$, page 45
A^{-1}	inverse of matrix A, page 39
I	identity matrix of appropriate dimensions, page 39
I_i	$i \times i$ identity matrix, page 63
$x^{\mathrm{T}}, A^{\mathrm{T}}$	transpose of vector x, or matrix A, page 23
inf	infimum. By convention, the infimum of an empty set is $+\infty$, page 56
lim	limit, page 78
max	maximum or maximize, page 23
min	minimum or minimize, page 23
$\nabla h(\tilde{x})$	$\nabla h(\tilde{x}) = \big(\frac{\partial h}{\partial x}(\tilde{x})\big)^{\mathrm{T}}$, gradient of scalar function $h(x)$ at point $\tilde{x}$, page 37
$\mathrm{sat}(\cdot)$	vector saturation function, page 21
$\mathrm{sgn}(\cdot)$	signum or sign function, page 104
sup	supremum. By convention, the supremum of an empty set is $-\infty$, page 123
$f\colon X \to Y$	f maps the domain X (e.g. $X = \mathbb{R}^n \times \mathbb{R}^m$) into the codomain Y, page 23
$x \mapsto f(x)$	x maps to $f(x)$, page 128
$\lvert\mathcal{I}\rvert$	cardinality of index set $\mathcal{I}$, page 38
$\mathcal{I}$	an index set consisting of (possibly non-consecutive) positive integers, page 38
$\mathcal{I}^*$	optimal index set to combinatorial optimization subproblem, page 41
$\mathcal{I}_i$	set of consecutive integers $\{1, 2, \ldots, i\}$ for some integer $i > 0$, page 23
$\sigma_{\mathcal{I}}$	index selection/ordering function, a bijection that assigns an index in $\mathcal{I}$ to each index in $\{1, 2, \ldots, \lvert\mathcal{I}\rvert\}$, page 38
K	unsaturated region in the controller state space, page 51
$K(y, r)$	input dependent unsaturated region in the controller state space, page 36
$R_{\mathcal{I}^*}(x)$	gradient projection operator, page 43
GPAW	gradient projection anti-windup, page 30
LMI	linear matrix inequality, page 21
LTI	linear time invariant, page 21
MIMO	multi-input-multi-output, page 28
ODE	ordinary differential equation, page 43
PID	proportional-integral-derivative, page 19
ROA	region of attraction, page 21
SISO	single-input-single-output, page 74

Theorems, lemmas, corollaries, propositions, claims, and facts, are numbered consecutively within each section. For example, within Section 4.2, the first lemma, first corollary, second lemma, and first theorem are Lemma 4.2.1, Corollary 4.2.2, Lemma 4.2.3, and Theorem 4.2.4, respectively. Definitions, assumptions, remarks, and conjectures, are each numbered independently within each chapter, while examples are numbered independently within each section. The end of proofs are indicated by the symbol ■; the end of definitions, assumptions, and remarks by □; and the end of examples by △.

Chapter 1

Introduction

It is a well-recognized fact that control saturation affects virtually all practical control systems. Examples include the minimum and maximum torque-generated/speed-attained by motors, the limiting open/close positions of control valves, the minimum and maximum cooling capacity of air-conditioners, the maximum acceleration/deceleration as well as steering limits in a car, and the deflection limits on an aircraft's control surfaces as well as its thrust limits. This fact is expounded forcefully by Bernstein and Michel in [1]:

> "All control actuation devices are subject to amplitude saturation. Force, torque, thrust, stroke, voltage, current, flow rate, and every conceivable physical input in every conceivable application of control technology is ultimately limited."

Given its prevalence, it is not surprising to find numerous books (e.g. [2–8]) and articles (e.g. articles in [9–11] and references in [1, 12–14]) devoted to the analysis of input-constrained[1] systems and their control design.

Control saturation leads to a phenomenon called *windup* [12–15]. Historically, windup is used to describe the integral state of the proportional-integral-derivative (PID) controller accumulating to anomalous levels[2] (hence appropriately called *integrator windup*) under control saturation [12], [17, Section 3.2, pp. 35 – 36]. However, it was later recognized that windup affects all dynamic controllers (even those without a pure integral state) when

[1] We will use the terms *control saturation*, *input saturation*, and *input constraint(s)* interchangeably. Note also that control saturation is one of numerous possible constraints (e.g. input rate saturation, state and output constraints) that may be imposed by practical applications.

[2] The integrator windup phenomenon is described in the 1965 patent [16], where integral action was referred to as *reset*:

> "However, difficulties are encountered when a controller having a reset capacitor is used to control a batch process, for example, a process wherein measured portions of primary chemicals are placed in a reaction vessel, and the temperature of this vessel subsequently is raised to a relatively high reaction temperature which must be maintained for a period of time. The difficulties result from the fact that, during the usually considerable time that the process condition is being brought up to set point, the deviation signal will be of substantial magnitude, and as a result an excessively large electrical charge will be built up on the reset capacitor. This result is often referred to as reset 'wind up.' After the process has reached set point, this charge will still be present, and will adversely affect the controller operation until the charge has been dissipated. Thus, especially because of the relatively long time-constants of reset circuits, considerable time may elapse before the controller is properly in control of the process."

subjected to control saturation [12, 15]. It was then interpreted as an inconsistency between the controller state and output [12, 18] arising from control saturation, among other possibilities.[3]

Windup causes performance degradation and may even induce instability [12–14]. In milder cases, it leads to sluggish closed-loop responses, large overshoots, and long settling times [17, Section 3.2, pp. 35 – 36]. Examples of disasters caused directly/indirectly by windup include the 1992 crash of the YF-22 fighter aircraft [19], the 1989 and 1993 crashes of the Saab Gripen JAS 39 fighter aircraft [20, 21], and the 1986 Chernobyl disaster [21]. All these examples, from the mild to the catastrophic, illustrate the importance of proper compensation for the deleterious effects of windup in practical control systems.

1.1 Control Design Strategies for Input-constrained Systems

Since the problem of control saturation or windup was recognized (at least as early as 1956 [14, 22]), numerous control design strategies have been proposed for input-constrained systems [2–8]. The simplest and most obvious strategy is to avoid driving the controller to saturation. For applications in which the control task is well-defined and characterized, e.g. in assembly lines where the control tasks are repetitive in nature, this can be achieved by adopting *oversized* actuators such that saturation will not occur (or the effects of windup are negligible) for the set of applicable reference trajectories. The drawback is extra costs of the oversized actuators and its associated supporting equipment (e.g. motor drivers must be sized up according to the size of the motors). In concert with the aim of avoiding saturation, numerous ad hoc schemes[4] (e.g. see the smoothing of reference input adopted in industrial robots as described in [24]) can be devised to modify the reference input to the controller. More generally, for this class of applications with specific well-defined control tasks, an *optimal control problem* [25] can be solved *offline* for reference trajectories such that their application to the closed-loop system yields responses achieving the control objective, in the presence of control saturation (and possibly other constraints).

It is clear that the preceding strategies are inadequate for many other applications in which the control tasks are not repetitive in nature and/or cannot be characterized well. An example is in fighter aircraft where the reference inputs to the flight control system (i.e. the controller) are provided by human pilots and not all possible trajectories can be explicitly characterized beforehand. For such applications, controllers must be designed to account for windup. These control design strategies generally fall under two broad classes, the *one-step* or the *two-step* approach [14], [17, p. 37].

In the *one-step* approach, controllers are designed to achieve some control objective while taking explicit account of the saturation constraints [14], [17, p. 37]. This approach encompasses a broad spectrum of methods (includes most methods in [2–8] except anti-windup methods) and is sometimes necessary when there are no other justifiable methods available to deal with the problem of windup, e.g. when the system/controller is *nonlinear* and no anti-windup method (to be discussed) exists for such systems. While this approach has its own merits, it is usually complex and often results in conservative designs with parameters that are hard to tune [14, 26, 27].

In the *two-step* approach, a nominal controller is first designed to achieve some nominal

[3] Any function (except the identity map) inserted between the output of the controller and the input to the plant produces a similar effect.

[4] In contrast to a systematic reference modification as provided by the reference governor [23].

performance, *ignoring* the effects of saturation. Then in the second step, modifications to account for saturation are incorporated, with the *design requirement* that whenever the saturation constraints are not violated, the closed-loop response must be governed by the nominal controller *only*, i.e. the compensated and uncompensated system responses must be identical. Whenever the nominal controller saturates, the control modifications attempt to minimize the effects of windup. Such an approach is called *anti-windup compensation* [12–14], which is the subject of this dissertation.

Anti-windup compensation is attractive because [14]:

- it provides a decoupling between nominal performance and constraint handling which greatly simplifies the design of the nominal controller;
- it can be *retrofitted* to existing controllers, which allows incremental "system upgrades" using applicable anti-windup schemes available up to the point in time.

For these reasons, anti-windup compensation is often the preferred choice among practitioners [14], in particular when the nominal controller is simple, e.g. PID controllers [17, p. 37].

In addition to the preceding, we note that *given* some anti-windup scheme, a corresponding one-step approach may be devised. To illustrate this point, suppose some anti-windup scheme has been developed, that comes with some sufficient conditions that, when satisfied by the uncompensated system, will yield an *anti-windup compensator* such that its application will yield some desired stability and performance properties. If these *anti-windup* sufficient conditions can be incorporated in the design of the *nominal controller*, then a unified *one-step* control design is achieved. Such an approach may yield better overall designs, since the nominal controller is not fixed a priori (which would have translated to a hard constraint on the achievable stability/performance due to anti-windup compensation), and nominal performance and constraint handling may be traded off systematically. Such an approach is adopted in [27–32].

1.2 Anti-windup Compensation

Given the practical appeal of anti-windup compensation and the firm foundation of linear systems theory [33], much work has been devoted to the development of anti-windup schemes for input-constrained linear time invariant (LTI) plants driven by LTI controllers. Surveys on this topic are [12–14]. The most recent of these surveys [14] provides a historical sketch of the development of anti-windup schemes as well as descriptions of modern anti-windup schemes that provide guarantees of global asymptotic stability, or estimates of the associated region of attraction (ROA) when only local asymptotic stability can be assured (which is the case when the open-loop plant is unstable [34]).

Typically, a structure for the anti-windup compensator is first *assumed*, which implicitly defines the number and structure of the *parameters* of the anti-windup compensator. Then the parameters (or gains) that define the anti-windup compensator are determined, usually by solving some optimization problem (typically a semidefinite program [35] or linear matrix inequality (LMI) problem [36, Section 2.2.1, p. 9]) offline. For an input-constrained LTI plant (with state x, input u, and measurement y) described by

$$\begin{aligned}\dot{x} &= Ax + B\,\mathrm{sat}(u),\\ y &= Cx + D\,\mathrm{sat}(u),\end{aligned}$$

where sat$(\cdot)$ denotes the saturation function, the *anti-windup compensated controller* is defined by two subsystems, typically of the form [14]

$$\tilde{\Sigma}_c\colon \begin{cases} \dot{x}_c = A_c x_c + B_c y + B_{cr} r + y_{aw1}, \\ \tilde{u} = C_c x_c + D_c y + D_{cr} r, \end{cases} \qquad \tilde{\Sigma}_{aw}\colon \begin{cases} \dot{x}_{aw} = A_{aw} x_{aw} + B_{aw} w, \\ y_{aw1} = C_{aw1} x_{aw} + D_{aw1} w, \\ y_{aw2} = C_{aw2} x_{aw} + D_{aw2} w, \end{cases}$$

where the controller output u and *anti-windup compensator* input w are

$$u = \tilde{u} + y_{aw2}, \qquad w = \text{sat}(u) - u.$$

The structure of the closed-loop anti-windup compensated system is shown in Fig. 1-1.

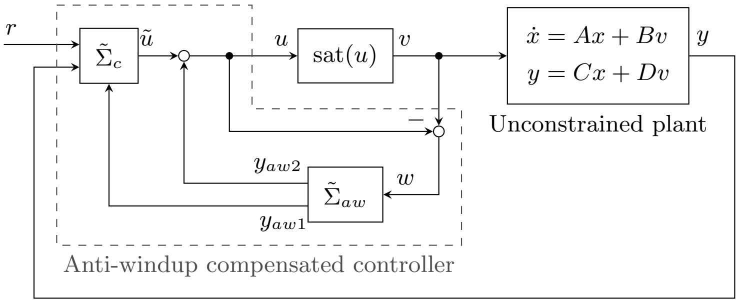


Figure 1-1: Structure of typical anti-windup compensated system.

Here, (A, B, C, D) define the plant, $(A_c, B_c, B_{cr}, C_c, D_c, D_{cr})$ define the *nominal controller*, and $(A_{aw}, B_{aw}, C_{aw1}, C_{aw2}, D_{aw1}, D_{aw2})$ define the *anti-windup compensator*, all of which are constant real matrices of appropriate dimensions. Subsystem $\tilde{\Sigma}_c$ *reduces* to the nominal controller (with state x_c, control output $u := \tilde{u}$, measurement input y, and exogenous input r) when $y_{aw1} \equiv 0$. It should be emphasized that we distinguish between the *anti-windup compensator* $(\tilde{\Sigma}_{aw})$ and the *anti-windup compensated controller*, which comprises of both $\tilde{\Sigma}_c$ and $\tilde{\Sigma}_{aw}$ (see Fig. 1-1). To be clear, the *anti-windup compensated controller* has state (x_c, x_{aw}), control output u, measurement input y, exogenous input r, and (y_{aw1}, y_{aw2}) are *internal* signals. Except for the anti-windup parameters $(A_{aw}, B_{aw}, C_{aw1}, C_{aw2}, D_{aw1}, D_{aw2})$, all other matrices are predefined by the plant and nominal controller. These unknowns are then determined (typically) by solving an optimization problem [14]. Observe that the typical assumed structure of the anti-windup compensator imposes 6 matrix parameters, $(A_{aw}, B_{aw}, C_{aw1}, C_{aw2}, D_{aw1}, D_{aw2})$.

1.2.1 Anti-windup Compensation for Nonlinear Systems – Open Problem

As discussed in [12–14], numerous anti-windup schemes are available when the unconstrained plant and controller are LTI. However, most practical control systems are inherently *nonlinear*. Excluding saturation, examples of common nonlinearities at the subsystem/actuator level are dead zone, hysteresis, and backlash [37, Section 1.2.7, pp. 18 – 23], [38, Section 5.2, pp. 169 – 172]. Classical examples of nonlinear systems are the tunnel diode circuit, the van der Pol oscillator, the pendulum, the buckling beam, and the Volterra-Lotka predator-prey model [39, Section 1.3, pp. 14 – 22]. Nonlinear systems encountered on a day-to-day basis include general mechanical systems governed by Euler-Lagrange equations [40, pp. 16 – 19], [41, Section 4.1, pp. 63 – 72], the nonlinear dynamics of the aircraft [42, Table 3.1, p. 105], the kinematic model of a simple car [43, Section 13.1.2.1, pp. 596 – 599], and the kinematic model of a car pulling trailers [43, Section 13.1.2.4, pp. 602 – 603]. Even when the unconstrained plant is LTI, adoption of any of the numerous nonlinear controllers in [37–41, 44–46] renders the nominal closed-loop system nonlinear.

Whether LTI models can approximate a system sufficiently well depends on the severity of the nonlinearities and the application. For high performance systems that need to operate over a large region of the state space, it is unlikely that LTI models alone can capture the dominant system behaviors, due to the richness of nonlinear phenomena [39, Section 1.1, pp. 1 – 4]. Moreover, adoption of any nonlinear controller likely imply that LTI controllers are inadequate for the control task. Control design for nonlinear systems is generally complex, and can be significantly simplified if the complexities due to saturation can be ignored at the outset. Clearly, these motivate the need for anti-windup schemes applicable to nonlinear systems/controllers. However, anti-windup compensation for nonlinear systems/controllers remains largely an *open problem* [14]. To set the stage for discussion on available anti-windup methods for nonlinear systems, the problem statement is presented next.

1.3 Problem Statement

For some integer $i > 0$, let $\mathcal{I}_i := \{1, 2, \dots, i\}$. Consider the input-constrained nonlinear plant

$$\Sigma_p\colon \begin{cases} \dot{x} = f(x, \mathrm{sat}(u)), \qquad x(0) = x_0, \\ y = g(x, \mathrm{sat}(u)), \end{cases} \tag{1.1}$$

where $x, x_0 \in \mathbb{R}^n$ are the state and initial state, $u = [u_1, u_2, \dots, u_m]^{\mathrm{T}} \in \mathbb{R}^m$ is the control input, $y \in \mathbb{R}^p$ is the measurement, $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ are functions governing the state evolution and output measurement respectively, and the vector saturation function $\mathrm{sat}\colon \mathbb{R}^m \to \mathbb{R}^m$ is defined by

$$\begin{gathered} \mathrm{sat}(u) = [\rho_1(u_1), \rho_2(u_2), \dots, \rho_m(u_m)]^{\mathrm{T}}, \\ \rho_i(u_i) = \max\{\min\{u_i, u_{\max,i}\}, u_{\min,i}\}, \qquad \forall i \in \mathcal{I}_m, \end{gathered} \tag{1.2}$$

for some $u_{\max,i}, u_{\min,i} \in \mathbb{R}$ satisfying $u_{\min,i} < u_{\max,i}$ for all $i \in \mathcal{I}_m$. Clearly, the *unconstrained plant* is described by (1.1) and (1.2) with $u_{\max,i} = -u_{\min,i} = \infty$. Define also the vectors of saturation limits $u_{\max} := [u_{\max,1}, u_{\max,2}, \dots, u_{\max,m}]^{\mathrm{T}}$ and $u_{\min} := [u_{\min,1}, u_{\min,2}, \dots, u_{\min,m}]^{\mathrm{T}}$ for ease of reference.

Remark 1.1. While it is customary to consider symmetric saturation constraints (e.g. see [14, 24]), i.e. $u_{\min,i} = -u_{\max,i}$ for all $i \in \mathcal{I}_m$, results thus obtained may not carry over to the case of *asymmetric* saturation constraints [47], one of practical importance. Numerical results in Section 3.6 show that the ROAs of the same system under symmetric and asymmetric saturation constraints can be qualitatively different, which suggests that asymmetric saturation constraints should be considered for full generality. □

Let $\mathcal{R}$ be a class of admissible reference signals evolving in $\mathbb{R}^{n_r}$, e.g. $\mathcal{R} \subset C^k([0, \infty), \mathbb{R}^{n_r})$, where $C^k([0, \infty), \mathbb{R}^{n_r})$ is the vector space of k times continuously differentiable functions

$[0, \infty) \to \mathbb{R}^{n_r}$. For any reference input $r \in \mathcal{R}$, let the *nominal controller* be described by

$$\Sigma_c\colon \begin{cases} \dot{x}_c = f_c(x_c, y, r), & x_c(0) = x_{c0}, \\ u_c = g_c(x_c, y, r), \end{cases} \tag{1.3}$$

where $x_c, x_{c0} \in \mathbb{R}^q$ are the state and initial state, $u_c \in \mathbb{R}^m$ is the controller output, $y \in \mathbb{R}^p$ is the measurement input, and $f_c\colon \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^{n_r} \to \mathbb{R}^q$, $g_c\colon \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^{n_r} \to \mathbb{R}^m$ govern the controller state evolution and output. Assume that the nominal controller (1.3) has been designed such that the *uncompensated closed-loop system* (also called the *uncompensated system* or the *nominal system*) defined by (1.1), (1.2), (1.3), and the relation $u \equiv u_c$ is well-posed, and achieves some nominal stability and performance.

Remark 1.2. Clearly, $r(t) \in \mathbb{R}^{n_r}$ means the *instantaneous* value of the *function* $r \in \mathcal{R}$, $r\colon [0, \infty) \to \mathbb{R}^{n_r}$, at time t. By an abuse of notation, we will also use r $(:= r(t))$ to denote the same instantaneous value of the function $r \in \mathcal{R}$, the meaning of which should be clear from context. Observe that r in (1.3) refers to the instantaneous function value. □

Remark 1.3. While we recognize the importance of robustness issues associated with the presence of noise, disturbances, time delays, and unmodeled dynamics in any control system, it appears superfluous at this point to consider these, given that the topic is still largely an open problem [14]. We leave the study of these (important) issues as future work (see Section 7.1.1). □

Problem 1 (Anti-windup Compensation for Nonlinear Systems)**.** The goal is to design an *anti-windup compensated controller* (see Section 1.2)

$$\Sigma_{aw}\colon \begin{cases} \dot{x}_{aw} = f_{aw}(x_{aw}, y, r), & x_{aw}(0) = x_{aw0}, \\ u_{aw} = g_{aw}(x_{aw}, y, r), \end{cases} \tag{1.4}$$

with state $x_{aw} \in \mathbb{R}^{\tilde{q}}$, inputs (y, r) and m dimensional output $u_{aw} \in \mathbb{R}^m$, and determine a state initialization x_{aw0}, such that the *anti-windup compensated system* defined by (1.1), (1.2), (1.4), and the relation $u \equiv u_{aw}$ satisfy:

(i) when $(r, x_0, x_{c0}) \in \mathcal{R} \times \mathbb{R}^n \times \mathbb{R}^q$ are such that the uncompensated system remains stable, then the anti-windup compensated system must also remain stable for $(r, x_0, x_{aw0}) \in \mathcal{R} \times \mathbb{R}^n \times \mathbb{R}^q$;

(ii) for every (r, x_0, x_{c0}) such that the controls never saturate for the uncompensated system, i.e. $\mathrm{sat}(u_c) \equiv u_c$, the control signal of the anti-windup compensated system for (r, x_0, x_{aw0}) satisfies $u_{aw} \equiv u_c$;

(iii) when (r, x_0, x_{c0}) are such that some controls saturate for the uncompensated system (i.e. there exists a non-trivial interval $[t_1, t_2]$, $t_1 < t_2$, such that $\mathrm{sat}(u_c(t)) \neq u_c(t)$ for all $t \in [t_1, t_2]$), then the performance of the anti-windup compensated system for (r, x_0, x_{aw0}) must be no worse than the uncompensated system. □

Remark 1.4. Examination of the preceding conditions shows that there always exists a trivial solution ($\Sigma_{aw} \equiv \Sigma_c$, $x_{aw0} = x_{c0}$) to Problem 1. To exclude the trivial solution, we can always add a 4th condition:

(iv) there exists $(r, x_0, x_{c0}) \in \mathcal{R} \times \mathbb{R}^n \times \mathbb{R}^q$ such that the performance of the anti-windup compensated system for $(r, x_0, x_{aw0}) \in \mathcal{R} \times \mathbb{R}^n \times \mathbb{R}^q$ is strictly better than the uncompensated system.

We exclude it because it has no real practical significance. □

Remark 1.5. Together, conditions (ii) and (iii) means that performance is never reduced with anti-windup compensation. Hence an equivalent formulation of condition (iii) is:

(iii) for all $(r, x_0, x_{c0}) \in \mathcal{R} \times \mathbb{R}^n \times \mathbb{R}^q$, the performance of the anti-windup compensated system for $(r, x_0, x_{aw0}) \in \mathcal{R} \times \mathbb{R}^n \times \mathbb{R}^q$ must be no worse than the uncompensated system. □

Remark 1.6. Numerous performance metrics can be specified, depending on the application. A common performance metric is based on L_2-gain [14, 48, 49], [41, pp. 18 – 19]. Also, observe that numerous stability criteria can be specified (in particular, see [46, pp. 207 – 212]), the suitability of which is again application dependent. □

Some observations in the statement of Problem 1:

- Condition (i) states that stability must never be compromised by anti-windup compensation. For a regulatory system, i.e. $\mathcal{R}$ is a set of constant functions $\mathcal{R} \subset \{\bar{r}\colon [0,\infty) \to \mathbb{R}^{n_r} \mid \dot{\bar{r}} \equiv 0\}$, this condition means that the ROA of the system can only be maintained/enlarged by anti-windup compensation.
- Condition (ii) states that nominal performance is recovered whenever no controls saturate, a classical requirement. When the nominal controller is a *minimal realization* [50], condition (ii) implies that in general, the order of the anti-windup compensated controller Σ_{aw} must be at least as large as that of the nominal controller Σ_c, i.e. $\tilde{q} \geq q$ (see (1.3) and (1.4)).
- The requirement for determination of the controller state initialization x_{aw0} is due to the recognition that usually, the controller state can be arbitrarily initialized. Such an initialization can be advantageous even if no explicit anti-windup scheme is adopted [51]. Moreover, condition (ii) imply that x_{aw0} must be dependent on x_{c0}.

1.3.1 Recovering the Anti-windup Compensator

In contrast to the typical anti-windup framework [14], the *structure* of the anti-windup compensated controller is not assumed in Problem 1. This is best seen by comparing Fig. 1-2 with Fig. 1-1, where the *anti-windup compensated controller* Σ_{aw} is free to assume any structure. While this provides much flexibility for the design of the anti-windup compensated controller, it may appear that one of the principal advantages of anti-windup schemes, namely their amenability to be *retrofitted* to an existing controller, is lost. We show here that an *anti-windup compensator* can always be obtained from the *anti-windup compensated controller* with either measurement of the nominal controller output, or knowledge of the nominal controller realization and initial state. Observe that these are reasonable requirements even to realize the conventional anti-windup compensator $\tilde{\Sigma}_{aw}$ in Fig. 1-1.

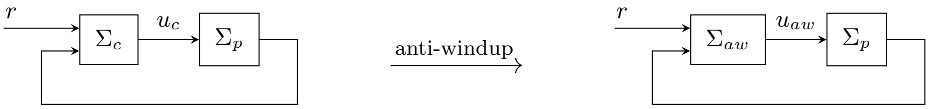


Figure 1-2: Illustration of general anti-windup problem.

Assume that an *anti-windup compensated controller* Σ_{aw} has been designed and is described by (1.4). When the output of the nominal controller u_c in (1.3) can be measured, an additive *anti-windup compensator* with output $\tilde{u}_{aw}$ can be defined by

$$\tilde{\Sigma}_{aw}\colon \begin{cases} \dot{x}_{aw} = f_{aw}(x_{aw}, y, r), & x_{aw}(0) = x_{aw0}, \\ \tilde{u}_{aw} = g_{aw}(x_{aw}, y, r) - u_c, \end{cases} \tag{1.5}$$

so that when the nominal controller's output u_c is added to the derived anti-windup compensator's output $\tilde{u}_{aw}$, we get the desired control signal $u = u_c + \tilde{u}_{aw} = g_{aw}(x_{aw}, y, r) = u_{aw}$.

Alternatively, if the realization of the nominal controller (1.3) and initial state x_{c0} are known, we can incorporate a model of the nominal controller and define the additive *anti-windup compensator* by

$$\tilde{\Sigma}_{aw}\colon \begin{cases} \dot{\tilde{x}}_c = f_c(\tilde{x}_c, y, r), & \tilde{x}_c(0) = x_{c0}, \\ \dot{x}_{aw} = f_{aw}(x_{aw}, y, r), & x_{aw}(0) = x_{aw0}, \\ \tilde{u}_{aw} = g_{aw}(x_{aw}, y, r) - g_c(\tilde{x}_c, y, r). \end{cases} \tag{1.6}$$

By the preceding construction, we have $\tilde{x}_c \equiv x_c$ (compare with (1.3)), so that similar addition yields

$$\begin{aligned} u = u_c + \tilde{u}_{aw} &= u_c + g_{aw}(x_{aw}, y, r) - g_c(\tilde{x}_c, y, r) = u_c + g_{aw}(x_{aw}, y, r) - g_c(x_c, y, r), \\ &= u_c + g_{aw}(x_{aw}, y, r) - u_c = g_{aw}(x_{aw}, y, r) = u_{aw}, \end{aligned}$$

as desired. Fig. 1-3 illustrates the resulting closed-loop system obtained from the derived additive anti-windup compensator $\tilde{\Sigma}_{aw}$. These are two of numerous possible realizations of anti-windup compensators that can be obtained from the *anti-windup compensated controller*. Each of these *realizations* comes with its own robustness issues, the study of which we leave as future work (see Section 7.1.2).

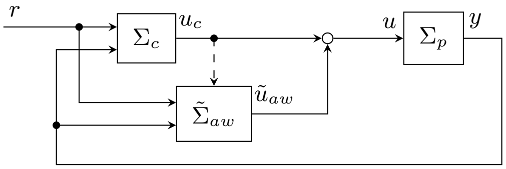


Figure 1-3: Closed-loop system with derived *anti-windup compensator*. The dashed line represents the case when measurement of u_c is available.

1.4 Literature Review

With Problem 1 as objective, we will focus only on literature relevant to anti-windup compensation for *nonlinear* systems, leaving out the vast majority of work on input-constrained LTI plants driven by LTI controllers [12–14].

1.4.1 Conditioning Technique

One of the early anti-windup schemes is the *conditioning technique* [51, 52], which is an extension of the *back-calculation method* reported in [53] (see also [18]). It is important because it forms the conceptual basis for numerous of the methods reviewed in this section, and it claims to be applicable to controllers of general structure [51]. When some controls saturate, the key idea is to determine a "realizable reference" input r_r such that the saturation constraints will be satisfied when the true reference[5] r is replaced by r_r. Conceptually, this newly determined realizable reference is then used, until the point where the true reference r causes no saturation constraint violations. The extension to nonlinear systems is presented in [51], where it was claimed:

> "It can be applied to any kind of controller (linear or not, discrete or continuous, time varying or not, single-input-single-output or multiple-input-multiple-output)."

While technically correct, the claim may be misleading. The first fundamental difficulty is that for general controllers of the form (1.3), a system of implicit nonlinear equations of the form $g_c(x_c, y, r_r) = \text{sat}(u_c)$ needs to be solved for r_r *online*. Not only would this be computationally prohibitive, neither existence nor uniqueness of solutions can be guaranteed in general [54, Section 2.1, pp. 15 – 16]. Moreover, when the nominal controller (1.3) is "strictly proper", i.e. the output equation has the form $u_c = g_c(x_c)$ and does not depend on r, the conditioning technique breaks down, in the sense that effectively, it yields *no anti-windup compensation*! This makes the conditioning technique as proposed in [51] ill-suited[6] as a candidate method to solve Problem 1.

1.4.2 Feedback Linearizable Nonlinear Systems

The majority of the literature on anti-windup compensation for nonlinear systems applies to feedback linearizable nonlinear systems [56–66]. Feedback linearization or dynamic inversion [44, Section 4.2, pp. 147 – 161] transforms the input-constrained nonlinear system into an equivalent input-constrained LTI system (with state-dependent constraints [56]) for which a nominal controller is designed. Then well established anti-windup schemes developed for input-constrained LTI systems are adopted or modified for anti-windup compensation. Note that this is a brief *oversimplified* description, ignoring the complications that arose and were addressed within these papers. These methods are appealing due to their conceptual simplicity (though by no means trivial), and relates well to established techniques, so insights from the experience on input-constrained LTI systems could be harnessed to some degree. While these methods are valuable in advancing the state-of-the-art, they are not viable candidates for *general* nonlinear plants and/or controllers since nonlinear systems are not generically feedback linearizable [67, 68], and the nominal controller is restricted to be a feedback linearizing controller.

1.4.3 Anti-windup Schemes for Particular Controllers

Various anti-windup schemes were proposed in [69–72] for some *particular* adaptive controllers, and in [73] for a sliding mode controller [38, Chapter 7, pp. 276 – 306]. These methods depend strongly on the underlying structure of the controller, and would not be applicable to general controllers of the form (1.3).

[5] For consistency in the discussions, we will use the notation established in Section 1.3 whenever possible, instead of notations used in particular papers.

[6] Note that the restrictions imposed on the nominal controller to avoid some of these limitations are severe (with respect to the nonlinear anti-windup problem), in requiring (i) the output function g_c (see (1.3)) to be *linear* in r; (ii) the dimensions of reference input and *control output* to coincide, i.e. $n_r = m$ (see the definition of r and u_c in (1.3)). See also [55] for a criticism of the conditioning technique when applied to input-constrained LTI systems.

1.4.4 Nonlinear Anti-windup for Euler-Lagrange Systems

An anti-windup scheme called *nonlinear anti-windup* was proposed in [74] and extended for Euler-Lagrange systems in [24]. Some remarkable properties of the method are:

- global asymptotic stability is assured for nominal controllers of general structure, i.e. of the form (1.3), as long as it globally stabilizes the *unconstrained system*;
- the realization of the anti-windup compensator depends only on the plant, and is independent of the nominal controller.

Due to the importance of Euler-Lagrange systems in modeling a large class of physical systems (see [40]), the *nonlinear anti-windup* scheme [24] is a very promising candidate to solve Problem 1. One difficulty in extending the method is the construction of the nonlinear function that defines the anti-windup compensator. While sufficient conditions and guidelines have been provided in [24] for Euler-Lagrange systems, it is likely a difficult task for general plants. Another minor point is that the state of the anti-windup compensator is required to be initialized to zero for nominal performance to be recoverable in the absence of saturation. While this removes one degree of freedom in controller state initialization (as stated in Problem 1, the determination of x_{aw0}) that may be exploited to improve the system's performance, it is likely that some initialization scheme can be devised appropriately.

1.4.5 Optimal Directionality Compensation

The *optimal directionality compensation* method was proposed in [75] for a class of *multi-input-multi-output* (MIMO) nonlinear systems, and extended in various ways in [76–81]. In *multi-input* systems, the vector input has a *direction* and magnitude, the *direction* being nonexistent in *single-input* systems. Recognizing that simply saturating (or clipping) the (vector) plant input or preserving the plant input direction [15] may not yield satisfactory results, performance degradation due to directionality in *multi-input* systems and windup are specifically distinguished. The key aspect is the design of an *optimal directionality compensator*. At each time instant, given the controller output u_c, the optimal directionality compensator finds a feasible control u (satisfying $u = \mathrm{sat}(u)$) that minimizes the difference between the system output (predicted at a short horizon into the future) due to u_c (in the absence of saturation), and the system output (predicted at the same short horizon into the future) due to u (in the presence of saturation). The control signal u is obtained as the solution to a constrained convex quadratic program [6, Section 2.5.6, p. 54] at each time instant, and applied to the plant. Note that the resulting control is generally different from that obtained by *direction preservation* [15]. Also, observe the similarities with *model predictive control* (MPC) [6, 82], where an optimization problem is solved online.

An important point is that the optimal directionality compensator does not modify the state or input of the controller directly, and essentially leaves them unaltered. Any changes in dynamic response of the controller and closed-loop system due to the directionality compensator is effected by changes in the plant response due to the newly determined control input. To enhance the system's performance, various (well-known) anti-windup schemes were added to the nominal controller, with the optimal directionality compensator remaining a distinct compensator.

One of the fundamental limitations in the method is that the system has to be square, i.e. has the same number of control inputs as measurement outputs. If this limitation can be eliminated and the method extended for general systems and controllers, this method can be a viable candidate to solve Problem 1.

1.4.6 Reference Governor

Distinct from conventional anti-windup schemes [14], the *reference governor* is a significant line of research, for which representative literature are [23, 83–99]. Among these, those that are applicable to nonlinear systems are [23, 95–99]. It is a *two-step* approach for constrained control that modifies the reference input (r in (1.3)) to the closed-loop system comprising a general nonlinear plant driven by a predesigned nominal controller of general structure. As such, it has conceptual similarities with the conditioning technique [51], the fundamental difference being the manner in which the true reference input is modified. For its implementation, it requires the *online* solution to an optimization problem, which renders it similar in some respects to MPC [82]. However, it solves a significantly simpler online optimization problem, and hence can be thought of as a simplified form of MPC.

Observe that control saturation constraints in (1.3) can be translated into constraints on the states and reference input of the closed-loop system. Under some assumptions, the reference governor can ensure satisfaction of constraints on closed-loop states and reference input [23, 95], so that it is more than adequate for anti-windup compensation.

Some difficulties of the reference governor are:

- some degree of conservatism were introduced to enable satisfaction of the hard constraints;
- for nonlinear systems/controllers, in general, a non-convex constrained nonlinear program needs to be solved *online*.

Despite these difficulties, it appears that the reference governor either solves Problem 1, or is a very likely candidate. The part that (at present) is ambiguous, is whether condition (i) of Problem 1 is fulfilled by the reference governor. It appears that due to some conservatism introduced, the reference governor must operate *within* the stability region of the *uncompensated* system. If its application will never *reduce* the stability region, it is very likely that it solves Problem 1 (with some qualifications). Otherwise, it is still a very likely candidate if this limitation can be resolved. We leave the investigation of these as future work (see Section 7.1.3).

1.4.7 Summary of Literature Review

Three likely candidate methods to solve the general anti-windup problem (Problem 1) were identified, namely, the nonlinear anti-windup scheme in [24], the optimal directionality compensation method in [79], and the reference governor in [23, 95]. Their primary characteristics were briefly described, and their current limitations have been identified. Structurally, anti-windup compensation is achieved by modification of:

- both the *measurement* input (y in (1.3)) and output of the nominal controller in the nonlinear anti-windup scheme [24];
- plant input (or equivalently, nominal controller output) in the optimal directionality compensator [79];
- *reference* input (r in (1.3)) in the reference governor [23, 95].

At present, it appears that the most likely candidate to solve Problem 1 is the reference governor, although more work is required to confirm this. The subject of this dissertation is the *gradient projection anti-windup* scheme. Similarities/relations with these existing methods are discussed in Section 2.9.

1.5 Dissertation Overview

The remainder of this dissertation is briefly described below.

Chapter 2 – Construction and Fundamental Properties

As mentioned previously, the subject of this dissertation is the *gradient projection anti-windup* (GPAW) scheme, constructed for general saturated nonlinear plants driven by nonlinear controllers. In Chapter 2, we construct the GPAW-compensated controller and show that it is a generalization of the well-known *conditional integration* method for PID-type controllers [53], [17, Section 3.3.2, p. 38]. This generalization depends critically on the projection operator that in turn is obtained by extension of Rosen's gradient projection method for nonlinear programming [100, 101] to continuous time. Fundamental properties like the *controller state-output consistency* property of GPAW-compensated controllers, as well as passivity and L_2-gain of the projection operator are established. The pertinent features of the GPAW scheme are demonstrated on a two-link robot driven by an adaptive sliding mode controller [102].

Chapter 3 – Input Constrained Planar LTI Systems

In Chapter 3, we study the simplest possible saturated feedback system, namely an input-constrained first order LTI plant driven by a first order LTI controller, where the objective is to regulate the system state about the origin. For this simple system, strong results on GPAW compensation are obtained, the most notable of which is that GPAW compensation (which reduces to the conditional integration method) can only maintain/enlarge the ROA of the uncompensated system. Qualitative weaknesses of some results in existing anti-windup literature are illustrated, which motivates a new paradigm to address the anti-windup problem, reflected in the statement of Problem 1 in Section 1.3. Numerical results in this chapter also imply that *asymmetric* saturation constraints should be considered for general input-constrained systems.

Chapter 4 – Geometric Properties and Region of Attraction Comparison Results

The main themes of Chapter 4 are geometric properties of GPAW-compensated systems and some ROA comparison results. Despite being primarily defined by a combinatorial optimization subproblem, we show that the GPAW-compensated controller can be equivalently defined by the online solution to either a convex quadratic program or a projection onto a convex polyhedral cone problem. This is significant because it allows a computationally attractive realization, and holds regardless of the nonlinearities in the plant or nominal controller. A geometric bounding condition that relates the vector fields of the nominal and GPAW-compensated controller is presented. Motivated by results in Chapter 3, ROA comparison results consistent with the new anti-windup paradigm are presented. The chapter ends with demonstrations of these ROA comparison results on some simple systems.

Chapter 5 – Input Constrained MIMO LTI Systems

Clearly, LTI models are widely used to approximate practical control systems. Saturated MIMO LTI plants driven by MIMO LTI controllers are studied in Chapter 5. An ROA comparison result of Chapter 4 is specialized to yield a stability result, which can be verified by solving an LMI problem. We show how the familiar similarity transformation can be applied to GPAW-compensated controllers, and that under some choice of the GPAW parameter, the GPAW-compensated system can be transformed into a *linear system with partial state constraints*, which has been studied in [103–106]. This link allows existing stability results to be applied to this class of GPAW-compensated systems, and vice versa.

Chapter 6 – Numerical Comparisons

In Chapter 6, we compare the GPAW scheme against two anti-windup schemes for nonlinear systems, and an anti-windup scheme for LTI systems, using some non-trivial examples available in the literature. Thus far, the stability results obtained in Chapters 4 and 5 are too conservative to be applied to these examples. Nevertheless, we show that even in the absence of stability results, ad hoc methods can be devised to determine the GPAW parameter such that the GPAW-compensated system achieves qualitatively similar performance as these state-of-the-art anti-windup schemes. These show that even in the absence of stability proofs, the GPAW scheme can be a viable *candidate* anti-windup method when no other anti-windup schemes are suitable.

Chapter 7 – Conclusions and Future Work

In Chapter 7, we summarize the main results and present possible areas for future research. This dissertation represents the first steps in the study of GPAW compensation, which has much potential to be developed into a formidable tool to address the problems caused by control saturation. Some possible links with other fields of controls and dynamical systems are identified, which may provide fruitful research on a theoretical as well as practical level.

Appendix A – Closed Form Expressions for Single-output GPAW-Compensated Controllers

Appendix A presents closed-form expressions for single-output GPAW-compensated controllers, which includes the case when the nominal controller is of PID-type. These closed-form expressions allow a computationally efficient implementation of the GPAW-compensated controller.

Appendix B – Closed Form Expressions for GPAW-Compensated Controllers with Output of Dimension Two

Appendix B presents closed-form expressions for GPAW-compensated controllers with output of dimension two. Computational results show that using these closed-form expressions requires on average less than 10% of computation power compared to the best case quadratic program realization of GPAW-compensated controllers.

Appendix C – Procedure to Apply GPAW Compensation

There are three equivalent ways to realize general GPAW-compensated controllers, detailed in different sections of this dissertation. For ease of reference, we summarize the procedure to apply GPAW compensation in Appendix C.

1.6 Contributions

Contributions of the research presented herein are listed below.

- **Developed General Purpose Anti-windup Scheme.** The GPAW scheme is a general purpose anti-windup scheme constructed for saturated nonlinear systems driven by output-feedback nonlinear controllers, and can be easily specialized as appropriate. It has clear geometric properties and is characterized by a passive projection operator with L_2-gain less than one. It achieves *controller state-output consistency*, a property unique to GPAW-compensated controllers, while being an implicit objective for most anti-windup schemes.

- **Motivated a New Paradigm for Anti-windup Problems.** We demonstrated the qualitative weaknesses of some existing results in the anti-windup literature, which motivated a new paradigm to address the anti-windup problem. This new paradigm is to search for results *relative* to the uncompensated system, such that it will give clear indications of benefits gained by adopting the anti-windup scheme.

- **Demonstrated Need to Consider Asymmetric Saturation Constraints for General Saturated Systems.** Numerical results showed that even for a simple saturated system, the ROAs induced by symmetric and *asymmetric* saturation constraints are fundamentally different. ROAs induced by symmetric saturation constraints are invalid when the saturation constraints become asymmetric, even when they are relaxed. This applies to general saturated systems and shows the need to consider the asymmetric case. We note that the majority of literature on saturated systems considers only the symmetric saturation case, and GPAW compensation applies whether the saturation constraints are symmetric or otherwise.

- **Developed ROA Comparison and Stability Results for GPAW-Compensated Systems.** Consistent with the new paradigm to search for results relative to the uncompensated system, the ROA comparison results are the first steps in this direction. These results are applicable to fairly general nonlinear systems and controllers, although they are likely conservative.

- **Demonstrated Viability of GPAW Scheme as a Candidate Anti-windup Scheme for General Systems.** The GPAW scheme was constructed for general nonlinear systems and controllers. Even when current stability results are too conservative to be applicable, we show that ad hoc methods can be devised to design the GPAW-compensated controller to yield (numerically) satisfactory solutions. This will appeal to practitioners in need of a candidate anti-windup scheme when otherwise, no suitable candidates exist.

- **Related GPAW-Compensated Systems to Projected Dynamical Systems and Linear Systems with Partial State Constraints.** For the constrained planar LTI system, we showed that it is in fact a *projected dynamical system* [107–110]. For constrained MIMO LTI systems, we established a link with *linear systems with partial state constraints* [103–106]. These links with existing research are strategic in nature, and will motivate much research that may allow cross utilization of ideas and methods in seemingly unrelated fields.

Chapter 2

Construction and Fundamental Properties

The subject of this dissertation is the *gradient projection anti-windup* (GPAW) scheme. In this chapter, we present the GPAW scheme by constructing the GPAW-compensated controller, together with some of its fundamental properties. The GPAW scheme is a generalization of the well known *conditional integration* method, which is described in Section 2.1. The extension of the conditional integration method to MIMO nonlinear systems requires the projection operator (Section 2.4), which is derived from the extension of the well-known gradient projection method for nonlinear programming (Section 2.2) to continuous time (Section 2.3). Using the projection operator, we construct the GPAW-compensated controller in Section 2.5. It is shown that under a mild restriction on the structure of the nominal controller, the projection operator enables the GPAW-compensated controller to achieve *controller state-output consistency*, a unique property among existing anti-windup schemes. In Section 2.6, we show how nominal controllers of general structure can be approximated arbitrarily well to have the required structure, so that GPAW compensation can be applied to the approximate controller yielding the same desirable properties. In Section 2.7, the projection operator is shown to be passive and with L_2-gain less than one. Some attractive features of the GPAW scheme are illustrated by a numerical example in Section 2.8. The chapter ends with a comparison of the GPAW scheme with existing anti-windup schemes in Section 2.9.

2.1 Conditional Integration

Two of the earliest known anti-windup schemes, namely the *back-calculation* method [17, Section 3.3.3, pp. 38 – 41] and the *conditional integration* method [17, Section 3.3.2, p. 38], were reported by Fertik and Ross in 1967 [53], 11 years after perhaps the first systematic study on the windup phenomenon was published in 1956 [22]. Both of these methods attempt to achieve controller state-output consistency (see Remark 2.1), by different but closely related ways (see Remark 2.4). Despite being recognized as perhaps the first paper on anti-windup compensation [12, 14], the methods reported in [53] should have been well-known among practitioners in varied forms, as evidenced by the numerous patents [16, 111–117] filed[1] prior to 1967. While the back-calculation method has been extended in various

[1] The patents [16, 111–117] are interesting, and provide a glimpse of the problems faced by the early controllers, as well as some early industrial solutions that were adopted to mitigate those problems. The

ways (e.g. to yield the *conditioning technique*, see Section 1.4.1) and continues to form the conceptual basis for numerous modern anti-windup schemes, much less effort has been devoted to the extension of the conditional integration method.[2] As will be seen, the GPAW scheme is in fact the natural extension of the conditional integration method reported by Fertik and Ross [53] for MIMO nonlinear controllers.

Remark 2.1. *Controller state-output consistency* is achieved when the controller output satisfies the saturation constraints inherently without additional limiting of the output, i.e. $\text{sat}(u_c) = u_c$. For the controller (1.3), this is achieved when (x_c, y, r) are such that $\text{sat}(g_c(x_c, y, r)) = g_c(x_c, y, r)$. Observe that this can be achieved by either modification of x_c, y, or r. The conditioning technique (see Section 1.4.1) and reference governor (see Section 1.4.6) introduce a variable r_r in place of r to achieve this, then attempts to drive r_r towards r. The standard anti-windup compensator (see Fig. 1-1) modifies the measurement y (i.e. introduce variable y_r and attempt to drive y_r towards y), and/or x_c through $\dot{x}_c$, and/or output u_c. The GPAW scheme only modifies x_c through its evolution, $\dot{x}_c$. □

Consider the description of the conditional integration method (called "accumulator technique" in [53]) by Fertik and Ross:

> "The overshoot problem discussed above for the accumulation method may be overcome (when velocity-limited) by ignoring the excess integral action when it is the same sign as the desired output."

For the PID controller described by

$$\dot{e}_i = e,$$
$$u = K_p e + K_i e_i + K_d \dot{e},$$

with saturation constraints $u_{\min} \leq u \leq u_{\max}$, $u_{\min} < 0 < u_{\max}$, this translates into the

back-calculation method was described for analog controllers in [16] (filed on May 18, 1961), which was also used for *bumpless transfer* [12] among multiple analog controllers. One form of conditional integration was described for analog controllers in [111] (filed on May 26, 1961). A form of conditional integration closest to the method proposed by Fertik and Ross in [53] was described for digital controllers in [114] (filed on Jan. 17, 1966):

> "A proportional gain channel operates to provide a term in accordance with the proportional gain constant and the error signal. The proportional term and the integral term are combined in the output register. The digital count in the output register is used to control the actuator through a digital to analog converter. The output register includes means for establishling [*sic*] limit conditions. These limit conditions are sensed, and in the event that the count is such that the controlled device is either fully opened or closed, the multi-rate sampler is inhibited during each scan, preventing accumulation of additional error pulses in the integral term until such time as control is again achieved."

Observe also that Charles W. Ross, co-author of [53], is also a co-inventor of patent [117] (filed on Mar. 14, 1968).

[2] The apparent lack of interest to extend the conditional integration method could be due to historical or technical reasons. We first note that the focus of [53] is the back-calculation method, while the conditional integration method is only mentioned in passing, described by only a single sentence quoted on this page. Early successes of anti-windup schemes derived from the back-calculation method, like the conditioning technique, could have diverted attention away from the less-proven conditional integration method. Moreover, the conditional integration method induces a *switched* or *hybrid* closed-loop system [118, 119] described by a differential equation with discontinuous right-hand side [120] (see Section 3.2). These topics being poorly understood in the early development of anti-windup schemes could be another reason for the lack of extensions of the conditional integration method.

anti-windup compensated controller[3]

$$\dot{e}_i = \begin{cases} 0, & \text{if } ((u \geq u_{\max}) \vee (u \leq u_{\min})) \wedge (eu > 0), \\ e, & \text{otherwise}, \end{cases}$$
$$u = K_p e + K_i e_i + K_d \dot{e},$$

where $A \vee B$ means "A OR B" and $A \wedge B$ means "A AND B" for *logical* statements A, B that evaluates to *true* or *false*. This can be rewritten as

$$\begin{aligned} \dot{e}_i &= \begin{cases} 0, & \text{if } (u \geq u_{\max}) \wedge (e > 0), \\ 0, & \text{if } (u \leq u_{\min}) \wedge (e < 0), \\ e, & \text{otherwise}, \end{cases} \\ u &= K_p e + K_i e_i + K_d \dot{e}. \end{aligned} \tag{2.1}$$

As can be seen, the fundamental idea is simple, intuitive, and elegant. It simply stops integrating only when such integration will cause the term $K_i e_i$ to increase/decrease in a manner to aggravate the existing saturation constraint violation.

Remark 2.2. Observe that the conditional integration method manipulates only the controller state (in this case, the integral state) e_i through its evolution $\dot{e}_i$. An equivalent interpretation is that it stops integration when the nominal state update will drive e_i further away from the unsaturated region $K(e, \dot{e}) = \{\bar{e}_i \in \mathbb{R} \mid \operatorname{sat}(K_p e + K_i \bar{e}_i + K_d \dot{e}) = K_p e + K_i \bar{e}_i + K_d \dot{e}\}$ (in this case an interval) under saturation. □

Now, consider extending this idea to m *decoupled* nonlinear controllers, described by

$$\begin{aligned} \dot{x}_{ci} &= f_{ci}(x_{ci}, y, r), \\ u_{ci} &= g_{ci}(x_{ci}, y, r), \end{aligned} \qquad i \in \mathcal{I}_m := \{1, 2, \ldots, m\},$$

subject to $u_{\min,i} \leq u_{ci} \leq u_{\max,i}$ for all $i \in \mathcal{I}_m$, where (y, r) are the measurement and reference input respectively. Observe that $(u_{ci}, \dot{x}_{ci})$ do not depend on x_{cj} for all $j \neq i$, and x_{ci} is analogous to the integral state of the PID controller, e_i. To apply the same idea (stop integration whenever it will aggravate saturation constraints) to these *decoupled* nonlinear controllers, we want to achieve $\dot{x}_{ci} = 0$ when $u_{ci} \geq u_{\max,i}$ and the nominal update will drive the state towards further constraint violations, i.e. when $\frac{\partial g_{ci}}{\partial x_{ci}}(x_{ci}, y, r) f_{ci}(x_{ci}, y, r) > 0$, and analogously when $u_{ci} \leq u_{\min,i}$. Then the anti-windup compensated controller is given by

$$\begin{aligned} \dot{x}_{ci} &= \begin{cases} 0, & \text{if } (u_{ci} \geq u_{\max,i}) \wedge \left(\frac{\partial g_{ci}}{\partial x_{ci}}(x_{ci}, y, r) f_{ci}(x_{ci}, y, r) > 0\right), \\ 0, & \text{if } (u_{ci} \leq u_{\min,i}) \wedge \left(\frac{\partial g_{ci}}{\partial x_{ci}}(x_{ci}, y, r) f_{ci}(x_{ci}, y, r) < 0\right), \\ f_{ci}(x_{ci}, y, r), & \text{otherwise}, \end{cases} \\ u_{ci} &= g_{ci}(x_{ci}, y, r), \end{aligned} \tag{2.2}$$

for all $i \in \mathcal{I}_m$. It is clear that the preceding also applies to single-output nonlinear controllers,[4] i.e. $m = 1$.

Remark 2.3. As shown in Appendix A, single-output GPAW-compensated controllers re-

[3] For the conditional integration to function as intended, we need $K_i \geq 0$.

[4] This is *not* the recommended way to apply GPAW compensation. The GPAW scheme exploits a subtle but important construction not present here. See Sections 2.5, 2.6, and Appendix C for details.

duces to a form similar to (2.2) (compare with the closed-form expressions (A.5)). This is expected because the GPAW scheme is a generalization of the conditional integration method. □

Now, consider the case for general *coupled* MIMO nonlinear controllers (1.3),

$$\begin{aligned}\dot{x}_c &= f_c(x_c, y, r),\\ u_c &= g_c(x_c, y, r),\end{aligned}$$

where $u_c = [u_{c1}, u_{c2}, \ldots, u_{cm}]^{\mathrm{T}}$, subject to $u_{\min,i} \le u_{ci} \le u_{\max,i}$ for all $i \in \mathcal{I}_m$. In extending the same idea, it is clear that we cannot selectively stop the integration of any particular element of its state x_c, since each element will in general affect more than one element of the controller output. Doing so may adversely affect those controller output elements that have not yet reached saturation. Furthermore, stopping integration on *all* elements of x_c whenever any u_{ci} is saturated is far too conservative. What is needed then, at a fixed point in time, is a way to *update the controller state vector in a manner as close as possible to the nominal update,*[5] *while attempting not to aggravate any existing saturation constraints.* In other words, at each fixed time, we want to find an $\tilde{f}$ as "close" to $f_c(x_c, y, r)$ as possible, such that the state evolution governed by $\dot{x}_c = \tilde{f}$ keeps within the unsaturated region $K(y, r) := \{\bar{x} \in \mathbb{R}^q \mid g_c(\bar{x}, y, r) = \mathrm{sat}(g_c(\bar{x}, y, r))\}$ as much as possible. One way to achieve this is by gradient projection, where we project the nominal update $f_c(x_c, y, r)$ onto $K(y, r)$.

Remark 2.4. Both the back-calculation and conditional integration methods attempt to achieve controller state-output consistency. The back-calculation method *enforces* this requirement (possibly by solving a system of nonlinear equations $g_c(x_c, y, r) = \mathrm{sat}(u_c)$ for x_c, y, or r), while the conditional integration method attempts to achieve this only by *not aggravating any existing constraint violations.* In this sense, the conditional integration method can be seen as a relaxation of the back-calculation method. Moreover, when the output equation is of general structure, i.e. $u_c = g_c(x_c, y, r)$ and $\left[\frac{\partial g_c}{\partial y}, \frac{\partial g_c}{\partial r}\right] \not\equiv 0 \in \mathbb{R}^{m\times(p+n_r)}$ (see (1.3)), it achieves controller state-output consistency only in an approximate sense.[6] □

2.2 Gradient Projection Method for Nonlinear Programming

To extend the conditional integration method to coupled MIMO nonlinear controllers, we need the gradient projection operator. Here, we describe the *gradient projection method* for nonlinear programming [100, 101]. It will be extended to continuous time in Section 2.3, after which, the desired operator is "extracted" in Section 2.4.

The gradient projection method [100, 101] solves constrained nonlinear programs of the form[7]

$$\begin{aligned}&\min_{x\in\mathbb{R}^q} J(x),\\ &\text{subject to} \qquad \tilde{h}(x) \le 0,\end{aligned} \tag{2.3}$$

where $x \in \mathbb{R}^q$ is the decision variable, $J(x)$ is a possibly nonlinear scalar function $J\colon \mathbb{R}^q \to \mathbb{R}$, and $\tilde{h}(x) = [\tilde{h}_1(x), \tilde{h}_2(x), \ldots, \tilde{h}_k(x)]^{\mathrm{T}}$ is a set of k possibly nonlinear functions $\tilde{h}\colon \mathbb{R}^q \to \mathbb{R}^k$. Note that $\tilde{h}(x) \le 0$ is to be interpreted element-wise as k scalar inequalities. In its basic

[5] Note that this does not mean *direction preservation* as described in [15].

[6] This is explained in Section 2.5, in particular, Remark 2.13.

[7] Here, we use $\tilde{h}$ instead of h to avoid notational conflicts in Section 2.5.

form, the gradient projection method is very powerful, with only very mild differentiability requirements on the functions J and $\tilde{h}$. However, some additional assumptions like convexity of J and/or $\tilde{h}$, boundedness of feasible region etc., must be imposed to ensure convergence to the global minimum [100, 101]. For our purposes, only the projection mechanism is of interest, so that analogously, only differentiability of $\tilde{h}$ needs to be assumed.

In the absence of any active constraints, the gradient projection method reduces to the steepest descent method [100, 101]. The key mechanism that enables the method to maintain feasibility is gradient projection. Each of the k inequalities $\tilde{h}_i(x) \leq 0$, $i \in \mathcal{I}_k$, defines a hypersurface $G_i = \{\bar{x} \in \mathbb{R}^q \mid \tilde{h}_i(\bar{x}) = 0\} \subset \mathbb{R}^q$ that forms the boundary of the feasible region

$$\tilde{K} = \{\bar{x} \in \mathbb{R}^q \mid \tilde{h}(\bar{x}) \leq 0\}.$$

On each point $\tilde{x}$ of the boundary of $\tilde{K}$, each hypersurface G_i that contains $\tilde{x}$ has an associated *supporting hyperplane* $H_i(\tilde{x})$ that is tangent to G_i at $\tilde{x}$. The normal of H_i at the point $\tilde{x}$ is the gradient of $\tilde{h}_i(x)$ at $\tilde{x}$ (denoted by $\nabla\tilde{h}_i(\tilde{x})$), which will point "away" from $\tilde{K}$. These are illustrated in Fig. 2-1.

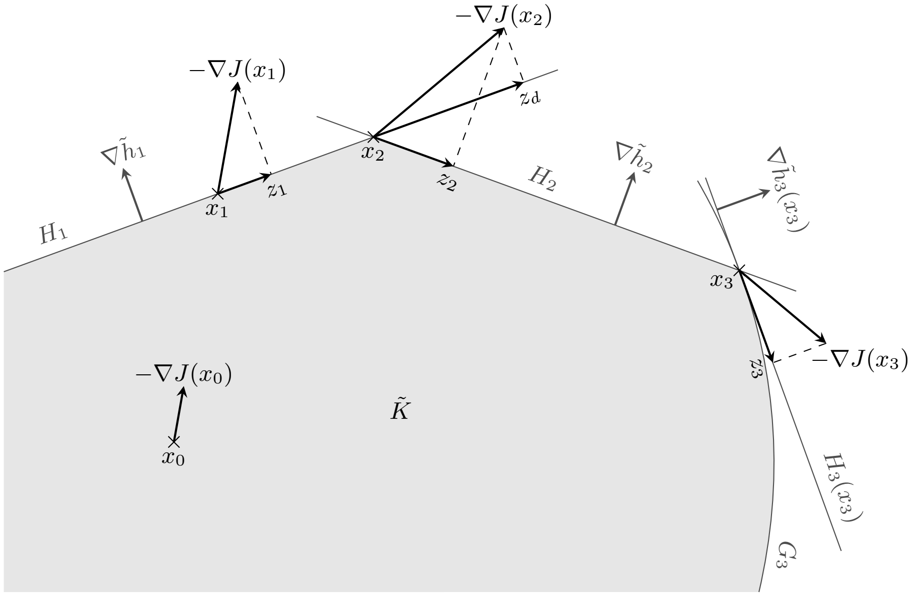


Figure 2-1: Visualization of the gradient projection method. $\tilde{K}$ is the feasible region, bounded by the hypersurfaces H_1, H_2, and G_3. The supporting hyperplane of G_3 at x_3 is $H_3(x_3)$. The projection of $-\nabla J(x_i)$ onto H_i yields z_i, while z_d is the projection of $-\nabla J(x_2)$ onto H_1. Notice that to maintain feasibility at x_2, it is sufficient to project onto H_2. In contrast, projection onto the intersection of *both* active constraints corresponding to H_1 and H_2 will yield the zero vector.

Similar to many optimization methods, the gradient projection method generates a sequence $\{x_n\}$, the limiting point of which would be the solution to the constrained nonlinear program (2.3). Consider now, the case where all $\tilde{h}_i(x)$ are affine functions of x [100]. Then G_i coincides with H_i, and the boundary of $\tilde{K}$ are all hyperplanes. At a particular point x_n that lies in the interior of $\tilde{K}$ (cf. x_0 in Fig. 2-1), the basic step is taken in a direction to decrease $J(x_n)$, i.e. in the negative gradient direction $-\nabla J(x_n)$, much like the steepest descent method [121, pp. 25 – 26]. When x_n lies on the boundary of $\tilde{K}$ (cf. x_1, x_2, and x_3 in Fig. 2-1), the step is taken in a direction "closest" to $-\nabla J(x_n)$ while enforcing $x_{n+1} \in \tilde{K}$. In this case, if $-\nabla J(x_n)$ points into the interior of $\tilde{K}$, the nominal direction is taken. Otherwise, $-\nabla J(x_n)$ is projected onto the intersection of the *smallest* set of *linearly independent* hyperplanes H_i that corresponds to active constraints ($\tilde{h}_i(x_n) \geq 0$) that can keep x_{n+1} within $\tilde{K}$ (cf. z_1, z_2, and z_3 in Fig. 2-1). The step is then taken in this new direction, and if some active constraints are nonlinear (cf. G_3 in Fig. 2-1), a correction is added to drive the new point x_{n+1} back to $\tilde{K}$ [101].

It is important to note that this *smallest* set of hyperplanes may exclude some active constraints, but nonetheless ensures that such exclusion will not cause further constraint violations. This case is illustrated in Fig. 2-1, at the point x_2. Notice here that taking a step in the direction of $-\nabla J(x_2)$ will violate both constraints corresponding to H_1 and H_2. If we project onto *both* of these active constraints, (in other words, the intersection of H_1 and H_2), the result is the zero vector, and no progress can be made. However, projecting onto H_2 alone, both constraints will be satisfied, and progress can be made in the direction z_2. Notice that projecting onto H_1 to get z_d is ineffective, since taking a step in this direction will violate the constraint corresponding to H_2.

The preceding summarizes the pertinent features of the gradient projection method relevant to the GPAW scheme. See [100, 101] for more details, including recursion relations that may allow a more computationally efficient implementation.

2.3 Continuous Time Gradient Projection Method

As shown in [122, Appendix B.4, pp. 788 – 791], the continuous time gradient projection method can be obtained by taking the limit as the stepsize of Rosen's gradient projection method [100, 101] is decreased to zero. The derived continuous time projection operator has been used successfully in the context of adaptive control to bound parameter estimates in some a priori known region in the parameter space [122, Sections 4.4, 8.4.2, and 8.5.5]. Another popular projection operator used in adaptive control is presented in [123]. While each has its merits, both of these methods are limited to projection with respect to a *single* inequality constraint. Here, we extend Rosen's gradient projection method [100, 101] to continuous time, the principal distinguishing property being its ability to accommodate *multiple* inequality constraints.

Recall the constrained nonlinear program (2.3) described by

$$\begin{aligned} &\min_{x\in\mathbb{R}^q} J(x), \\ &\text{subject to} \qquad \tilde{h}(x) \leq 0, \end{aligned} \tag{2.4}$$

where $x \in \mathbb{R}^q$ is the decision variable, $J(x)$ is the scalar objective function $J\colon \mathbb{R}^q \to \mathbb{R}$, and $\tilde{h}(x) = [\tilde{h}_1(x), \tilde{h}_2(x), \dots, \tilde{h}_k(x)]^{\mathrm{T}}$ is the function $\tilde{h}\colon \mathbb{R}^q \to \mathbb{R}^k$ defining the k scalar inequality constraints. Let $\mathcal{I}_k := \{1, 2, \dots, k\}$ be the set of indices corresponding to the k constraints in problem (2.4), and $\mathcal{I} \subset \mathcal{I}_k$ be some index set of cardinality $s := |\mathcal{I}|$ ($\leq k$). For $s > 0$, i.e. $\mathcal{I} \neq \emptyset$, let $\sigma_{\mathcal{I}}\colon \mathcal{I}_s \to \mathcal{I}$ be a (non-unique) bijection that assigns an integer in $\mathcal{I}$ to each integer in $\mathcal{I}_s$ ($= \{1, 2, \dots, s\}$). Define the $q \times \max\{s, 1\}$ matrix

$$N_{\mathcal{I}}(x) = \begin{cases} [\nabla\tilde{h}_{\sigma_{\mathcal{I}}(1)}(x), \nabla\tilde{h}_{\sigma_{\mathcal{I}}(2)}(x), \dots, \nabla\tilde{h}_{\sigma_{\mathcal{I}}(s)}(x)], & \text{if } s > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2.5}$$

where $\nabla\tilde{h}_i(\tilde{x}) = \left(\frac{\partial\tilde{h}_i}{\partial x}(\tilde{x})\right)^{\mathrm{T}} \in \mathbb{R}^q$ is the gradient of $\tilde{h}_i$ evaluated at the point $\tilde{x} \in \mathbb{R}^q$. Observe that when $s = 0$, i.e. $\mathcal{I} = \emptyset$, the matrix $N_{\mathcal{I}}(x)$ reduces to the zero vector $0 \in \mathbb{R}^{q\times 1}$. When $0 < s \leq k$, i.e. $\mathcal{I} \neq \emptyset$, the matrix $N_{\mathcal{I}}(x) \in \mathbb{R}^{q\times s}$ is the concatenation of those gradient vectors $\nabla\tilde{h}_i(x)$ whose indices are in $\mathcal{I}$, in some order determined by the map $\sigma_{\mathcal{I}}$.

Remark 2.5. Recall that a bijection is a function that establishes a one-to-one correspondence between two sets. Note also that *any* chosen bijection $\sigma_{\mathcal{I}}\colon \mathcal{I}_s \to \mathcal{I}$ suffices. For example, we can take the ascending order map defined recursively by

$$\sigma_{\mathcal{I}}(i) := \min\bigl(\mathcal{I} \setminus \textstyle\bigcup_{j=1}^{i-1}\{\sigma_{\mathcal{I}}(j)\}\bigr), \qquad \forall i \in \mathcal{I}_s.$$

A different choice of $\sigma_{\mathcal{I}}$ only results in a rearrangement of the columns of $N_{\mathcal{I}}(x)$. Our final matrix of interest, $P_{\mathcal{I}}(x)$ defined below, will be invariant with respect to such rearrangements (see Remark 2.6). Note that $\sigma_{\mathcal{I}}$ is only a construction to allow the index selection/ordering operations to be stated in a compact mathematical form, it is itself of no real significance. Indeed, as shown in Appendix A, $\sigma_{\mathcal{I}}$ is not needed to construct single-output GPAW-compensated controllers. □

For any full rank $N_{\mathcal{I}}(x)$, i.e. $\mathcal{I} \neq \emptyset$ and $\operatorname{rank}(N_{\mathcal{I}}(x)) = \min\{s, q\} = s = |\mathcal{I}|$ (> 0), define the symmetric $q \times q$ projection matrix [100, Theorem 1]

$$P_{\mathcal{I}}(x) = \begin{cases} I - N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}}N_{\mathcal{I}})^{-1}N_{\mathcal{I}}^{\mathrm{T}}(x), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise,} \end{cases} \tag{2.6}$$

where $N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}}N_{\mathcal{I}})^{-1}N_{\mathcal{I}}^{\mathrm{T}}(x) := N_{\mathcal{I}}(x)(N_{\mathcal{I}}^{\mathrm{T}}(x)N_{\mathcal{I}}(x))^{-1}N_{\mathcal{I}}^{\mathrm{T}}(x)$. Since $\operatorname{rank}(N_{\mathcal{I}}^{\mathrm{T}}N_{\mathcal{I}}(x)) = \operatorname{rank}(N_{\mathcal{I}}(x))$ [124, p. 13], whenever $N_{\mathcal{I}}(x)$ is full rank, the $s\times s$ inverse matrix $(N_{\mathcal{I}}^{\mathrm{T}}N_{\mathcal{I}}(x))^{-1}$ exists, and $P_{\mathcal{I}}(x)$ is well defined. As shown in [100, Theorem 1], when $N_{\mathcal{I}}(x)$ is full rank, $P_{\mathcal{I}}(x)$ takes any $z \in \mathbb{R}^q$ into the orthogonal complement of the subspace spanned by the columns of $N_{\mathcal{I}}(x)$. In other words, for any $x \in \mathbb{R}^q$ lying on the constraint boundary and any $z \in \mathbb{R}^q$, $P_{\mathcal{I}}(x)z$ will be parallel to the intersection of all supporting hyperplanes $H_i(x)$ (see Fig. 2-1) whose indices are in $\mathcal{I}$.

Remark 2.6. Clearly, $P_{\mathcal{I}}(x)$ is independent of the bijection $\sigma_{\mathcal{I}}$ when $\mathcal{I} = \emptyset$. To see that $P_{\mathcal{I}}(x)$ is invariant with respect to the choice of $\sigma_{\mathcal{I}}$ when $\mathcal{I} \neq \emptyset$, let $\tilde{N}_{\mathcal{I}}(x)$ be defined by another bijection $\tilde{\sigma}_{\mathcal{I}}\colon \mathcal{I}_s \to \mathcal{I}$, i.e. $\tilde{N}_{\mathcal{I}}(x) = [\nabla\tilde{h}_{\tilde{\sigma}_{\mathcal{I}}(1)}(x), \nabla\tilde{h}_{\tilde{\sigma}_{\mathcal{I}}(2)}(x), \ldots, \nabla\tilde{h}_{\tilde{\sigma}_{\mathcal{I}}(s)}(x)]$. In effect, $\tilde{\sigma}_{\mathcal{I}}$ induces a rearrangement of the columns of $N_{\mathcal{I}}(x)$, so that we can write $\tilde{N}_{\mathcal{I}}(x) = N_{\mathcal{I}}(x)P$ for some nonsingular *permutation matrix* P [124, pp. 25 – 26]. Then when $N_{\mathcal{I}}(x)$ is full rank, we have

$$\tilde{N}_{\mathcal{I}}(\tilde{N}_{\mathcal{I}}^{\mathrm{T}}\tilde{N}_{\mathcal{I}})^{-1}\tilde{N}_{\mathcal{I}}^{\mathrm{T}}(x) = N_{\mathcal{I}}P(P^{\mathrm{T}}N_{\mathcal{I}}^{\mathrm{T}}N_{\mathcal{I}}P)^{-1}P^{\mathrm{T}}N_{\mathcal{I}}^{\mathrm{T}}(x) = N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}}N_{\mathcal{I}})^{-1}N_{\mathcal{I}}^{\mathrm{T}}(x).$$

The preceding with (2.6) show that different choices of $\sigma_{\mathcal{I}}$ do not alter $P_{\mathcal{I}}(x)$. □

For any fixed $x \in \mathbb{R}^q$, define the sets

$$\mathcal{I}_{\mathrm{act}} := \mathcal{I}_{\mathrm{act}}(x) = \{i \in \mathcal{I}_k \mid \tilde{h}_i(x) \geq 0\}, \qquad \mathcal{J} := \{\mathcal{I} \subset \mathcal{I}_{\mathrm{act}} \mid |\mathcal{I}| \leq q\}. \tag{2.7}$$

It can be seen that $\mathcal{I}_{\mathrm{act}}$ is the set of indices of all active constraints, and $\mathcal{J}$ is the set of all subsets of $\mathcal{I}_{\mathrm{act}}$ with cardinality less than or equal to q. For any fixed $x \in \mathbb{R}^q$, define the following *combinatorial* optimization subproblem

$$\begin{aligned} &\max_{\mathcal{I}\in\mathcal{J}} \|P_{\mathcal{I}}(x)\nabla J(x)\|, \\ \text{subject to} \qquad &\operatorname{rank}(N_{\mathcal{I}}(x)) = |\mathcal{I}|, \\ &N^{\mathrm{T}}_{\mathcal{I}_{\text{act}}}(x)P_{\mathcal{I}}(x)\nabla J(x) \geq 0. \end{aligned} \tag{2.8}$$

Notice that $\operatorname{rank}(N_{\mathcal{I}}(x)) = |\mathcal{I}|$ holds if and only if either $\mathcal{I} = \emptyset$ or $N_{\mathcal{I}}(x)$ is full rank. Moreover, the condition $N^{\mathrm{T}}_{\mathcal{I}_{\text{act}}}(x)P_{\mathcal{I}}(x)\nabla J(x) \geq 0$ holds if and only if either x is in the interior of $\tilde{K} = \{\bar{x} \in \mathbb{R}^q \mid \tilde{h}(\bar{x}) \leq 0\}$ (for which $\mathcal{I}_{\text{act}} = \emptyset$ and $N_{\mathcal{I}_{\text{act}}} = 0 \in \mathbb{R}^q$), or x is on the boundary of $\tilde{K}$ and $-P_{\mathcal{I}}(x)\nabla J(x)$ points into $\tilde{K}$ from x. Finally, since $\operatorname{rank}(N_{\mathcal{I}}(x)) \leq q$ for all $\mathcal{I} \subset \mathcal{I}_{\text{act}}$, $\mathcal{J}$ as defined in (2.7) is an exhaustive set of candidate solutions to subproblem (2.8). In summary, subproblem (2.8) is to find a subset of $\mathcal{I}_{\text{act}}$ such that the supporting hyperplanes whose indices are in this subset are linearly independent, the projection of $-\nabla J(x)$ onto the intersection of these hyperplanes is maximal in magnitude, and when x is evolved in the projection $-P_{\mathcal{I}}(x)\nabla J(x)$, no constraints will be violated. The following result asserts the existence of solutions to subproblem (2.8).

Proposition 2.3.1 (Existence of Solutions to Combinatorial Optimization Subproblem)**.** *For any fixed $x \in \mathbb{R}^q$, there exists a solution to subproblem* (2.8)*.*

Proof. It is sufficient to show that there always exists a feasible (not necessarily optimal) solution to subproblem (2.8). If $v := \operatorname{rank}(N_{\mathcal{I}_{\text{act}}}(x)) = 0$ (which includes the case $\mathcal{I}_{\text{act}} = \emptyset$), then $N_{\mathcal{I}_{\text{act}}}(x)$ must be a $q \times \max\{|\mathcal{I}_{\text{act}}|, 1\}$ zero matrix. In this case, it can be verified that $\mathcal{I} = \emptyset$ is a feasible solution (in fact, the only feasible solution) to subproblem (2.8).

If $v > 0$ (necessarily, $v \leq \min\{|\mathcal{I}_{\text{act}}|, q\}$), then $N_{\mathcal{I}_{\text{act}}}(x)$ has exactly v *linearly independent* columns, so that there exists $\mathcal{I} \subset \mathcal{I}_{\text{act}}$ such that $\operatorname{rank}(N_{\mathcal{I}}(x)) = v = |\mathcal{I}|$, satisfying the first constraint of subproblem (2.8). Any column of $N_{\mathcal{I}_{\text{act}}}(x)$ can then be expressed as a linear combination of the columns of $N_{\mathcal{I}}(x)$, so that $N_{\mathcal{I}_{\text{act}}}(x) = N_{\mathcal{I}}(x)\Psi$ for some $\Psi \in \mathbb{R}^{v\times|\mathcal{I}_{\text{act}}|}$. From (2.6), we have $N^{\mathrm{T}}_{\mathcal{I}}(x)P_{\mathcal{I}}(x) = 0$. Since $N^{\mathrm{T}}_{\mathcal{I}_{\text{act}}}(x)P_{\mathcal{I}}(x) = \Psi^{\mathrm{T}}N^{\mathrm{T}}_{\mathcal{I}}(x)P_{\mathcal{I}}(x) = 0$, the second constraint of subproblem (2.8) holds, which shows $\mathcal{I}$ to be a feasible solution. ■

Remark 2.7. As noted in the proof of Proposition 2.3.1, $N^{\mathrm{T}}_{\mathcal{I}}(x)P_{\mathcal{I}}(x) = 0$ for any well defined $P_{\mathcal{I}}(x)$. Note that when $\mathcal{I} = \emptyset$, we have $N^{\mathrm{T}}_{\mathcal{I}}(x)P_{\mathcal{I}}(x) = 0 \cdot I = 0$ (see (2.5) and (2.6)). Writing $\mathcal{I}_{\text{act}} = (\mathcal{I}_{\text{act}} \setminus \mathcal{I}) \cup \mathcal{I}$, the second constraint of subproblem (2.8) is equivalent to $N^{\mathrm{T}}_{\mathcal{I}_{\text{act}}\setminus\mathcal{I}}(x)P_{\mathcal{I}}(x)\nabla J(x) \geq 0$ and $N^{\mathrm{T}}_{\mathcal{I}}(x)P_{\mathcal{I}}(x)\nabla J(x) \geq 0$ combined. Since $N^{\mathrm{T}}_{\mathcal{I}}(x)P_{\mathcal{I}}(x)\nabla J(x) \equiv 0$, the second constraint in subproblem (2.8) can be replaced by $N^{\mathrm{T}}_{\mathcal{I}_{\text{act}}\setminus\mathcal{I}}(x)P_{\mathcal{I}}(x)\nabla J(x) \geq 0$ without affecting the solution. This yields possibly less conditions to verify ($|\mathcal{I}_{\text{act}}\setminus\mathcal{I}| \leq |\mathcal{I}_{\text{act}}|$), and may result in some (marginal) savings in computation. Note that Proposition 2.3.1 can be easily adapted for this variant. □

Remark 2.8. It can be verified that for a finite number of constraints k (see (2.4)), there is only a finite number of candidate solutions given by

$$|\mathcal{J}| = \sum_{i=0}^{\min\{q,|\mathcal{I}_{\text{act}}|\}} \binom{|\mathcal{I}_{\text{act}}|}{i} < \infty,$$

where $|\mathcal{I}_{\text{act}}| \leq k$ and $\binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{k(k-1)\ldots 2\cdot 1}$ is the binomial coefficient [125, pp. 824 – 825]. Hence an optimal solution to subproblem (2.8) can always be found by an exhaustive search algorithm, and solvability is not an issue. See Section 4.1 as well as Appendices A and B for more properties of the combinatorial optimization subproblem (2.8), and alternative solution methods. □

At each fixed time, let $\mathcal{I}^*$ be a solution to subproblem (2.8). The continuous time gradient projection method is then given by the update

$$\dot{x} = -P_{\mathcal{I}^*}(x)\nabla J(x), \qquad x(0) = x_{ig},$$

where x_{ig} is the initial guess. Proposition 2.3.1 ensures the existence of $\mathcal{I}^*$, while Remark 2.8 shows subproblem 2.8 to be solvable. Together, they show that the continuous time gradient projection method is well defined.

Our interest in the continuous time gradient projection method lies only in the projection operator. As such, we leave the analyses to show that it indeed solves problem (2.4) as future work (see Section 7.1.4). The following example demonstrates that the continuous time gradient projection method can solve a convex nonlinear program. Moreover, when appropriately initialized, it maintains feasibility at all times.

Example 2.3.1. Consider the convex nonlinear program

$$\begin{aligned} &\min_{x\in\mathbb{R}^2} J(x) = x^{\mathrm{T}}Q_1 x, \\ \text{subject to} \qquad &\tilde{h}_1(x) = a_1^{\mathrm{T}}x + b_1 \leq 0, \\ &\tilde{h}_2(x) = (x - x_{cen})^{\mathrm{T}}Q_2(x - x_{cen}) + b_2 \leq 0, \end{aligned}$$

where

$$\begin{aligned} Q_1 &= \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix} > 0, \qquad a_1 = \begin{bmatrix} -4 \\ -1 \end{bmatrix}, \qquad b_1 = 4, \\ Q_2 &= \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} > 0, \qquad x_{cen} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \qquad b_2 = -10. \end{aligned}$$

The feasible region $\tilde{K} = \{\bar{x} \in \mathbb{R}^2 \mid \tilde{h}_1(\bar{x}) \leq 0, \tilde{h}_2(\bar{x}) \leq 0\}$ is bounded by the line G_1 and ellipse G_2, defined by

$$\begin{aligned} G_1 &= \{(\bar{x}_1, \bar{x}_2) \in \mathbb{R}^2 \mid 4\bar{x}_1 + \bar{x}_2 = 4\}, \\ G_2 &= \{\bar{x} \in \mathbb{R}^2 \mid (\bar{x} - x_{cen})^{\mathrm{T}}Q_2(\bar{x} - x_{cen}) + b_2 = 0\}, \end{aligned}$$

and illustrated in Fig. 2-2. It can be verified from [121, Proposition 3.3.2, p. 320] and [121, Proposition 2.1.1, p. 193] that the unique global minimum is $x^* = [1, 1]^{\mathrm{T}}$.

The solution trajectory of the continuous time gradient projection method is defined by

$$\dot{x} = -P_{\mathcal{I}^*}(x)\nabla J(x) = -2P_{\mathcal{I}^*}(x)Q_1 x, \qquad x(0) = x_{ig},$$

where at each fixed time, $\mathcal{I}^*$ is a solution to the combinatorial optimization subproblem

$$\begin{aligned} &\min_{\mathcal{I}\in\mathcal{J}} \|P_{\mathcal{I}}(x)Q_1 x\|, \\ \text{subject to} \qquad &\operatorname{rank}(N_{\mathcal{I}}(x)) = |\mathcal{I}|, \\ &N_{\mathcal{I}_{\text{act}}}^{\mathrm{T}} P_{\mathcal{I}}(x)Q_1 x \geq 0, \end{aligned}$$

with $\mathcal{I}_{\text{act}}$, $\mathcal{J}$, $P_{\mathcal{I}}(x)$, and $N_{\mathcal{I}}(x)$ defined as before.

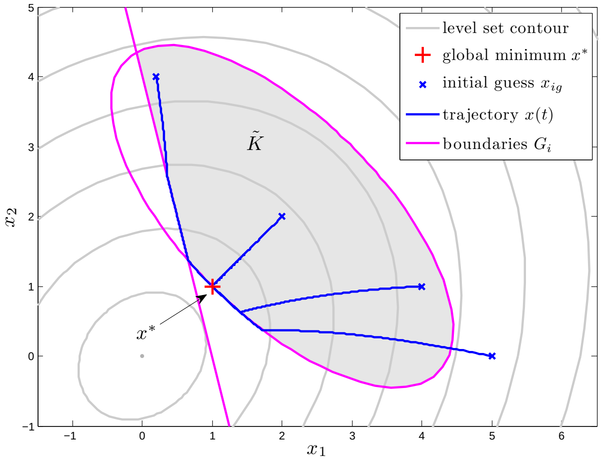


Figure 2-2: Continuous time gradient projection method applied to a convex nonlinear program. All solutions starting within the feasible region $\tilde{K}$ converged to the global minimum x^*, maintaining feasibility at all times. Observe that the solution starting from $(5, 0)$, outside the feasible region, also converged to x^* in this instance.

The simulation results in Fig. 2-2 show that all trajectories considered converged to the global minimum x^*. It can be seen that the trajectories "slide" along the constraint boundaries when the nominal update would have caused constraint violations. Observe that when started within the feasible region $\tilde{K}$ ($x_{ig} \in \{(0.2, 4), (2, 2), (4, 1)\}$), feasibility is maintained at all times, i.e. $x(t) \in \tilde{K}$ for all $t \geq 0$. Starting from the point $x_{ig} = (5, 0)$ outside the feasible region, the solution also converged to the global minimum. Observe that the projection mechanism allows solutions to *enter*, but never to leave the feasible region $\tilde{K}$. In general, convergence to the global minimum starting from the infeasible region is not guaranteed. △

2.3.1 Scaled Continuous Time Gradient Projection Method

Here, we derive a scaled version of the continuous time gradient projection method in similar manner as in [122, Appendix B.4, pp. 788 – 791]. When the resulting projection operator (see Section 2.4) is used in the GPAW scheme, the scaled variant introduces a parameter that can be used to tune the GPAW-compensated controller. To obtain the scaled method, define the transformed decision variable $\tilde{x}$ by $x = \Phi\tilde{x}$, for some *nonsingular* scaling matrix $\Phi \in \mathbb{R}^{q \times q}$. In this new coordinate system, the constrained nonlinear program (2.4) becomes

$$\begin{aligned} &\min_{\tilde{x} \in \mathbb{R}^q} J(\Phi\tilde{x}), \\ &\text{subject to} \qquad \tilde{h}(\Phi\tilde{x}) \leq 0, \end{aligned} \tag{2.9}$$

with associated gradient vectors and transformed matrix

$$\nabla_{\tilde{x}} J(\Phi\tilde{x}) := \left(\frac{\partial J(\Phi\tilde{x})}{\partial \tilde{x}}\right)^{\mathrm{T}} = \left(\frac{\partial J(\Phi\tilde{x})}{\partial x}\Phi\right)^{\mathrm{T}} = \Phi^{\mathrm{T}} \nabla J(\Phi\tilde{x}), \tag{2.10}$$

$$\nabla_{\tilde{x}} \tilde{h}_i(\Phi\tilde{x}) := \left(\frac{\partial \tilde{h}_i(\Phi\tilde{x})}{\partial \tilde{x}}\right)^{\mathrm{T}} = \left(\frac{\partial \tilde{h}_i(\Phi\tilde{x})}{\partial x}\Phi\right)^{\mathrm{T}} = \Phi^{\mathrm{T}} \nabla \tilde{h}_i(\Phi\tilde{x}), \qquad \forall i \in \mathcal{I}_k,$$

$$\tilde{N}_{\mathcal{I}}(\Phi\tilde{x}) = \Phi^{\mathrm{T}} N_{\mathcal{I}}(\Phi\tilde{x}). \tag{2.11}$$

For a full rank $N_{\mathcal{I}}(\Phi\tilde{x})$, the corresponding projection matrix (obtained from (2.6) by replacing $N_{\mathcal{I}}(x)$ with $\tilde{N}_{\mathcal{I}}(x)$ and using (2.11)), is

$$\tilde{P}_{\mathcal{I}}(\Phi\tilde{x}) = \begin{cases} I - \Phi^{\mathrm{T}} N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}} \Gamma N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(\Phi\tilde{x})\Phi, & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise}, \end{cases} \tag{2.12}$$

where $\Gamma := \Phi\Phi^{\mathrm{T}} \in \mathbb{R}^{q\times q}$ is symmetric positive definite. Written in terms of x, subproblem (2.8) becomes

$$\begin{aligned} &\max_{\mathcal{I}\in\mathcal{J}} \|\tilde{P}_{\mathcal{I}}(x)\nabla_{\tilde{x}} J(x)\|, \\ \text{subject to} \qquad &\operatorname{rank}(N_{\mathcal{I}}(x)) = |\mathcal{I}|, \\ &\tilde{N}_{\mathcal{I}_{\text{act}}}^{\mathrm{T}}(x)\tilde{P}_{\mathcal{I}}(x)\nabla_{\tilde{x}} J(x) \geq 0, \end{aligned} \tag{2.13}$$

where we have used $\operatorname{rank}(\tilde{N}_{\mathcal{I}}(x)) = \operatorname{rank}(N_{\mathcal{I}}(x))$.

At each fixed time, let $\mathcal{I}^*$ be a solution to subproblem (2.13). The scaled continuous time gradient projection method for problem (2.9) is then given by the update

$$\dot{\tilde{x}} = -\tilde{P}_{\mathcal{I}^*}(\Phi\tilde{x})\nabla_{\tilde{x}} J(\Phi\tilde{x}), \qquad \tilde{x}(0) = \Phi^{-1}x_{ig}.$$

Using $x = \Phi\tilde{x}$ and (2.10), this becomes

$$\dot{x} = -\Phi\tilde{P}_{\mathcal{I}^*}(x)\Phi^{\mathrm{T}}\nabla J(x), \qquad x(0) = x_{ig}, \tag{2.14}$$

in the original coordinates.

Remark 2.9. The effect of introducing scaling is to perform the projection in the transformed space. This changes the way the solution evolves on the boundary of the feasible region. □

2.4 Projection Operator

Section 2.3 extended Rosen's gradient projection method [100, 101] to continuous time, while Section 2.3.1 derived a scaled variant. Here, we show that the continuous time gradient projection method can be separated into two components: a part analogous to the steepest descent method for optimization, and a projection operator for constraint satisfaction. The projection operator is then shown to be applicable to a larger class of constrained problems governed by an ordinary differential equation (ODE).

Using (2.12), we can define

$$R_{\mathcal{I}}(x) := \Phi\tilde{P}_{\mathcal{I}}(x)\Phi^{-1} = \begin{cases} I - \Gamma N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}}\Gamma N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(x), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise}, \end{cases} \tag{2.15}$$

where $\Gamma = \Gamma^{\mathrm{T}} = \Phi\Phi^{\mathrm{T}} > 0$ is defined in (2.12). The scaled continuous time update (2.14) can then be expressed as

$$\dot{x} = -\Phi\tilde{P}_{\mathcal{I}^*}(x)\Phi^{-1}\Phi\Phi^{\mathrm{T}}\nabla J(x) = -R_{\mathcal{I}^*}(x)\Gamma\nabla J(x), \qquad x(0) = x_{ig}, \tag{2.16}$$

where $\mathcal{I}^*$ is a solution to subproblem (2.13).

Observe from (2.7) that when no constraints are active, i.e. $\mathcal{I}_{\text{act}} = \emptyset$ and $\mathcal{J} = \{\emptyset\}$, the unique optimal solution to subproblem (2.13) is $\mathcal{I}^* = \emptyset$, so that by (2.15), the update (2.16) reduces to

$$\dot{x} = -\Gamma\nabla J(x).$$

It can be verified that this update is equivalent to the scaled continuous time steepest descent update [122, Appendix B.2, pp. 785 – 786]. From (2.16), we see that the scaled continuous time gradient projection method is composed of two parts: the nominal steepest descent algorithm for minimization of the objective function, i.e. $-\Gamma\nabla J(x)$, and the *projection operator* $R_{\mathcal{I}^*}(x)$ for constraint satisfaction. When x is on the constraint boundary at time t_0, the projection operator $R_{\mathcal{I}^*}(x)$ projects the nominal update $-\Gamma\nabla J(x)$ such that its evolution governed by (2.16) ensures

$$x(t) \in \tilde{K} = \{\bar{x} \in \mathbb{R}^q \mid \tilde{h}(\bar{x}) \leq 0\}, \qquad \forall t \geq t_0. \tag{2.17}$$

Conceivably, the projection operator can also be used to enforce a set of constraints

$$\tilde{h}(x) = [\tilde{h}_1(x), \tilde{h}_2(x), \ldots, \tilde{h}_k(x)]^{\mathrm{T}} \leq 0, \tag{2.18}$$

for some process governed by an ODE of the form

$$\dot{x} = \tilde{f}(t, x), \qquad x(0) = x_{ig}. \tag{2.19}$$

Applying the projection operator to this constrained process, i.e. replacing $-\Gamma\nabla J(x)$ in (2.16) by $\tilde{f}(t, x)$, results in the update

$$\dot{x} = R_{\mathcal{I}^*}(x)\tilde{f}(t, x), \qquad x(0) = x_{ig}, \tag{2.20}$$

where $\mathcal{I}^*$ is a solution to an analogous combinatorial optimization subproblem.

This analogous subproblem is obtained from (2.13) by replacing $-\Gamma\nabla J(x)$ with $\tilde{f}(t, x)$, or equivalently (see (2.10)), replacing $\nabla_{\tilde{x}} J(x)$ with $-\Phi^{-1}\tilde{f}(t, x)$. Using (2.15) and (2.11), this replacement yields the subproblem

$$\begin{aligned} &\max_{\mathcal{I}\in\mathcal{J}} \|\Phi^{-1} R_{\mathcal{I}}(x)\tilde{f}(t, x)\|, \\ \text{subject to} \qquad &\mathrm{rank}(N_{\mathcal{I}}(x)) = |\mathcal{I}|, \\ &N_{\mathcal{I}_{\mathrm{act}}}^{\mathrm{T}}(x) R_{\mathcal{I}}(x)\tilde{f}(t, x) \leq 0, \end{aligned} \tag{2.21}$$

where $\mathcal{I}_{\mathrm{act}}$, $\mathcal{J}$ are defined in (2.7), $N_{\mathcal{I}}(x)$ is defined in (2.5), reproduced below

$$\mathcal{I}_{\mathrm{act}} := \mathcal{I}_{\mathrm{act}}(x) = \{i \in \mathcal{I}_k \mid \tilde{h}_i(x) \geq 0\}, \qquad \mathcal{J} := \{\mathcal{I} \subset \mathcal{I}_{\mathrm{act}} \mid |\mathcal{I}| \leq q\},$$
$$N_{\mathcal{I}}(x) = \begin{cases} [\nabla\tilde{h}_{\sigma_{\mathcal{I}}(1)}(x), \nabla\tilde{h}_{\sigma_{\mathcal{I}}(2)}(x), \ldots, \nabla\tilde{h}_{\sigma_{\mathcal{I}}(|\mathcal{I}|)}(x)], & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases}$$

and the bijection $\sigma_{\mathcal{I}}$ is described in Remark 2.5.

Recognizing that maximizing a positive semidefinite function (like the objective function of subproblem (2.21)) is equivalent to maximizing its square, we see that an equivalent subproblem for the definition of $\mathcal{I}^*$ in (2.20) is

$$\begin{aligned}\max_{\mathcal{I}\in\mathcal{J}} \; & \tilde{f}^{\mathrm{T}}(t,x)\Gamma^{-1}R_{\mathcal{I}}(x)\tilde{f}(t,x),\\ \text{subject to} \quad & \operatorname{rank}(N_{\mathcal{I}}(x)) = |\mathcal{I}|,\\ & N^{\mathrm{T}}_{\mathcal{I}_{\mathrm{act}}}(x)R_{\mathcal{I}}(x)\tilde{f}(t,x) \le 0,\end{aligned} \tag{2.22}$$

since from (2.15), we have $R^{\mathrm{T}}_{\mathcal{I}}(x)\Gamma^{-1} = \Gamma^{-1}R_{\mathcal{I}}(x)$ and $R^2_{\mathcal{I}}(x) = R_{\mathcal{I}}(x)$ (i.e. $R_{\mathcal{I}}(x)$ is idempotent [126, p. 697]), and the square of the objective function of subproblem (2.21) is

$$\begin{aligned}\|\Phi^{-1}R_{\mathcal{I}}(x)\tilde{f}(t,x)\|^2 &= \tilde{f}^{\mathrm{T}}(t,x)R^{\mathrm{T}}_{\mathcal{I}}(x)\Phi^{-\mathrm{T}}\Phi^{-1}R_{\mathcal{I}}(x)\tilde{f}(t,x),\\ &= \tilde{f}^{\mathrm{T}}(t,x)R^{\mathrm{T}}_{\mathcal{I}}(x)\Gamma^{-1}R_{\mathcal{I}}(x)\tilde{f}(t,x),\\ &= \tilde{f}^{\mathrm{T}}(t,x)\Gamma^{-1}R^2_{\mathcal{I}}(x)\tilde{f}(t,x) = \tilde{f}^{\mathrm{T}}(t,x)\Gamma^{-1}R_{\mathcal{I}}(x)\tilde{f}(t,x),\end{aligned}$$

where $A^{-\mathrm{T}} := (A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}$ for any nonsingular matrix A. It can be seen that when no constraints are active, i.e. $\mathcal{I}_{\mathrm{act}} = \emptyset$ and $\mathcal{J} = \{\emptyset\}$, the unique optimal solution to subproblems (2.21) and (2.22) is $\mathcal{I}^* = \emptyset$, so that by (2.15), system (2.20) reduces to (2.19).

Remark 2.10. Proposition 2.3.1 as well as Remarks 2.7 and 2.8 apply to subproblems (2.21) and (2.22) with minor changes. Observe that the projection operator has a single parameter, a nonsingular matrix Φ when adopting subproblem (2.21), or a symmetric positive definite matrix Γ when adopting subproblem (2.22). □

The following result shows that the projection operator maintains feasibility of the constraints (2.18) for all future times once the solution of (2.20) enters the feasible region $\tilde{K}$. In other words, if there exists a $t_0 \in \mathbb{R}$ such that $x(t_0) \in \tilde{K}$, then (2.17) holds.

Proposition 2.4.1 (Feasibility Maintenance Property of Projection Operator)**.** *Let $x(t)$ be the solution of system* (2.20)*, $\mathcal{I}^*$ a solution of either subproblem* (2.21) *or* (2.22)*. If there exists a $T \in \mathbb{R}$ such that $\tilde{h}(x(T)) \le 0$, then $\tilde{h}(x(t)) \le 0$ holds for all $t \ge T$.*

Proof. Clearly, $\tilde{h}(x(t)) = [\tilde{h}_1(x(t)), \tilde{h}_2(x(t)), \ldots, \tilde{h}_k(x(t))]^{\mathrm{T}} \le 0$ holds if and only if its elements satisfy $\tilde{h}_i(x(t)) \le 0$ for all $i \in \mathcal{I}_k$. By assumption, $\tilde{h}(x(T)) \le 0$. It is sufficient to show that for all $i \in \mathcal{I}_k$, whenever $\tilde{h}_i(x(t)) = 0$, then $\dot{\tilde{h}}_i(x(t)) \le 0$. Taking the time derivative yields

$$\dot{\tilde{h}}_i(x(t)) = \frac{\partial \tilde{h}_i(x(t))}{\partial x}\dot{x}(t) = \nabla\tilde{h}^{\mathrm{T}}_i(x(t))R_{\mathcal{I}^*}(x(t))\tilde{f}(t,x(t)).$$

If $\tilde{h}_i(x(t)) = 0$, then $i \in \mathcal{I}_{\mathrm{act}}$. Concatenating all gradient vectors with indices in $\mathcal{I}_{\mathrm{act}}$, we need to show that $N^{\mathrm{T}}_{\mathcal{I}_{\mathrm{act}}}(x(t))R_{\mathcal{I}^*}(x(t))\tilde{f}(t,x(t)) \le 0$. This follows immediately from the fact that $\mathcal{I}^*$ is a solution to subproblem (2.21) (or (2.22)), and the second constraint in (2.21) (or (2.22)). ■

Remark 2.11. An equivalent statement of Proposition 2.4.1 is that $\tilde{K}$ is a *positively invariant set* [37, p. 127] for the system (2.20). □

In summary, the projection operator $R_{\mathcal{I}^*}(x)$ is defined by (2.15) and an optimal solution $\mathcal{I}^*$ to the combinatorial optimization subproblem (2.21) (or equivalently, (2.22)). The projection operator applied to an ODE, i.e. (2.20), maintains feasibility of the constraints (2.18) once the solution enters the feasible region. In the next section, we will use the projection operator for anti-windup compensation. Clearly, it can also be used for other purposes, like bounding the parameter estimates of adaptive controllers in some region, as in [122, Sections 4.4, 8.4.2, and 8.5.5]. We leave the exploitation of the projection operator for other applications as future work (see Section 7.1.5).

Remark 2.12. *Projected dynamical systems* (PDS) [107–110] is a significant line of independent research that has attracted economists, physicists, and mathematicians, among others. It is clear that there is a close relationship between system (2.20) and the PDS. Indeed, it will be shown in Section 3.2 that the particular GPAW-compensated system considered is in fact a PDS. We leave the investigation of the link in the general case as future work (see Section 7.1.6). □

2.5 Gradient Projection Anti-windup (GPAW) Scheme

In this section, we construct the *gradient projection anti-windup* (GPAW)-compensated controller using the projection operator developed in Section 2.4. Recall that, classically, windup is interpreted as an inconsistency between the controller state and output, i.e. $\text{sat}(u) \not\equiv u$. The GPAW scheme aims to reduce/eliminate this inconsistency, which can be achieved if the GPAW controller state can be constrained at all times to lie in the unsaturated region

$$K(y,r) = \{\bar{x} \in \mathbb{R}^q \mid \text{sat}(g_c(\bar{x},y,r)) = g_c(\bar{x},y,r)\}. \tag{2.23}$$

Proposition 2.4.1 suggests that application of the projection operator on the nominal controller may achieve this objective.

Recall the nominal controller (1.3)

$$\begin{aligned} \dot{x}_c &= f_c(x_c,y,r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c,y,r). \end{aligned} \tag{2.24}$$

To apply the projection operator on the nominal controller, we make the following identifications between quantities in (2.24) and those in (2.19) and (2.18):

$$\begin{gathered} x \sim x_c, \qquad x_{ig} \sim x_{c0}, \qquad \tilde{f}(t,x) \sim f_c(x_c,y,r), \\ \tilde{h}(x) \sim \begin{bmatrix} g_c(x_c,y,r) - u_{\max} \\ -g_c(x_c,y,r) + u_{\min} \end{bmatrix}, \end{gathered} \tag{2.25}$$

where $u_{\max}$ and $u_{\min}$ are the vectors of saturation limits defined by the vector saturation function (1.2). Notice that the identification of $\tilde{f}$ with f_c is motivated by the need to recover nominal performance when no saturation constraints are active, and that (2.25) implies $\tilde{K} = \{\bar{x} \in \mathbb{R}^q \mid \tilde{h}(\bar{x}) \le 0\} \sim K(y,r)$. The *GPAW-compensated controller*[8] (see (2.20) and (2.19)) will then be described by

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g,y,r) f_c(x_g,y,r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g,y,r), \end{aligned}$$

where (x_g, u_g) are the state and output respectively, and $R_{\mathcal{I}^*}(x_g,y,r)$ remains to be defined.

[8] See Section 1.2 for a discussion on the distinction between the *anti-windup compensator* and the *anti-windup compensated controller*. See also Section 1.3.1 on how to recover an anti-windup compensator from an anti-windup compensated controller.

Analogous to the conditional integration method, the GPAW scheme only modifies the controller state x_g through its evolution $\dot{x}_g$, and the output equation remains unaltered with g_c defined by the nominal controller (2.24) (see Remark 2.2).

If we were to carry through the steps[9] needed to define $R_{\mathcal{I}^*}(x_g, y, r)$, the GPAW-compensated controller thus obtained (i.e. by applying the projection operator on nominal controllers of the form (2.24)) can achieve controller state-output consistency only in an approximate sense, as corroborated by numerical results in [127]. This is due to the structure of the output equation $u_g = g_c(x_g, y, r)$ having u_g dependent on measurement y and reference input r. For simplicity in the present discussion, restrict consideration to single-output controllers, i.e. u_g is scalar. Under saturation, i.e. $u_g = u_{\max}$ or $u_g = u_{\min}$, controller state-output consistency requires $\dot{u}_g = 0$ when the nominal update will aggravate the existing saturation constraint. The projection operator modifies x_g through $\dot{x}_g$, which, under these conditions, can be shown[10] to induce $\frac{\partial g_c(x_g,y,r)}{\partial x_c}\dot{x}_g = 0$. Hence

$$\dot{u}_g = \frac{\partial g_c(x_g, y, r)}{\partial x_c}\dot{x}_g + \frac{\partial g_c(x_g, y, r)}{\partial y}\dot{y} + \frac{\partial g_c(x_g, y, r)}{\partial r}\dot{r} = \frac{\partial g_c(x_g, y, r)}{\partial y}\dot{y} + \frac{\partial g_c(x_g, y, r)}{\partial r}\dot{r},$$

so that in general, $\dot{u}_g$ will be non-zero for arbitrary $(\dot{y}, \dot{r})$. It can also be surmised that the GPAW scheme will be ineffective when

$$\left\|\frac{\partial g_c(x_c, y, r)}{\partial x_c}\dot{x}_c\right\| \ll \left\|\frac{\partial g_c(x_c, y, r)}{\partial y}\dot{y} + \frac{\partial g_c(x_c, y, r)}{\partial r}\dot{r}\right\|,$$

holds for the uncompensated controller on the saturation constraint boundaries, since the projection operator only acts on the controller state.

Remark 2.13. Observe also that the proof of Proposition 2.4.1 breaks down when $\tilde{h}$ does not depend on x alone. This would be analogous to the case when g_c does not depend on x_g alone (for the GPAW-compensated controller). □

The preceding discussion motivates us to restrict consideration to "strictly proper"[11] nominal controllers of the form

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c), \end{aligned} \tag{2.26}$$

where g_c *depends only on the controller state* x_c. Here, as in (2.24), $x_c, x_{c0} \in \mathbb{R}^q$ are the controller state and initial state, $u_c \in \mathbb{R}^m$ is the controller output, $y \in \mathbb{R}^p$ is the measurement, and $r := r(t) \in \mathbb{R}^{n_r}$ is the instantaneous reference input.

Remark 2.14. For nominal controllers of the form (2.24), it is shown in Section 2.6 that an arbitrarily close approximate controller that has the required structure of (2.26) can be constructed. □

As before, we apply the projection operator of Section 2.4 on the nominal controller (2.26) by making the following identifications with quantities in (2.19) and (2.18):

$$x \sim x_c, \qquad x_{ig} \sim x_{c0}, \qquad \tilde{f}(t, x) \sim f_c(x_c, y, r), \qquad \tilde{h}(x) \sim \begin{bmatrix} g_c(x_c) - u_{\max} \\ -g_c(x_c) + u_{\min} \end{bmatrix}.$$

[9] The construction of $R_{\mathcal{I}^*}(x_g, y, r)$ will be analogous to the construction of $R_{\mathcal{I}^*}(x)$ in Section 2.4.

[10] See the proof of Proposition 2.4.1.

[11] For LTI nominal controllers, observe that this restricts the nominal controllers to have strictly proper transfer functions.

In contrast to (2.25), observe that a crucial difference here is that the last vector defined by g_c (being dependent solely on x_c) maps directly to $\tilde{h}$ (which depends solely on x).

The resulting GPAW-compensated controller then has the form (see (2.20) and (2.19))

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g), \end{aligned} \tag{2.27}$$

where g_c is defined by the nominal controller (2.26) and $R_{\mathcal{I}^*}(x_g, y, r)$ is to be defined. Observe that apart from the definition of an independent state x_g, the only difference with (2.26) is the introduction of the projection operator $R_{\mathcal{I}^*}(x_g, y, r)$, which is a state-and-input dependent $q \times q$ matrix.

Next, we describe the construction of the projection operator, which is largely analogous to the constructions in Sections 2.3, 2.3.1, and 2.4, made explicit here for clarity. Let g_c in (2.27) be decomposed into its m elements, $g_c = [g_{c1}, g_{c2}, \ldots, g_{cm}]^{\mathrm{T}}$, and define the $2m$ saturation constraint functions h_i by (see (1.2))

$$\begin{aligned} h_i(x_g) &= g_{ci}(x_g) - u_{\max,i}, \\ h_{i+m}(x_g) &= -g_{ci}(x_g) + u_{\min,i}, \end{aligned} \qquad \forall i \in \mathcal{I}_m := \{1, 2, \ldots, m\}. \tag{2.28}$$

Observe that $h_i(x_g) \leq 0$ for all $i \in \mathcal{I}_m$ imply $\operatorname{sat}(g_c(x_g)) = g_c(x_g)$. Assuming differentiability of g_c, these constraint functions have gradients

$$\nabla h_i(x_g) = \nabla g_{ci}(x_g), \qquad \nabla h_{i+m}(x_g) = -\nabla g_{ci}(x_g) = -\nabla h_i(x_g), \qquad \forall i \in \mathcal{I}_m.$$

For any index set $\mathcal{I} \subset \mathcal{I}_{2m}$, define the $q \times \max\{|\mathcal{I}|, 1\}$ matrix

$$N_{\mathcal{I}}(x_g) = \begin{cases} [\nabla h_{\sigma_{\mathcal{I}}(1)}(x_g), \nabla h_{\sigma_{\mathcal{I}}(2)}(x_g), \ldots, \nabla h_{\sigma_{\mathcal{I}}(|\mathcal{I}|)}(x_g)], & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases} \tag{2.29}$$

where $\sigma_{\mathcal{I}} \colon \{1, 2, \ldots, |\mathcal{I}|\} \to \mathcal{I}$ is a chosen (non-unique) bijection (described in Remark 2.5) that assigns an integer in $\mathcal{I}$ to each integer in $\{1, 2, \ldots, |\mathcal{I}|\}$. For any $\mathcal{I} \subset \mathcal{I}_{2m}$ such that $\operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$, define the projection matrix $R_{\mathcal{I}} \colon \mathbb{R}^q \to \mathbb{R}^{q \times q}$

$$R_{\mathcal{I}}(x_g) = \begin{cases} I - \Gamma N_{\mathcal{I}} (N_{\mathcal{I}}^{\mathrm{T}} \Gamma N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise}, \end{cases} \tag{2.30}$$

where $\Gamma \in \mathbb{R}^{q \times q}$ is the *single* GPAW parameter, chosen to be symmetric positive definite.

Remark 2.15. An attractive feature of the GPAW scheme is that it has only a *single* symmetric positive definite matrix parameter, in contrast to the typical anti-windup scheme which introduced 6 matrix parameters (see Section 1.2). Moreover, observe that multiplying a non-zero scalar to Γ does not change $R_{\mathcal{I}}(x_g)$, so that we can always normalize Γ by such scalar multiplications (nominally a positive scalar to preserve positive definiteness). □

Define the index set of active saturation constraints, and the candidate solution set

$$\mathcal{I}_{\text{sat}} := \mathcal{I}_{\text{sat}}(x_g) = \{i \in \mathcal{I}_{2m} \mid h_i(x_g) \geq 0\}, \qquad \mathcal{J} := \{\mathcal{I} \subset \mathcal{I}_{\text{sat}} \mid |\mathcal{I}| \leq q\}.$$

With (x_g, y, r) fixed, the analogue of subproblem (2.22) becomes

$$\begin{aligned} \max_{\mathcal{I}\in\mathcal{J}} F(\mathcal{I}) = f_c^{\mathrm{T}}(x_g,y,r)\Gamma^{-1}R_{\mathcal{I}}(x_g)f_c(x_g,y,r), \\ \text{subject to} \qquad \operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|, \\ N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}(x_g)R_{\mathcal{I}}(x_g)f_c(x_g,y,r) \le 0. \end{aligned} \tag{2.31}$$

The following result asserts existence of solutions to subproblem (2.31), and is analogous to Proposition 2.3.1. While their proofs are similar, we prove it explicitly because the GPAW scheme is the main theme of this dissertation, and to avoid any potential ambiguities caused by notational differences.

Proposition 2.5.1 (Existence of Solutions to Combinatorial Optimization Subproblem)**.** *For any fixed $(x_g, y, r) \in \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^{n_r}$, there exists a solution to subproblem* (2.31)*.*

Proof. It is sufficient to show that there always exists a feasible (not necessarily optimal) solution to subproblem (2.31). If $v := \operatorname{rank}(N_{\mathcal{I}_{\mathrm{sat}}}(x_g)) = 0$ (which includes the case $\mathcal{I}_{\mathrm{sat}} = \emptyset$), then $N_{\mathcal{I}_{\mathrm{sat}}}(x_g)$ must be a $q \times \max\{|\mathcal{I}_{\mathrm{sat}}|, 1\}$ zero matrix. In this case, it can be verified that $\mathcal{I} = \emptyset$ is a feasible solution (in fact, the only feasible solution) to subproblem (2.31).

If $v > 0$ (necessarily, $v \le \min\{|\mathcal{I}_{\mathrm{sat}}|, q\}$), then $N_{\mathcal{I}_{\mathrm{sat}}}(x_g)$ has exactly v *linearly independent* columns, so that there exists $\mathcal{I} \subset \mathcal{I}_{\mathrm{sat}}$ such that $\operatorname{rank}(N_{\mathcal{I}}(x_g)) = v = |\mathcal{I}|$, satisfying the first constraint of subproblem (2.31). Any column of $N_{\mathcal{I}_{\mathrm{sat}}}(x_g)$ can then be expressed as a linear combination of the columns of $N_{\mathcal{I}}(x_g)$, so that $N_{\mathcal{I}_{\mathrm{sat}}}(x_g) = N_{\mathcal{I}}(x_g)\Psi$ for some $\Psi \in \mathbb{R}^{v\times|\mathcal{I}_{\mathrm{sat}}|}$. From (2.30), we have $N_{\mathcal{I}}^{\mathrm{T}}(x_g)R_{\mathcal{I}}(x_g) = 0$. Since $N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}(x_g)R_{\mathcal{I}}(x_g) = \Psi^{\mathrm{T}}N_{\mathcal{I}}^{\mathrm{T}}(x_g)R_{\mathcal{I}}(x_g) = 0$, the second constraint of subproblem (2.31) holds, which shows $\mathcal{I}$ to be a feasible solution. ■

Remark 2.16. Observe that Remarks 2.7 and 2.8 apply to subproblem (2.31) with minor changes. Specifically, for a finite dimensional control output $u_g \in \mathbb{R}^m$, $m < \infty$, there can only be a finite number of candidate solutions, so that an optimal solution to subproblem (2.31) can always be found by an exhaustive search algorithm. See Section 4.1, Appendices A and B for more properties of the combinatorial optimization subproblem (2.31), and alternative solution methods. □

The next result[12] states a property of the objective function of subproblem (2.31) which follows immediately from [100, equation (3.20)]. It is useful when deriving closed-form expressions for the GPAW-compensated controller, as demonstrated in Appendices A and B.

Proposition 2.5.2. *If some index set $\tilde{\mathcal{I}}_1 \subset \mathcal{I}_{\mathrm{sat}}$ is such that* $\operatorname{rank}(N_{\tilde{\mathcal{I}}_1}(x_g)) = |\tilde{\mathcal{I}}_1|$*, then the objective function of subproblem* (2.31) *satisfies $F(\tilde{\mathcal{I}}_2) \ge F(\tilde{\mathcal{I}}_1)$ for any $\tilde{\mathcal{I}}_2 \subset \tilde{\mathcal{I}}_1$.*

Proof. Since $\tilde{\mathcal{I}}_2 \subset \tilde{\mathcal{I}}_1$, the rank condition $\operatorname{rank}(N_{\tilde{\mathcal{I}}_1}(x_g)) = |\tilde{\mathcal{I}}_1|$ imply $\operatorname{rank}(N_{\tilde{\mathcal{I}}_2}(x_g)) = |\tilde{\mathcal{I}}_2|$, and ensures that the projection matrices $R_{\tilde{\mathcal{I}}_1}(x_g)$ and $R_{\tilde{\mathcal{I}}_2}(x_g)$ (and hence $F(\tilde{\mathcal{I}}_1)$ and $F(\tilde{\mathcal{I}}_2)$) are well defined. If $\tilde{\mathcal{I}}_1 = \tilde{\mathcal{I}}_2$, then $F(\tilde{\mathcal{I}}_1) = F(\tilde{\mathcal{I}}_2)$ and the conclusion holds trivially. Assume $\tilde{\mathcal{I}}_2$ is a strict subset of $\tilde{\mathcal{I}}_1$. Since the GPAW parameter $\Gamma \in \mathbb{R}^{q\times q}$ is positive definite, it can always be decomposed as $\Gamma = \Phi\Phi^{\mathrm{T}}$ for some nonsingular $\Phi \in \mathbb{R}^{q\times q}$ [124, Theorem 7.2.7,

[12] In Proposition 2.5.2, we use accented symbols $\tilde{\mathcal{I}}_1$, $\tilde{\mathcal{I}}_2$, to avoid confusion with $\mathcal{I}_i := \{1, 2, \ldots, i\}$.

p. 406]. For any $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$ such that $\text{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$, define[13]

$$P_{\mathcal{I}}(x_g) := \Phi^{-1} R_{\mathcal{I}}(x_g) \Phi = \begin{cases} I - \tilde{N}_{\mathcal{I}} (\tilde{N}_{\mathcal{I}}^{\text{T}} \tilde{N}_{\mathcal{I}})^{-1} \tilde{N}_{\mathcal{I}}^{\text{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise,} \end{cases}$$

where $\tilde{N}_{\mathcal{I}}(x_g) := \Phi^{\text{T}} N_{\mathcal{I}}(x_g)$. From its preceding definition, the matrix $P_{\mathcal{I}}(x_g)$ is a projection matrix [100, Theorem 1], and hence satisfies[14] [100, equation (3.20)]

$$\|P_{\mathcal{I} \cup \{j\}}(x_g) z\| \leq \|P_{\mathcal{I}}(x_g) z\| \leq \|z\|, \qquad \forall \mathcal{I} \subset \mathcal{I}_{\text{sat}}, \forall j \in \mathcal{I}_{\text{sat}} \setminus \mathcal{I}, \forall z \in \mathbb{R}^q. \tag{2.32}$$

Since $\mathcal{I}_{\text{sat}}$ is a finite set (due to $m < \infty$), the set difference $\tilde{\mathcal{I}}_1 \setminus \tilde{\mathcal{I}}_2$ is also finite. Let $\tilde{\mathcal{I}}_1 \setminus \tilde{\mathcal{I}}_2 = \{i_1, i_2, \ldots, i_k\}$, where i_j for $j \in \{1, 2, \ldots, k\}$ ($k = |\tilde{\mathcal{I}}_1| - |\tilde{\mathcal{I}}_2|$) are its elements. Then $\tilde{\mathcal{I}}_1 = \tilde{\mathcal{I}}_2 \cup \{i_1\} \cup \{i_2\} \cup \cdots \cup \{i_k\}$. From (2.32), we have

$$\|P_{\tilde{\mathcal{I}}_1}(x_g) z\| = \|P_{\tilde{\mathcal{I}}_2 \cup \{i_1\} \cup \cdots \cup \{i_k\}}(x_g) z\| \leq \cdots \leq \|P_{\tilde{\mathcal{I}}_2 \cup \{i_1\}}(x_g) z\| \leq \|P_{\tilde{\mathcal{I}}_2}(x_g) z\|, \qquad \forall z \in \mathbb{R}^q.$$

The conclusion $F(\tilde{\mathcal{I}}_2) \geq F(\tilde{\mathcal{I}}_1)$ follows from the preceding by observing that

$$F(\mathcal{I}) = f_c^{\text{T}} \Gamma^{-1} R_{\mathcal{I}} f_c = f_c^{\text{T}} \Phi^{-\text{T}} P_{\mathcal{I}} \Phi^{-1} f_c = f_c^{\text{T}} \Phi^{-\text{T}} P_{\mathcal{I}}^{\text{T}} P_{\mathcal{I}} \Phi^{-1} f_c = \|P_{\mathcal{I}} \Phi^{-1} f_c\|^2 = \|P_{\mathcal{I}} z\|^2,$$

where $z := \Phi^{-1} f_c$, and the function arguments have been dropped. ■

Remark 2.17. Proposition 2.5.2 implies that $\mathcal{I} = \emptyset$ is an optimal solution to subproblem (2.31) whenever it is feasible. Since $\text{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$ always holds for $\mathcal{I} = \emptyset$, it is an optimal solution whenever $N_{\mathcal{I}_{\text{sat}}}^{\text{T}}(x_g) f_c(x_g, y, r) \leq 0$. In particular, this happens when no constraints are active, i.e. $\mathcal{I}_{\text{sat}} = \emptyset$, or when $\text{rank}(N_{\mathcal{I}_{\text{sat}}}) = 0$, so that $N_{\mathcal{I}_{\text{sat}}} = 0$. Observe that when $\mathcal{I}^* = \emptyset$, the nominal update is recovered (see (2.27) and (2.30)). □

At each fixed time, so that (x_g, y, r) is fixed, the GPAW-compensated controller is defined by (2.27) with $R_{\mathcal{I}^*}(x_g, y, r)$ defined by (2.30) and a solution $\mathcal{I}^*$ to subproblem (2.31). Note that $\mathcal{I}^*$ depends on (x_g, y, r). As mentioned previously, the GPAW scheme aims to achieve *controller state-output consistency*, which is established in the following theorem. To the best of our knowledge, it is a unique property among anti-windup schemes.[15] This result is analogous to Proposition 2.4.1, proven explicitly to avoid ambiguities.

Theorem 2.5.3 (Controller State-output Consistency)**.** *Consider the GPAW-compensated controller defined by* (2.27), (2.30)*, and a solution $\mathcal{I}^*$ to subproblem* (2.31)*. If there exists a $T \in \mathbb{R}$ such that* $\text{sat}(u_g(T)) = u_g(T)$*, then* $\text{sat}(u_g(t)) = u_g(t)$ *holds for all $t \geq T$.*

Proof. From (1.2), $u_g = g_c(x_g)$ (see (2.27)), and (2.28), it can be seen that $\text{sat}(u_g(t)) = u_g(t)$ holds if and only if $h_i(x_g(t)) \leq 0$ for all $i \in \mathcal{I}_{2m}$. By assumption, we have $\text{sat}(u_g(T)) = u_g(T)$, so that $h_i(x_g(T)) \leq 0$. It is sufficient to show that for all $i \in \mathcal{I}_{2m}$, whenever $h_i(x_g(t)) = 0$, then $\dot{h}_i(x_g(t)) \leq 0$. Taking the time derivative and using (2.27) yields

$$\dot{h}_i(x_g(t)) = \frac{\partial h_i(x_g(t))}{\partial x_g} \dot{x}_g(t) = \nabla h_i^{\text{T}}(x_g(t)) R_{\mathcal{I}^*} f_c(t, x_g(t)),$$

[13] This is analogous to $\tilde{P}_{\mathcal{I}}(x)$ in (2.15).

[14] This is written using our notation rather than that in [100].

[15] Clearly, specializations of the GPAW scheme (like the conditional integration method) also possess this property for the same class of nominal controllers. However, note that it must be a true specialization of the GPAW scheme. Not all variants of the conditional integration method (e.g. see [17, Section 3.3.2, p. 38]) possess this property.

where $R_{\mathcal{I}^*} f_c(t, x_g(t)) := R_{\mathcal{I}^*}(x_g(t), y(t), r(t)) f_c(x_g(t), y(t), r(t))$. If $h_i(x_g(t)) = 0$, then $i \in \mathcal{I}_{\text{sat}}$. Concatenating all gradient vectors with indices in $\mathcal{I}_{\text{sat}}$, we need to show that

$$N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}}(x_g(t)) R_{\mathcal{I}^*} f_c(t, x_g(t)) \leq 0.$$

This follows immediately from the fact that $\mathcal{I}^*$ is a solution to subproblem (2.31), and the second constraint in (2.31). ■

Remark 2.18. An equivalent statement of Theorem 2.5.3 is that the unsaturated region

$$K = \{\bar{x} \in \mathbb{R}^q \mid \operatorname{sat}(g_c(\bar{x})) = g_c(\bar{x})\} = \{\bar{x} \in \mathbb{R}^q \mid h_i(\bar{x}) \leq 0, \forall i \in \mathcal{I}_{2m}\}, \tag{2.33}$$

is a *positively invariant set* [37, p. 127] for the solution of the GPAW-compensated controller (2.27). Notice that the second equivalent form written in terms of h_i follows from the observation that $\operatorname{sat}(g_c(x_g)) = g_c(x_g)$ holds if and only if $h_i(x_g) \leq 0$ for all $i \in \mathcal{I}_{2m}$ (see (2.28) and (1.2)). □

Remark 2.19. Observe that usually, the controller state can be arbitrarily initialized. If it is initialized such that $\operatorname{sat}(g_c(x_{c0})) = g_c(x_{c0})$, Theorem 2.5.3 shows that controller state-output consistency is achieved for all $t \geq 0$. Even when the nominal controller's state cannot be arbitrarily initialized, the same can be achieved by initializing the *anti-windup compensator's* state x_{aw0} appropriately (see (1.5) and (1.6)). □

Remark 2.20. It is clear that we can always impose more *controller state* constraints $h_j(x_g) \leq 0$, $j \in \{2m+1, \ldots, 2m+k\}$. The proof of Theorem 2.5.3 can be readily modified to show that these additional constraints will be enforced exactly for all future times once they are satisfied at any time instant. If we impose more general constraints of the form $h_j(x_g, x, r) \leq 0$, with (x, r) the plant state and reference input respectively, these constraints will be satisfied approximately, i.e. not enforced exactly. It can be shown that even with the introduction of these constraints, those that depend only on the controller state will be enforced exactly by the projection operator. These observations suggest that the projection operator may be used for general constrained control [6], with ability to enforce a subset of constraints exactly, and others approximately. Indeed, we will demonstrate this by way of example in Section 2.8.3. □

Theorem 2.5.3 established the controller state-output consistency property of GPAW-compensated controllers when the nominal controller is of the form (2.26), i.e. with g_c depending only on x_c. This is a significant result because:

- most anti-windup schemes [12, 14] aim to achieve controller state-output consistency (see Remark 2.21). To the best of our knowledge, this has been achieved only in an approximate sense to date, whereas the GPAW scheme achieved it *exactly*;
- it applies to a large class of *nonlinear* controllers, and is achieved irrespective of the plant dynamics.[16]

Remark 2.21. Typically, when a controller is driven by some error signal, the control objective would be achieved *exactly* if the error signal is identically zero. In most anti-windup schemes, whether the early ad hoc schemes [12] or modern schemes [14], the anti-windup compensator is driven by the signal $(\operatorname{sat}(u) - u)$ (see Fig. 1-1). When controller state-output

[16] In particular, being independent of plant dynamics implies that GPAW compensation can be applied to *adaptive controllers* [122] [38, Chapter 8, pp. 311 – 389] without much additional complications.

consistency is achieved, i.e. $\mathrm{sat}(u) \equiv u$, this signal would be identically zero. Hence, controller state-output consistency is an implicit objective in anti-windup compensators driven by the signal $(\mathrm{sat}(u) - u)$. □

One implication of controller state-output consistency is that it allows the saturation function to be *eliminated* in the description of the closed-loop system when the controller state is appropriately initialized. The closed-loop system comprising the plant (1.1)

$$\begin{aligned} \dot{x} &= f(x, \mathrm{sat}(u)), \qquad x(0) = x_0, \\ y &= g(x, \mathrm{sat}(u)), \end{aligned} \tag{2.34}$$

and GPAW-compensated controller (2.27) (with $u := u_g$) can be written as

$$\begin{aligned} \dot{x} &= f(x, \mathrm{sat}(g_c(x_g))), && x(0) = x_0, \\ \dot{x}_g &= R_{\mathcal{I}^*} f_c(x_g, g(x, \mathrm{sat}(g_c(x_g))), r(t)), && x_g(0) = x_{c0}, \end{aligned} \tag{2.35}$$

where $R_{\mathcal{I}^*} f_c(x_g, y, r) := R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r)$, and we have used $r(t)$ to emphasize the time-varying nature of the system (see Remark 1.2). When the controller state is initialized such that $\mathrm{sat}(g_c(x_{c0})) = g_c(x_{c0})$ (see Remark 2.19), the closed-loop system reduces by Theorem 2.5.3 to

$$\begin{aligned} \dot{x} &= f(x, g_c(x_g)), && x(0) = x_0, \\ \dot{x}_g &= R_{\mathcal{I}^*} f_c(x_g, g(x, g_c(x_g)), r(t)), && x_g(0) = x_{c0}, \end{aligned} \tag{2.36}$$

for all $t \geq 0$, where the $\mathrm{sat}(\cdot)$ function has been *eliminated*. This shows that all complications due to saturation are accounted for by the projection operator $R_{\mathcal{I}^*}(x_g, y, r)$. The elimination of the saturation function in the system description provides significant simplifications when deriving stability results, examples of which are shown in Sections 4.4 and 5.2.

The effect of applying GPAW compensation is illustrated in Fig. 2-3, which is similar to Fig. 2-1. Here, K is the unsaturated region in the controller state space, $x_{gi} := x_g(t_i)$, and $f_{ci} := f_c(x_g(t_i), y(t_i), r(t_i))$ represents the nominal controller's vector field at time t_i. At each time instant, the combinatorial optimization subproblem (2.31) determines an optimal combination of constraints to project f_{ci}, yielding the vector $\tilde{f}_{gi} := R_{\mathcal{I}^*}(x_g(t_i), y(t_i), r(t_i)) f_c(x_g(t_i), y(t_i), r(t_i))$ that defines the GPAW-compensated controller. Observe that the unsaturated region K (2.33) does not depend on inputs (y, r), and the controller state can be constrained within it exactly. For nominal controllers of the form (2.24), the resulting unsaturated region $K(y, r)$ (2.23) varies with inputs (y, r). In this case, variations of $K(y, r)$ may cause the controller state to depart it even under projection.

We end this section with some observations:

(i) apart from the restriction on the form of the output function g_c of the nominal controller, the only assumption needed to construct the projection operator (and hence the GPAW-compensated controller) is differentiability of g_c;

(ii) by construction (see Remark 2.17), the GPAW-compensated controller satisfies condition (ii) of Problem 1 (see Section 1.3);

(iii) as shown in Section 3.1 and Appendix A, application of the GPAW scheme to some nominal controllers yields the conditional integration compensated controller. This shows that indeed, the GPAW scheme is a generalization of the conditional integration method.

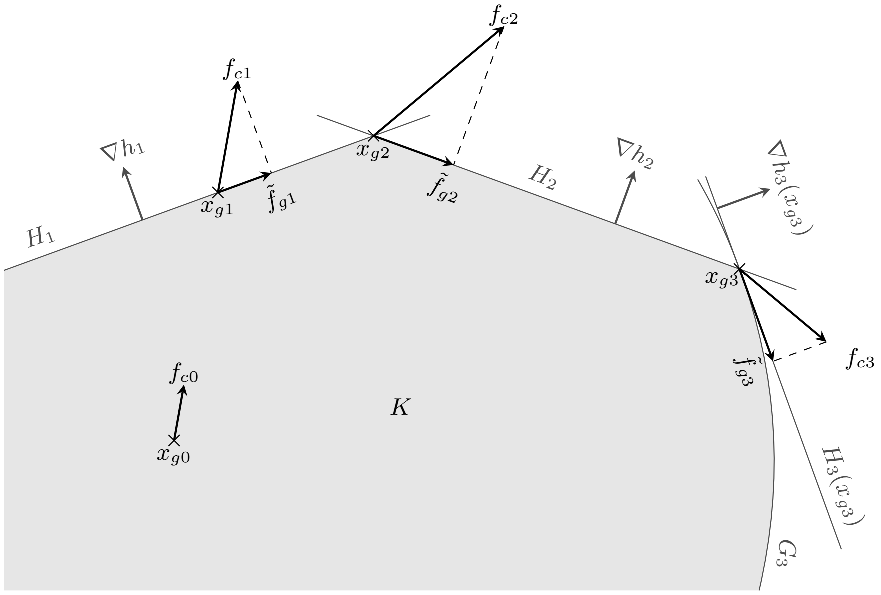


Figure 2-3: Visualization of the GPAW scheme. K is the unsaturated region, bounded by hypersurfaces H_1, H_2, and G_3. Here, $x_{gi} := x_g(t_i)$, the supporting hyperplane of G_3 at x_{g3} is $H_3(x_{g3})$, and the projection of $f_{ci} := f_c(x_g(t_i), y(t_i), r(t_i))$ onto H_i yields $\tilde{f}_{gi} := R_{\mathcal{I}^*}(x_g(t_i), y(t_i), r(t_i)) f_c(x_g(t_i), y(t_i), r(t_i))$. Observe that at each time instant, the combinatorial optimization subproblem (2.31) determines an optimal combination of active constraints to project onto, in particular at the points x_{g2} and x_{g3}.

2.6 Approximate Nominal Controller

In Section 2.5, restrictions to nominal controllers of the form (2.26) were made to achieve controller state-output consistency. Here, we show that for nominal controllers of general structure (2.24), an arbitrarily close approximating controller can be constructed that has the required structure of (2.26), i.e. with output equation depending only on its state. Then GPAW compensation can be applied to this approximate controller, yielding the same desirable properties. Note that this construction is not unique, and similar ideas have been discussed in [128, Remark 9].

The main idea is to replace the signal components in the controller output equation that are not part of the controller state by its low-pass filtered signal, and design the low-pass filter such that its bandwidth is much larger than the effective bandwidth of the closed-loop system. It is clear that the approximation will be enhanced as the bandwidth of the low-pass filter is increased. Importantly, the main purpose of this low-pass filter is *not* for noise rejection or performance/robustness enhancements.

Consider the nominal controller

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c, y), \end{aligned} \tag{2.37}$$

whose output equation depends not only on the state, but on measurement y as well. For simplicity, we have assumed that the output equation is not dependent on the reference input r. If it indeed does, the treatment is similar.

Remark 2.22. When g_c depends on the measurement y as in (2.37), the closed-loop system comprising the plant (2.34) and controller (2.37) with $u := u_c$ will contain an *algebraic loop* whenever $\frac{\partial g}{\partial u}\frac{\partial g_c}{\partial y} \not\equiv 0$. □

Consider augmenting the controller state to be $\tilde{x}_c := (x_c, \tilde{y})$, with $\tilde{y} = y$. Then, by replacing y with $\tilde{y}$ in the controller output equation, we have $u_c = g_c(x_c, \tilde{y}) = g_c(\tilde{x}_c)$, which is the desired form in (2.26). The state equation of the augmented controller with state $\tilde{x}_c$ needs to satisfy

$$\dot{x}_c = f_c(x_c, y, r), \qquad \dot{\tilde{y}} = \dot{y}. \tag{2.38}$$

Clearly, if the functions f and g in (2.34) are known *exactly*, realization of (2.38) is straightforward,[17] by taking the time derivative of y in (2.34) and using the knowledge of f and g. We avoid making such a conservative assumption by using an *approximation*.

Consider $\tilde{y}$ obtained as the output of an *exponentially stable, unity DC gain* low-pass filter with input y, parameterized by $a \in (0, \infty)$

$$\dot{\tilde{y}} = a(y - \tilde{y}), \qquad \tilde{y}(0) = y(0).$$

It can be seen that $\tilde{y}(t) \to y(t)$ for all $t \geq 0$ as $a \to \infty$, so that the solution of the approximating controller

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), & x_c(0) &= x_{c0}, \\ \dot{\tilde{y}} &= a(y - \tilde{y}), & \tilde{y}(0) &= y(0), \\ u_c &= g_c(x_c, \tilde{y}), \end{aligned} \tag{2.39}$$

can be made arbitrarily close to the nominal controller (2.37). While this can be shown formally for any fixed $y\colon [0,\infty) \to \mathbb{R}^p$ and $r\colon [0,\infty) \to \mathbb{R}^{n_r}$ using singular perturbation theory [37, Chapter 11, pp. 423 – 459], the larger question is the effect of the approximation on the *closed-loop system*, which we discuss next.

The closed-loop system described by the feedback interconnection of the plant (2.34) and approximate controller (2.39) with $u := u_c$ is described by

$$\begin{aligned} \dot{x} &= f(x, \mathrm{sat}(g_c(x_c, \tilde{y}))), \\ \dot{x}_c &= f_c(x_c, g(x, \mathrm{sat}(g_c(x_c, \tilde{y}))), r), \\ \epsilon\dot{\tilde{y}} &= g(x, \mathrm{sat}(g_c(x_c, \tilde{y}))) - \tilde{y}, \end{aligned} \tag{2.40}$$

where $\epsilon := \frac{1}{a}$. Observe that when $\epsilon = 0$, we recover the *exact* closed-loop system obtained with controller (2.37), which corresponds to the reduced system in the singular perturbation framework. System (2.40) is referred as the approximate system when $\epsilon > 0$, and the exact system when $\epsilon = 0$. When we assume existence and uniqueness of solutions[18] to the *exact* system, then (2.40) is a standard singular perturbation model [37, p. 424]. It can be shown

[17] When the closed-loop system contains an algebraic loop, there are additional difficulties on well-posedness of the feedback interconnection.

[18] Recall that in the anti-windup context, the nominal controller has been designed to achieve some desired performance. Existence and uniqueness of solutions to the closed-loop system is usually guaranteed even when not explicitly sought in the control design.

that if g and g_c are such that the eigenvalue condition[19] [37, p. 433]

$$\mathrm{Re}\left(\lambda\left(\frac{\partial g}{\partial u}\frac{\partial g_c}{\partial y}(x, x_c) - I\right)\right) < 0,$$

holds uniformly for all (x, x_c) in some domain, then the origin of the associated boundary layer model for the singular perturbation model (2.40) is exponentially stable. With this, and assuming existence and uniqueness of solutions of the exact system, [37, Theorem 11.1, p. 434] shows that on any finite time interval, the solution of the approximate system can be made arbitrarily close to the solution of the exact system when ϵ is sufficiently small (or equivalently, a is sufficiently large). When the equilibrium of the *exact* system is *exponentially stable*, [37, Theorem 11.2, pp. 439 – 440] shows that the result extends to infinite intervals.

Observe that for input-constrained LTI systems driven by LTI controllers, local exponential stability is usually guaranteed, so that the infinite time approximation result holds. If the exact system is not exponentially stable and the finite time approximation result indicated above is not sufficient, repeating the analysis with the approximate controller may be required. Because the approximation can be made arbitrarily well, it is likely that the approximate controller will be able to achieve the control objectives as well.

Remark 2.23. The approximate controller (2.39) requires the augmentation of the controller state. We note that the $(q + m)$-th order controller with state (x_c, u_c) and output u_c,

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), & x_c(0) &= x_{c0}, \\ \dot{u}_c &= \frac{\partial g_c(x_c)}{\partial x_c} f_c(x_c, y, r), & u_c(0) &= g_c(x_{c0}), \end{aligned}$$

is a non-minimal (equivalent) realization of the nominal controller (2.26). For the preceding augmented controller, it can be shown that GPAW compensation with parameter $\Gamma = I \in \mathbb{R}^{(q+m)\times(q+m)}$ yields effectively *no anti-windup compensation.* This suggests that controller state augmentation should always be done with caution, and using a minimal realization [50] would likely be more appropriate. We leave the study of the implications of controller state augmentation as future work (see Section 7.1.10). □

In summary, for controllers of general structure (2.24), an arbitrarily close approximating controller can be constructed that has the form of (2.26), where the output equation depends only on the controller state. Then GPAW compensation can be applied to the approximate controller yielding controller state-output consistency (see Theorem 2.5.3). The approximate controller constructed in this section will be of higher order than the exact nominal controller. In this sense, application of the GPAW scheme on the approximate controller can be seen to be analogous to the case of employing *dynamic* anti-windup compensators, where additional states are employed. Sections 2.8.1, 6.1.2, and 6.2.3 illustrate how the construction presented can be applied in modified form.

Other than the basic construction presented here and Section 2.5, alternative ways to realize GPAW-compensated controllers are shown in Section 4.1 as well as Appendices A and B. For ease of reference, we summarize the procedure to apply GPAW compensation in Appendix C.

[19] Here, $\mathrm{Re}(\lambda(A))$ means real part of all eigenvalues of matrix A, and $\frac{\partial g}{\partial u}\frac{\partial g_c}{\partial y}(x, x_c)$ is shorthand for $\frac{\partial g}{\partial u}(x, \mathrm{sat}(g_c(x_c, \bar{h}(x, x_c))))\frac{\partial g_c}{\partial y}(x_c, \bar{h}(x, x_c))$ where $\tilde{y} = \bar{h}(x, x_c)$ is an isolated real root of the equation $g(x, \mathrm{sat}(g_c(x_c, \tilde{y}))) - \tilde{y} = 0$.

2.7 Passivity and L_2-Gain of Projection Operator

In this section, we define the *principal* and *complementary projection operators* induced by the projection operator $R_{\mathcal{I}^*}(x_g, y, r)$ in (2.27), and show them to be *passive* and with L_2-gain less than one. These results can be used in conjunction with numerous passivity based results (e.g. [37, Chapter 6, pp. 227 – 259], [40, 41, 129]) to establish stability of GPAW-compensated systems. First, we recall the definitions of passivity, sector bounds, and L_2-gain.

Definition 2.1 (Passive Memoryless Systems [37, Definition 6.1, p. 231])**.** The memoryless nonlinear system $y = h(t, u)$ is said to be

(i) *passive* if $u^{\mathrm{T}} y \geq 0$ for all (t, u);
(ii) *lossless* if $u^{\mathrm{T}} y = 0$ for all (t, u);
(iii) *output strictly passive* if $u^{\mathrm{T}} y \geq y^{\mathrm{T}} \varphi(y)$ for some function φ and all (t, u), and $y^{\mathrm{T}} \varphi(y) > 0$ for all $y \neq 0$. □

Definition 2.2 (Sector Bounds [37, Definition 6.2, pp. 232 – 233])**.** A memoryless function $h\colon [0, \infty) \times \mathbb{R}^q \to \mathbb{R}^q$ is said to belong to the sector

(i) $[0, \infty]$ if $u^{\mathrm{T}} h(t, u) \geq 0$ for all (t, u);
(ii) $[0, F]$ with $F = F^{\mathrm{T}} > 0$ if $h^{\mathrm{T}}(t, u)(h(t, u) - Fu) \leq 0$ for all (t, u). □

For the definition of the L_2-gain, recall that $L_{2e}(\mathbb{R}^q)$ is the extended L_2-space which consists of all measurable functions $\beta\colon [0, \infty) \to \mathbb{R}^q$ such that [41, pp. 2 – 3]

$$\|\beta_T\|_2^2 := \int_0^T \|\beta(t)\|^2 \, dt < \infty, \qquad \forall T \in [0, \infty).$$

A map $G\colon L_{2e}(\mathbb{R}^q) \to L_{2e}(\mathbb{R}^q)$ has *finite L_2-gain* if there exist finite constants $\gamma, b \in \mathbb{R}$ such that [41, Definition 1.2.1, p. 4, Lemma 2.2.13, p. 19]

$$\|(G(u))_T\|_2^2 \leq \gamma^2 \|u_T\|_2^2 + b, \qquad \forall u \in L_{2e}(\mathbb{R}^q), \forall T \geq 0. \tag{2.41}$$

Definition 2.3 (L_2-Gain [41, Lemma 2.2.13, p. 19])**.** The L_2-gain of $G\colon L_{2e}(\mathbb{R}^q) \to L_{2e}(\mathbb{R}^q)$ is defined as

$$\gamma_2(G) := \inf\{\gamma \in [0, \infty) \mid \exists b \in \mathbb{R} \text{ such that (2.41) holds}\},$$

with the infimum of an empty set being $+\infty$. □

Next, we define the *principal* and *complementary projection operators* induced by the projection operator $R_{\mathcal{I}^*}(x_g, y, r)$ in (2.27). Since the GPAW parameter $\Gamma \in \mathbb{R}^{q \times q}$ is symmetric positive definite, it can always be decomposed as $\Gamma = \Phi\Phi^{\mathrm{T}}$ for some nonsingular $\Phi \in \mathbb{R}^{q \times q}$ [124, Theorem 7.2.7, p. 406]. For any $\mathcal{I}$ such that $\operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$, the matrix $R_{\mathcal{I}}(x_g)$ (2.30) is well defined.[20] For any well defined $R_{\mathcal{I}}(x_g)$, let the *principal projection matrix*[21] $P_{\mathcal{I}}(x_g)$ and *complementary projection matrix* $S_{\mathcal{I}}(x_g)$ induced by $R_{\mathcal{I}}(x_g)$ be

$$P_{\mathcal{I}}(x_g) := \Phi^{-1} R_{\mathcal{I}}(x_g) \Phi, \qquad S_{\mathcal{I}}(x_g) := I - P_{\mathcal{I}}(x_g), \tag{2.42}$$

so that

$$R_{\mathcal{I}}(x_g) = \Phi P_{\mathcal{I}}(x_g) \Phi^{-1} = I - \Phi S_{\mathcal{I}}(x_g) \Phi^{-1}. \tag{2.43}$$

[20] The embedded inverse $(N_{\mathcal{I}}^{\mathrm{T}} \Gamma N_{\mathcal{I}}(x_g))^{-1}$ exists, or is not needed.

[21] This is the same $P_{\mathcal{I}}(x_g)$ in the proof of Proposition 2.5.2.

It can be verified from (2.30) that $P_{\mathcal{I}}(x_g)$ and $S_{\mathcal{I}}(x_g)$ take the explicit forms

$$P_{\mathcal{I}}(x_g) = \begin{cases} I - \tilde{N}_{\mathcal{I}}(\tilde{N}_{\mathcal{I}}^{\mathrm{T}}\tilde{N}_{\mathcal{I}})^{-1}\tilde{N}_{\mathcal{I}}^{\mathrm{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise,} \end{cases} \tag{2.44}$$

$$S_{\mathcal{I}}(x_g) = \begin{cases} \tilde{N}_{\mathcal{I}}(\tilde{N}_{\mathcal{I}}^{\mathrm{T}}\tilde{N}_{\mathcal{I}})^{-1}\tilde{N}_{\mathcal{I}}^{\mathrm{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases}$$

where $\tilde{N}_{\mathcal{I}}(x_g) := \Phi^{\mathrm{T}} N_{\mathcal{I}}(x_g)$. From [100, Lemma 1, Theorem 1], it can be seen that both are projection matrices. Moreover, it can be verified from the preceding definitions that they are idempotent [126, p. 697] and satisfy

$$P_{\mathcal{I}}(x_g) = P_{\mathcal{I}}^2(x_g) = P_{\mathcal{I}}^{\mathrm{T}} P_{\mathcal{I}}(x_g), \qquad S_{\mathcal{I}}(x_g) = S_{\mathcal{I}}^2(x_g) = S_{\mathcal{I}}^{\mathrm{T}} S_{\mathcal{I}}(x_g). \tag{2.45}$$

The *principal* and *complementary projection operators* are then given by

$$P_{\mathcal{I}^*}(x_g, y, r) = \Phi^{-1} R_{\mathcal{I}^*}(x_g, y, r)\Phi, \qquad S_{\mathcal{I}^*}(x_g, y, r) = I - P_{\mathcal{I}^*}(x_g, y, r),$$

respectively, where $\mathcal{I}^*$ is a solution to subproblem (2.31). It can be seen from (2.31) that $\operatorname{rank}(N_{\mathcal{I}^*}(x_g)) = |\mathcal{I}^*|$, so that all projection operators $R_{\mathcal{I}^*}(x_g, y, r)$, $P_{\mathcal{I}^*}(x_g, y, r)$, and $S_{\mathcal{I}^*}(x_g, y, r)$, are well defined. The next result states that the principal and complementary projection operators[22] are passive.

Proposition 2.7.1 (Passivity of Principal and Complementary Projection Operators)**.** *The projection operators*

(i) $w_1 \colon \mathbb{R}^q \to \mathbb{R}^q$ *defined by* $w_1 = P_{\mathcal{I}^*}(x_g, y, r)v_1$*; and*
(ii) $w_2 \colon \mathbb{R}^q \to \mathbb{R}^q$ *defined by* $w_2 = S_{\mathcal{I}^*}(x_g, y, r)v_2$*,*

are output strictly passive, satisfying $v_i^{\mathrm{T}} w_i = w_i^{\mathrm{T}} w_i \geq 0$ *for* $i \in \{1, 2\}$*.*

Proof. Consider case (i). From (2.45), we have

$$v_1^{\mathrm{T}} w_1 = v_1^{\mathrm{T}} P_{\mathcal{I}^*}(x_g, y, r)v_1 = v_1^{\mathrm{T}} P_{\mathcal{I}^*}^{\mathrm{T}}(x_g, y, r)P_{\mathcal{I}^*}(x_g, y, r)v_1 = w_1^{\mathrm{T}} w_1 \geq 0.$$

Defining $\varphi(w_1) := w_1$ for all w_1, we have $w_1^{\mathrm{T}}\varphi(w_1) = w_1^{\mathrm{T}} w_1 > 0$ for all $w_1 \neq 0$, which proves the conclusion for case (i). Case (ii) is proved similarly by invoking the idempotence of $S_{\mathcal{I}^*}(x_g, y, r)$ (2.45). ■

From Proposition 2.7.1, we have $w_i^{\mathrm{T}}(w_i - Iv_i) = 0 \leq 0$ for $i \in \{1, 2\}$, so that both w_1 and w_2 belong to the sector $[0, I]$ with I the identity matrix (see Definition 2.2). This implies that they also belong to the sector $[0, \infty]$. Being in the sector $[0, I]$ allows the use of output-feedback (see Fig. 2-4) to yield passive maps that belong to the sector $[0, \infty]$ [37, pp. 255 – 259]. By defining $\tilde{v}_i := v_i - w_i$ and $\tilde{w}_i := w_i$, it can be seen that $\tilde{v}_i^{\mathrm{T}}\tilde{w}_i = 0 \geq 0$, which show the systems in Fig. 2-4 to be lossless (and passive). The L_2-gain bound of the projection maps are presented next.

Proposition 2.7.2 (L_2-Gain Bound of Projection Maps)**.** *For any well defined trajectory* $(x_g(t), y(t), r(t)) \in \mathbb{R}^{q+p+n_r}$ *for all* $t \in [0, \infty)$*, let* $\mathcal{I}^*(t)$ *be a solution to subproblem* (2.31) *at time* t*. The* L_2*-gain of the projection maps*

[22] Both the map and its associated matrix are called *operators*.

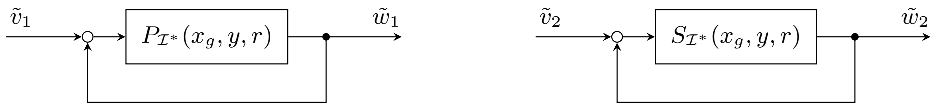

$\tilde{v}_1$
$P_{\mathcal{I}^*}(x_g, y, r)$
$\tilde{w}_1$
$\tilde{v}_2$
$S_{\mathcal{I}^*}(x_g, y, r)$
$\tilde{w}_2$

Figure 2-4: Unity positive output feedback yields lossless (passive) operators belonging to the sector $[0, \infty]$, with $\tilde{v}_i^{\mathrm{T}} \tilde{w}_i = 0$ for $i \in \{1, 2\}$.

(i) $w_1 \colon L_{2e}(\mathbb{R}^q) \to L_{2e}(\mathbb{R}^q)$ *defined by* $w_1(t) = P_{\mathcal{I}^*(t)}(x_g(t), y(t), r(t)) v_1(t)$ *for all* $t \in [0, \infty)$*; and*

(ii) $w_2 \colon L_{2e}(\mathbb{R}^q) \to L_{2e}(\mathbb{R}^q)$ *defined by* $w_2(t) = S_{\mathcal{I}^*(t)}(x_g(t), y(t), r(t)) v_2(t)$ *for all* $t \in [0, \infty)$*,*

satisfy the bound $\gamma_2(w_i) \leq 1$ *for* $i \in \{1, 2\}$*, where* γ_2 *is defined in Definition 2.3.*

Proof. The proof is similar to that of [37, Lemma 6.5, p. 242]. By Proposition 2.7.1, for $i \in \{1, 2\}$ and each $t \in [0, \infty)$, we have $v_i^{\mathrm{T}}(t) w_i(t) = w_i^{\mathrm{T}}(t) w_i(t)$, so that

$$\begin{aligned} 0 &= w_i^{\mathrm{T}}(t) w_i(t) - v_i^{\mathrm{T}}(t) w_i(t), \\ &= \tfrac{1}{2}(w_i(t) - v_i(t))^{\mathrm{T}}(w_i(t) - v_i(t)) + \tfrac{1}{2}\big(w_i^{\mathrm{T}}(t) w_i(t) - v_i^{\mathrm{T}}(t) v_i(t)\big), \\ &\geq \tfrac{1}{2}\big(w_i^{\mathrm{T}}(t) w_i(t) - v_i^{\mathrm{T}}(t) v_i(t)\big). \end{aligned}$$

Rearranging terms and taking the integral over $[0, T]$ yields

$$\int_0^T \|w_i(t)\|^2 \, dt \leq \int_0^T \|v_i(t)\|^2 \, dt, \qquad \forall T \geq 0,$$

so that $\|(w_i)_T\|_2^2 \leq \|(v_i)_T\|_2^2$ for all $T \geq 0$. Hence (2.41) holds with $b = 0$ and $\gamma = 1$, which implies the conclusion (see Definition 2.3). ■

Remark 2.24. Proposition 2.7.2 can be used to derive small-gain type [37, Theorem 5.6, p. 218] stability conditions for the GPAW-compensated system. □

Remark 2.25. Since $\Gamma = \Phi\Phi^{\mathrm{T}}$, it can be shown that the L_2-gain of the projection operator $R_{\mathcal{I}^*}(x_g, y, r) = \Phi P_{\mathcal{I}^*}(x_g, y, r)\Phi^{-1}$ is bounded above by $\|\Phi\|\|\Phi^{-1}\| = \kappa(\Phi)$, where $\kappa(A)$ is the *condition number* of matrix A [124, p. 336]. This can be related to the condition number of $\Gamma = \Gamma^{\mathrm{T}} > 0$ through

$$\begin{aligned} \kappa(\Phi) = \|\Phi\|\|\Phi^{-1}\| = \sqrt{\lambda_{\max}(\Phi^{\mathrm{T}}\Phi)}\sqrt{\lambda_{\max}(\Phi^{-\mathrm{T}}\Phi^{-1})} &= \frac{\sqrt{\lambda_{\max}(\Phi^{\mathrm{T}}\Phi)}}{\sqrt{\lambda_{\min}(\Phi\Phi^{\mathrm{T}})}} = \frac{\sqrt{\lambda_{\max}(\Phi\Phi^{\mathrm{T}})}}{\sqrt{\lambda_{\min}(\Phi\Phi^{\mathrm{T}})}}, \\ = \sqrt{\frac{\lambda_{\max}(\Gamma)}{\lambda_{\min}(\Gamma)}} = \left(\frac{\sqrt{\lambda_{\max}^2(\Gamma)}}{\sqrt{\lambda_{\min}^2(\Gamma)}}\right)^{\frac{1}{2}} &= \left(\frac{\sqrt{\lambda_{\max}(\Gamma^2)}}{\sqrt{\lambda_{\min}(\Gamma^2)}}\right)^{\frac{1}{2}} = \left(\frac{\sqrt{\lambda_{\max}(\Gamma^{\mathrm{T}}\Gamma)}}{\sqrt{\lambda_{\min}(\Gamma\Gamma^{\mathrm{T}})}}\right)^{\frac{1}{2}} = \sqrt{\kappa(\Gamma)}, \end{aligned}$$

where we have used $\lambda_{\max}(A) = \frac{1}{\lambda_{\min}(A^{-1})}$ for $A = A^{\mathrm{T}} > 0$, $\lambda(A^{\mathrm{T}}A) = \lambda(AA^{\mathrm{T}})$, and $\lambda^2(A) = \lambda(A^2)$. While this bound may not be tight, it suggests that the GPAW parameter should be chosen so that it is not ill-conditioned, i.e. choose Γ so that $\kappa(\Gamma) \not\gg 1$. □

The GPAW-compensated system (2.35) is illustrated in Fig. 2-5 with the projection operator $R_{\mathcal{I}^*}(x_g, y, r) = \Phi P_{\mathcal{I}^*}(x_g, y, r)\Phi^{-1}$. Observe that the nominal uncompensated system is obtained by replacing $R_{\mathcal{I}^*}(x_g, y, r)$ with the identity matrix. We leave the development of results based on Propositions 2.7.1 and 2.7.2 as future work (see Section 7.1.7).

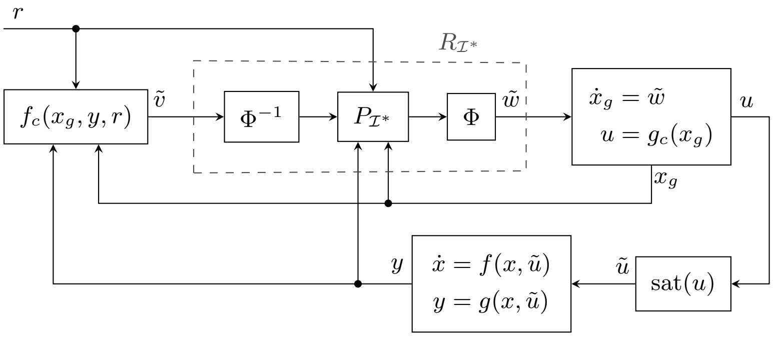


Figure 2-5: Closed-loop GPAW-compensated system as the uncompensated system modified by a passive operator $P_{\mathcal{I}^*}$ (with L_2-gain less than 1) and two constant transformation matrices, Φ, Φ^{-1}.

2.8 GPAW Compensation on a Two-link Robot Driven by an Adaptive Sliding Mode Controller

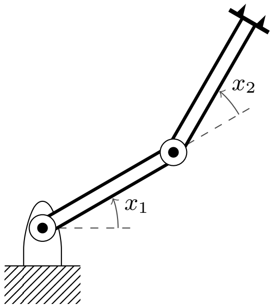


Figure 2-6: Nonlinear two-link robot.

To illustrate the pertinent features of GPAW compensation, we apply the GPAW scheme on a two-link robot (illustrated in Fig. 2-6) driven by an adaptive sliding mode controller [38, pp. 404 – 408], [102]. The model[23] of a saturated nonlinear two-link robot is given by [38, p. 393]

$$H(x_t)\ddot{x}_t + C(x_t, \dot{x}_t)\dot{x}_t = \text{sat}(u), \tag{2.46}$$

where sat(·) is the saturation function in (1.2), the truncated or partial state $x_t = [x_1, x_2]^{\mathrm{T}}$, and $\dot{x}_t$, $\ddot{x}_t$, are the vectors of joint angles, velocities, and accelerations respectively, and $u = [u_1, u_2]^{\mathrm{T}}$ is the vector of input joint torques. In (2.46), $H(x_t) \in \mathbb{R}^{2\times 2}$ is the symmetric inertia matrix, and $C(x_t, \dot{x}_t)\dot{x}_t \in \mathbb{R}^2$ is the vector of centripetal and Coriolis torques, defined by [38, p. 396]

$$H(x_t) = \begin{bmatrix} H_{11}(x_2) & H_{12}(x_2) \\ H_{12}(x_2) & H_{22} \end{bmatrix}, \qquad C(x_t, \dot{x}_t) = \tilde{h}(x_2) \begin{bmatrix} -\dot{x}_2 & -\dot{x}_1 - \dot{x}_2 \\ \dot{x}_1 & 0 \end{bmatrix}, \tag{2.47}$$

$$H_{11}(x_2) = h_{11}^{\mathrm{T}} a, \qquad H_{12}(x_2) = h_{12}^{\mathrm{T}} a, \qquad H_{22} = h_{22}^{\mathrm{T}} a, \qquad \tilde{h}(x_2) = h_0^{\mathrm{T}} a,$$

$$h_{11} := h_{11}(x_2) = [1, 0, 2\cos x_2, 2\sin x_2]^{\mathrm{T}}, \qquad h_{12} := h_{12}(x_2) = [0, 1, \cos x_2, \sin x_2]^{\mathrm{T}},$$

$$h_{22} := h_{22}(x_2) = [0, 1, 0, 0]^{\mathrm{T}}, \qquad h_0 := h_0(x_2) = [0, 0, \sin x_2, -\cos x_2]^{\mathrm{T}},$$

[23] The movement of the two-link robot is confined to the horizontal plane so that gravitational torques are not involved. Moreover, we have used $(x_t, x_1, x_2, \text{sat}(u))$ in place of (q, q_1, q_2, τ) in [38, p. 393].

and $a := [a_1, a_2, a_3, a_4]^{\mathrm{T}} = [3.34, 0.97, 3\sqrt{3}/5, 0.6]^{\mathrm{T}}$ is a vector of constant system parameters. As observed in [38, p. 394], the inertia matrix $H(x_t)$ is uniformly positive definite, i.e. there exists a constant $\alpha > 0$ such that $H(x_t) \geq \alpha I$ for all $x_t \in \mathbb{R}^2$, so that the inverse $H^{-1}(x_t)$ exists for all $x_t \in \mathbb{R}^2$. Then system (2.46) can be written in the form of (2.34), with state $x := [x_1, x_2, x_3, x_4]^{\mathrm{T}} := [x_1, x_2, \dot{x}_1, \dot{x}_2]^{\mathrm{T}} = (x_t, \dot{x}_t)$ and

$$f(x, \mathrm{sat}(u)) = \begin{bmatrix} x_3, \\ x_4, \\ H^{-1}(x_t)(-C(x_t, \dot{x}_t)\dot{x}_t + \mathrm{sat}(u)) \end{bmatrix}, \qquad g(x, \mathrm{sat}(u)) = x.$$

Observe that the definition of the output function g implies the entire state x is available to the controller.

An adaptive sliding mode controller for the two-link robot is described by [38, Section 9.2.1, pp. 404 – 408]

$$\begin{aligned} \dot{\hat{a}} &= -\Theta Y^{\mathrm{T}} s, \qquad \hat{a}(0) = \hat{a}_0, \\ u_c &= Y\hat{a} - K_D s, \end{aligned} \tag{2.48}$$

where[24]

$$s = \dot{e}_t + \Lambda e_t = \dot{x}_t - \dot{x}_{tr}, \qquad e_t = x_t - x_{td}, \qquad \dot{x}_{tr} = \dot{x}_{td} - \Lambda e_t = [\dot{x}_{tr1}, \dot{x}_{tr2}]^{\mathrm{T}}, \tag{2.49}$$

and the matrix $Y := Y(x, \dot{x}_{tr}, \ddot{x}_{tr})$ is defined by

$$H(x_t)\ddot{x}_{tr} + C(x_t, \dot{x}_t)\dot{x}_{tr} = Y(x, \dot{x}_{tr}, \ddot{x}_{tr})a. \tag{2.50}$$

Here, $\Theta = \Theta^{\mathrm{T}} > 0 \in \mathbb{R}^{4\times 4}$, $K_D = K_D^{\mathrm{T}} > 0 \in \mathbb{R}^{2\times 2}$, and $\Lambda = \Lambda^{\mathrm{T}} > 0 \in \mathbb{R}^{2\times 2}$ are chosen constant controller parameters (or gains), and $x_{td}(t) \in \mathbb{R}^2$, $t \in [0, \infty)$, is the desired reference trajectory for the joint angles x_t. We fix the controller parameters as

$$\Theta = \mathrm{diag}(30, 1, 10, 10), \qquad K_D = 10I, \qquad \Lambda = I, \tag{2.51}$$

for all (derivative) controllers to be defined, where $\mathrm{diag}(x_1, \ldots, x_i)$ is the diagonal matrix with $(x_1, \ldots, x_i)$ as its ordered diagonal entries. The reference trajectory is fixed to be the sinusoid[25]

$$x_{td}(t) = \pi(1 - \cos(2\pi t)) \left[\tfrac{1}{6}, \tfrac{1}{4}\right]^{\mathrm{T}}, \tag{2.52}$$

which may represent a pick-and-place operation. Using (2.47) to expand the left-hand-side of (2.50) yields

$$\begin{aligned} &\begin{bmatrix} h_{11}^{\mathrm{T}}a & h_{12}^{\mathrm{T}}a \\ h_{12}^{\mathrm{T}}a & h_{22}^{\mathrm{T}}a \end{bmatrix} \begin{bmatrix} \ddot{x}_{tr1} \\ \ddot{x}_{tr2} \end{bmatrix} + h_0^{\mathrm{T}}a \begin{bmatrix} -\dot{x}_2 & -\dot{x}_1 - \dot{x}_2 \\ \dot{x}_1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_{tr1} \\ \dot{x}_{tr2} \end{bmatrix} \\ &\qquad = \begin{bmatrix} \ddot{x}_{tr1}h_{11}^{\mathrm{T}} + \ddot{x}_{tr2}h_{12}^{\mathrm{T}} \\ \ddot{x}_{tr1}h_{12}^{\mathrm{T}} + \ddot{x}_{tr2}h_{22}^{\mathrm{T}} \end{bmatrix} a + \begin{bmatrix} -\dot{x}_2 & -\dot{x}_1 - \dot{x}_2 \\ \dot{x}_1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_{tr1} \\ \dot{x}_{tr2} \end{bmatrix} h_0^{\mathrm{T}}a, \\ &\qquad = \begin{bmatrix} \ddot{x}_{tr1}h_{11}^{\mathrm{T}} + \ddot{x}_{tr2}h_{12}^{\mathrm{T}} - (\dot{x}_{tr1}\dot{x}_2 + \dot{x}_{tr2}(\dot{x}_1 + \dot{x}_2))h_0^{\mathrm{T}} \\ \ddot{x}_{tr1}h_{12}^{\mathrm{T}} + \ddot{x}_{tr2}h_{22}^{\mathrm{T}} + \dot{x}_{tr1}\dot{x}_1 h_0^{\mathrm{T}} \end{bmatrix} a = Ya, \end{aligned}$$

[24] We have used $(\Theta, u_c, e_t, \dot{x}_{tr}, x_{td})$ in place of $(\Gamma, \tau, \tilde{q}, \dot{q}_r, q_d)$ in the description of the adaptive sliding mode controller in [38, Section 9.2.1, pp. 404 – 408].

[25] This choice of reference trajectory implies $\dot{x}_{td}(t) = 2\pi^2 \sin(2\pi t)[\frac{1}{6}, \frac{1}{4}]^{\mathrm{T}}$ and $\ddot{x}_{td}(t) = 4\pi^3 \cos(2\pi t)[\frac{1}{6}, \frac{1}{4}]^{\mathrm{T}}$.

which gives the expressions for elements of the matrix $Y = [y_{ij}] \in \mathbb{R}^{2\times 4}$ as

$$\begin{aligned}
&y_{11} = \ddot{x}_{tr1}, \qquad y_{12} = \ddot{x}_{tr2}, \qquad\qquad y_{23} = \ddot{x}_{tr1}\cos x_2 + \dot{x}_{tr1}\dot{x}_1 \sin x_2,\\
&y_{21} = 0, \qquad\quad y_{22} = \ddot{x}_{tr1} + \ddot{x}_{tr2}, \qquad y_{24} = \ddot{x}_{tr1}\sin x_2 - \dot{x}_{tr1}\dot{x}_1 \cos x_2,\\
&\quad y_{13} = (2\ddot{x}_{tr1} + \ddot{x}_{tr2})\cos x_2 - (\dot{x}_{tr1}\dot{x}_2 + \dot{x}_{tr2}(\dot{x}_1 + \dot{x}_2))\sin x_2,\\
&\quad y_{14} = (2\ddot{x}_{tr1} + \ddot{x}_{tr2})\sin x_2 + (\dot{x}_{tr1}\dot{x}_2 + \dot{x}_{tr2}(\dot{x}_1 + \dot{x}_2))\cos x_2.
\end{aligned} \tag{2.53}$$

The adaptive sliding mode controller (2.48) was designed particularly for the case when the system parameters $a = [a_1, a_2, a_3, a_4]^{\mathrm{T}}$ are *unknown*. Observe that the reference trajectory and its time derivatives $(x_{td}, \dot{x}_{td}, \ddot{x}_{td})$ are required to realize (2.48). The closed-loop system comprising the plant (2.46) and controller (2.48) with $u = u_c$ will be called the *nominal system* and denoted by Σ_n. If the plant is *unconstrained*, i.e. $u_{\max,i} = -u_{\min,i} = \infty$ for $i \in \{1, 2\}$, the same closed-loop system (2.46), (2.48), will be called the *unconstrained system* and denoted by Σ_u. As shown in [38, pp. 405 – 406], the *unconstrained system* Σ_u achieves global stability and convergence of tracking error, i.e. boundedness of $(x(t), \hat{a}(t))$ for all $t \in [0, \infty)$ and $(e_t(t), \dot{e}_t(t)) \to (0, 0)$ as $t \to \infty$.

Now, observe from (2.49) that s is a function of $(x_t, \dot{x}_t, x_{td}, \dot{x}_{td}) = (x, x_{td}, \dot{x}_{td})$, and $\dot{x}_{tr}$ is a function of $(x_t, x_{td}, \dot{x}_{td})$. In turn, these show $\ddot{x}_{tr}$ to be a function of $(\dot{x}_t, \dot{x}_{td}, \ddot{x}_{td})$ and $Y = Y(x, \dot{x}_{tr}, \ddot{x}_{tr})$ to be a function of $(x, x_{td}, \dot{x}_{td}, \ddot{x}_{td})$. By defining[26]

$$x_c := \hat{a}, \qquad y := x, \qquad r := (x_{td}, \dot{x}_{td}, \ddot{x}_{td}),$$

it is clear that the adaptive sliding mode controller (2.48) is of the form (2.24), with output equation being

$$g_c(x_c, y, r) = Y\hat{a} - K_D s = Y(y, r)x_c - K_D s(y, r). \tag{2.54}$$

However, to achieve *controller state-output consistency* (see Theorem 2.5.3), the nominal controller needs to be of the form (2.26), which we discuss next.

2.8.1 Approximate Nominal Controller

We will use techniques similar to those in Section 2.6 to derive an approximate nominal controller of the form (2.26) from the adaptive sliding mode controller (2.48). First, observe from the output equation (2.54) that we need to approximate Y and s. If approximations of both measurement and reference input $(y, r) = (x, x_{td}, \dot{x}_{td}, \ddot{x}_{td}) \in \mathbb{R}^{10}$ are used as in Section 2.6, we need to augment the controller state by 10 additional state variables. Also recall Remark 2.23 which cautions against excessive controller state augmentation. Examine (2.53) to see that the matrix Y can be approximated if we have estimates of the 5 signals $(\ddot{x}_{tr1}, \ddot{x}_{tr2}, x_2, \dot{x}_{tr1}\dot{x}_1, \dot{x}_{tr1}\dot{x}_2 + \dot{x}_{tr2}(\dot{x}_1 + \dot{x}_2))$. Adding 2 more signals, namely the elements of s, requires a total of 7 (< 10) augmented state variables.

Hence define the approximation $x_{caug} := [x_{c5}, x_{c6}, \ldots, x_{c11}]^{\mathrm{T}}$ of these 7 signals by a parameter $b > 0$ as

$$\begin{aligned}
\dot{x}_{caug} = b(z_{in} - x_{caug}), \qquad x_{caug}(0) = z_{in}(0) := z_{in}(y(0), r(0)),\\
z_{in} := [\ddot{x}_{tr1}, \ddot{x}_{tr2}, x_2, \dot{x}_{tr1}\dot{x}_1, \dot{x}_{tr1}\dot{x}_2 + \dot{x}_{tr2}(\dot{x}_1 + \dot{x}_2), s^{\mathrm{T}}(y, r)]^{\mathrm{T}}.
\end{aligned} \tag{2.55}$$

The matrix Y can then be approximated as $\hat{Y} = [\hat{y}_{ij}]$, whose elements are defined by

[26] Note that y is the measurement, and does not correspond to elements of $Y = [y_{ij}]$.

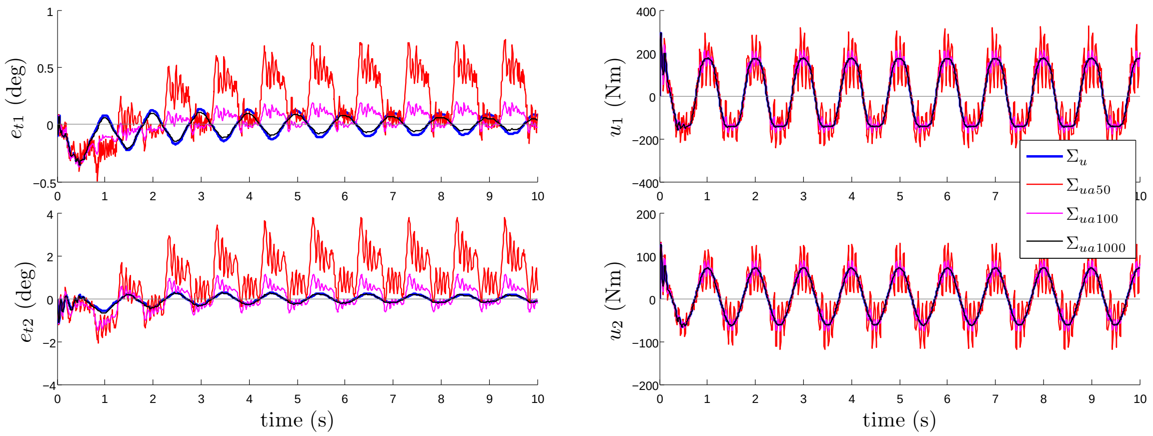


Figure 2-7: Time responses of unconstrained approximate system Σ_{ua} for $b \in \{50, 100, 1000\}$ compared to the response of the unconstrained system Σ_u. Observe that the responses of systems Σ_u and Σ_{ua1000} are almost indistinguishable.

(compare with (2.53))

$$\begin{aligned}
&\hat{y}_{11} = x_{c5}, \qquad \hat{y}_{12} = x_{c6}, \qquad \hat{y}_{23} = x_{c5}\cos x_{c7} + x_{c8}\sin x_{c7},\\
&\hat{y}_{21} = 0, \qquad \hat{y}_{22} = x_{c5} + x_{c6}, \qquad \hat{y}_{24} = x_{c5}\sin x_{c7} - x_{c8}\cos x_{c7},\\
&\hat{y}_{13} = (2x_{c5} + x_{c6})\cos x_{c7} - x_{c9}\sin x_{c7},\\
&\hat{y}_{14} = (2x_{c5} + x_{c6})\sin x_{c7} + x_{c9}\cos x_{c7}.
\end{aligned} \tag{2.56}$$

Since $\hat{s} := [x_{c10}, x_{c11}]^{\mathrm{T}}$ is an approximation of $s(y, r)$, we can define the augmented state $\tilde{x}_c := (x_c, x_{caug}) \in \mathbb{R}^{11}$ to write the augmented nominal controller as

$$\begin{aligned}
\dot{x}_c &= -\Theta Y^{\mathrm{T}}(y, r)s(y, r), & x_c(0) &= \hat{a}_0,\\
\dot{x}_{caug} &= b(z_{in}(y, r) - x_{caug}), & x_{caug}(0) &= z_{in}(0),\\
u_c &= \hat{Y}(x_{caug})x_c - K_D\hat{s}(x_{caug}) = g_c(\tilde{x}_c).
\end{aligned} \tag{2.57}$$

It is clear that the approximation will be improved as $b \to \infty$. The closed-loop system comprising the plant (2.46) and approximate controller (2.57) with $u = u_c$ will be called the *nominal approximate system* and denoted by Σ_{na}. If the plant is *unconstrained*, i.e. $u_{\max,i} = -u_{\min,i} = \infty$ for $i \in \{1, 2\}$, the same closed-loop system (2.46), (2.57) will be called the *unconstrained approximate system* and denoted by Σ_{ua}.

To choose the approximation parameter b, we simulate and compare the time responses of the unconstrained system Σ_u and unconstrained approximate system Σ_{ua} for different values of b. It was found numerically that setting $b = 27$ yields an unstable Σ_{ua}, while with $b = 28$, system Σ_{ua} is stable but with poor performance. Systems Σ_u and Σ_{ua} were compared for values of $b \in (\{50, 60, \ldots, 90, 100\} \cup \{150, 200, \ldots, 950, 1000\})$. Time responses of three cases corresponding to $b \in \{50, 100, 1000\}$ are shown in Fig. 2-7, with system Σ_{ua} denoted by $(\Sigma_{ua50}, \Sigma_{ua100}, \Sigma_{ua1000})$ respectively. In Fig. 2-7, the tracking errors $e_t = [e_{t1}, e_{t2}]^{\mathrm{T}}$ (see (2.49)) are shown together with the joint torques $u = [u_1, u_2]^{\mathrm{T}}$. Observe that the time responses of Σ_u and Σ_{ua1000} (for $b = 1000$) are almost indistinguishable in Fig. 2-7.

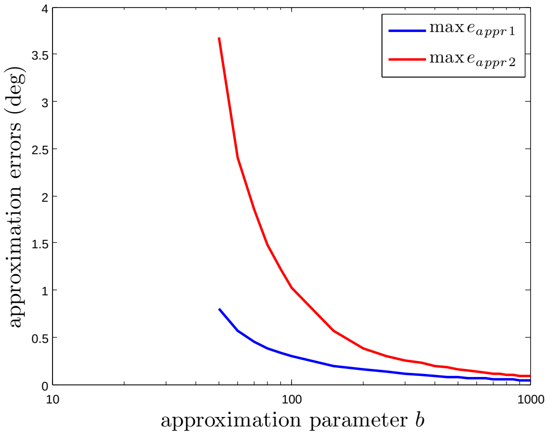


Figure 2-8: Approximation errors against approximation parameter (in logarithmic scale).

Define the maximum absolute approximation errors as[27]

$$\max e_{appr1} := \max_{t\in[5,10]} |x_{tu1}(t) - x_{tua1}(t)|, \qquad \max e_{appr2} := \max_{t\in[5,10]} |x_{tu2}(t) - x_{tua2}(t)|,$$

where $x_{tu} = [x_{tu1}, x_{tu2}]^{\mathrm{T}}$ and $x_{tua} = [x_{tua1}, x_{tua2}]^{\mathrm{T}}$ are the plant joint angles corresponding to systems Σ_u and Σ_{ua} respectively. The maximum absolute approximation errors are shown in Fig. 2-8 as b is varied between 50 and 1000. Clearly, the approximation errors decrease as b is increased, corroborating the claims of Section 2.6. Using Fig. 2-8 as a guide, we choose and fix the approximation parameter as $b = 1000$.

2.8.2 GPAW-Compensated Controller

Applying GPAW compensation to the augmented nominal controller (2.57) yields the GPAW-compensated controller (2.27) where (see (2.57))

$$\begin{aligned} f_c(x_g, y, r) &:= \begin{bmatrix} -\Theta Y^{\mathrm{T}}(y,r)s(y,r) \\ b(z_{in}(y,r) - \begin{bmatrix} 0 & I_7 \end{bmatrix} x_g) \end{bmatrix}, \\ g_c(x_g) &:= \hat{Y}(\begin{bmatrix} 0 & I_7 \end{bmatrix} x_g) \begin{bmatrix} I_4 & 0 \end{bmatrix} x_g - K_D \begin{bmatrix} 0 & I_2 \end{bmatrix} x_g, \end{aligned} \tag{2.58}$$

and I_j (for some positive integer j) is the $j \times j$ identity matrix. The projection operator $R_{\mathcal{I}^*}(x_g, y, r)$ in (2.27) is defined by a chosen GPAW parameter $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{11\times 11}$.

Our next goal is to rewrite the GPAW-compensated controller defined by Γ, (2.27), and (2.58), in closed-form, using (B.5) in Appendix B. Examining (B.5) shows that we only need to make explicit the expressions for the elements of $g_c = [g_{c1}, g_{c2}]^{\mathrm{T}}$ and their respective gradients ∇g_{c1} and ∇g_{c2}. First, with $x_g = [x_{g1}, x_{g2}, \ldots, x_{g11}]^{\mathrm{T}}$ and $K_D = [K_{Dij}]$, the elements of g_c in (2.58) are (see (2.56))

$$\begin{aligned} g_{c1}(x_g) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \left(\hat{Y}(\begin{bmatrix} 0 & I_7 \end{bmatrix} x_g) \begin{bmatrix} I_4 & 0 \end{bmatrix} x_g - K_D \begin{bmatrix} 0 & I_2 \end{bmatrix} x_g\right), \\ &= \hat{y}_{11}x_{g1} + \hat{y}_{12}x_{g2} + \hat{y}_{13}x_{g3} + \hat{y}_{14}x_{g4} - K_{D11}x_{g10} - K_{D12}x_{g11}, \\ &= x_{g1}x_{g5} + x_{g2}x_{g6} + x_{g3}((2x_{g5} + x_{g6})\cos x_{g7} - x_{g9}\sin x_{g7}) \\ &\qquad + x_{g4}((2x_{g5} + x_{g6})\sin x_{g7} + x_{g9}\cos x_{g7}) - K_{D11}x_{g10} - K_{D12}x_{g11}, \end{aligned}$$

[27] Note that the maximization over the interval $[5, 10]$ is to reflect the steady state approximation errors.

$$\begin{aligned} g_{c2}(x_g) &= \begin{bmatrix} 0 & 1 \end{bmatrix} \big(\hat{Y}(\begin{bmatrix} 0 & I_7 \end{bmatrix} x_g) \begin{bmatrix} I_4 & 0 \end{bmatrix} x_g - K_D \begin{bmatrix} 0 & I_2 \end{bmatrix} x_g\big), \\ &= \hat{y}_{21}x_{g1} + \hat{y}_{22}x_{g2} + \hat{y}_{23}x_{g3} + \hat{y}_{24}x_{g4} - K_{D21}x_{g10} - K_{D22}x_{g11}, \\ &= x_{g2}(x_{g5} + x_{g6}) + x_{g3}(x_{g5}\cos x_{g7} + x_{g8}\sin x_{g7}) \\ &\qquad + x_{g4}(x_{g5}\sin x_{g7} - x_{g8}\cos x_{g7}) - K_{D21}x_{g10} - K_{D22}x_{g11}. \end{aligned}$$

Their respective gradients $\nabla g_{c1} = [\nabla g_{c1,1}, \ldots, \nabla g_{c1,11}]^{\mathrm{T}}$, $\nabla g_{c2} = [\nabla g_{c2,1}, \ldots, \nabla g_{c2,11}]^{\mathrm{T}}$, can then be evaluated, whose elements are

$$\begin{aligned} \nabla g_{c1,1} &= x_{g5}, & \nabla g_{c2,1} &= 0, \\ \nabla g_{c1,2} &= x_{g6}, & \nabla g_{c2,2} &= x_{g5} + x_{g6}, \\ \nabla g_{c1,3} &= (2x_{g5} + x_{g6})\cos x_{g7} - x_{g9}\sin x_{g7}, & \nabla g_{c2,3} &= x_{g5}\cos x_{g7} + x_{g8}\sin x_{g7}, \\ \nabla g_{c1,4} &= (2x_{g5} + x_{g6})\sin x_{g7} + x_{g9}\cos x_{g7}, & \nabla g_{c2,4} &= x_{g5}\sin x_{g7} - x_{g8}\cos x_{g7}, \\ \nabla g_{c1,5} &= x_{g1} + 2x_{g3}\cos x_{g7} + 2x_{g4}\sin x_{g7}, & \nabla g_{c2,5} &= x_{g2} + x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7}, \\ \nabla g_{c1,6} &= x_{g2} + x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7}, & \nabla g_{c2,6} &= x_{g2}, \\ \nabla g_{c1,7} &= (-x_{g3}x_{g9} + x_{g4}(2x_{g5} + x_{g6}))\cos x_{g7} & \nabla g_{c2,7} &= (x_{g3}x_{g8} + x_{g4}x_{g5})\cos x_{g7} \\ &\quad - (x_{g3}(2x_{g5} + x_{g6}) + x_{g4}x_{g9})\sin x_{g7}, & &\quad + (-x_{g3}x_{g5} + x_{g4}x_{g8})\sin x_{g7}, \\ \nabla g_{c1,8} &= 0, & \nabla g_{c2,8} &= x_{g3}\sin x_{g7} - x_{g4}\cos x_{g7}, \\ \nabla g_{c1,9} &= -x_{g3}\sin x_{g7} + x_{g4}\cos x_{g7}, & \nabla g_{c2,9} &= 0, \\ \nabla g_{c1,10} &= -K_{D11}, & \nabla g_{c2,10} &= -K_{D21}, \\ \nabla g_{c1,11} &= -K_{D12}, & \nabla g_{c2,11} &= -K_{D22}. \end{aligned}$$

With the preceding expressions for g_{c1}, g_{c2}, ∇g_{c1}, and ∇g_{c2}, the closed-form expressions for the GPAW-compensated controller (2.27), (2.58), is given by (B.5). The closed-loop system comprising the plant (2.46) and GPAW-compensated controller (2.27), (2.58), with $u = u_g$ will be called the *GPAW-compensated system* and denoted by Σ_g.

2.8.3 Constrained Control

Recall that the inertia matrix $H(x_t)$ in (2.46) is uniformly positive definite, i.e. $H(x_t) \geq \alpha I$ for some $\alpha > 0$, and is defined by the system parameters a and joint angle x_2 (see (2.47)). Moreover, the controller state $\hat{a}$ in (2.48) has the interpretation as an estimate of the *unknown* system parameters a [38, p. 404]. This induces an *estimate* of the inertia matrix, defined by $\hat{a}$ instead of a (see (2.47)). For the augmented approximate controller (2.57) and GPAW-compensated controller (2.27) defined by (2.58), the partial states x_c in (2.57) and $[I_4, 0]x_g$ in (2.58) will have the same interpretation as estimates of the system parameter a. Here, we use ideas described in Remark 2.20 to enforce the positive *semidefiniteness* of the inertia matrix *estimate*, with the goal of enhancing system performance. Positive *semidefiniteness* of the inertia matrix estimate is a reasonable goal when $\alpha > 0$ in the relation $H(x_t) \geq \alpha I$ is *unknown.*

First, we derive necessary and sufficient conditions for a 2×2 symmetric matrix $A = [A_{ij}]$ to be positive *semidefinite.*[28] Recall that $A = A^{\mathrm{T}}$ is positive semidefinite if and only if all its eigenvalues are non-negative [124, Theorem 7.2.1, p. 402]. Its eigenvalues λ_1, λ_2 are roots

[28] The conditions presented yields two inequalities. Other necessary and sufficient conditions are possible, with perhaps more than two inequalities.

of the characteristic polynomial $\det(\lambda I - A) = 0$, which can be verified to be

$$\begin{aligned}\lambda_1(A) &= \tfrac{1}{2}\Big(A_{11} + A_{22} + \sqrt{(A_{11} + A_{22})^2 - 4(A_{11}A_{22} - A_{12}^2)}\Big),\\ \lambda_2(A) &= \tfrac{1}{2}\Big(A_{11} + A_{22} - \sqrt{(A_{11} + A_{22})^2 - 4(A_{11}A_{22} - A_{12}^2)}\Big).\end{aligned} \tag{2.59}$$

Clearly, we have $\lambda_1 \geq \lambda_2$, and A will be positive semidefinite if and only if $\lambda_2 \geq 0$. By inspection of the expression for λ_2, we see that $\lambda_2 \geq 0$ (and $A = A^{\mathrm{T}}$ is positive semidefinite) if and only if

$$A_{11} + A_{22} \geq 0, \qquad A_{11}A_{22} - A_{12}^2 \geq 0. \tag{2.60}$$

Next, we add constraints to the basic formulation of the GPAW-compensated controller to enforce the positive semidefiniteness of the inertia matrix estimate. Observe from (2.55) that x_2 is the third component of $z_{in} \in \mathbb{R}^7$. Decomposing the state as $x_g = [x_{g1}, x_{g2}, \ldots, x_{g11}]^{\mathrm{T}}$, it can be seen from the definition of f_c in (2.58) that x_{g7} (the third component of $[0, I_7]x_g$) is an estimate of the third component of z_{in}, so that it is an estimate of the joint angle x_2. Then we can define the inertia matrix *estimate* $\hat{H}(x_g)$ as (compare with (2.47))

$$\hat{H}(x_g) = \begin{bmatrix} \hat{H}_{11}(x_g) & \hat{H}_{12}(x_g) \\ \hat{H}_{12}(x_g) & \hat{H}_{22}(x_g) \end{bmatrix}, \qquad \begin{aligned} \hat{H}_{11}(x_g) &= x_{g1} + 2(x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7}),\\ \hat{H}_{12}(x_g) &= x_{g2} + x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7},\\ \hat{H}_{22}(x_g) &= x_{g2}.\end{aligned} \tag{2.61}$$

From (2.60), necessary and sufficient conditions for the inertia matrix estimate $\hat{H}(x_g)$ to be positive semidefinite are

$$\hat{H}_{11}(x_g) + \hat{H}_{22}(x_g) \geq 0, \qquad \hat{H}_{11}(x_g)\hat{H}_{22}(x_g) - \hat{H}_{12}^2(x_g) \geq 0.$$

To enforce positive semidefiniteness of $\hat{H}(x_g)$, define the two constraints

$$\begin{aligned} h_5(x_g) &= -\hat{H}_{11}(x_g) - \hat{H}_{22}(x_g),\\ &= -x_{g1} - x_{g2} - 2(x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7}) \leq 0,\\ h_6(x_g) &= -\hat{H}_{11}(x_g)\hat{H}_{22}(x_g) + \hat{H}_{12}^2(x_g),\\ &= x_{g2}(x_{g2} - x_{g1}) + (x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7})^2 \leq 0,\end{aligned} \tag{2.62}$$

in addition to the four saturation constraints $h_i(x_g) \leq 0$, $i \in \mathcal{I}_4$, defined by (2.28). The respective gradients $\nabla h_5(x_g) = [\nabla h_{5,1}, \ldots, \nabla h_{5,11}]^{\mathrm{T}}$ and $\nabla h_6(x_g) = [\nabla h_{6,1}, \ldots, \nabla h_{6,11}]^{\mathrm{T}}$ can be evaluated, whose non-zero elements can be verified to be

$$\begin{aligned} \nabla h_{5,1} &= -1, & \nabla h_{6,1} &= -x_{g2},\\ \nabla h_{5,2} &= -1, & \nabla h_{6,2} &= -x_{g1} + 2x_{g2},\\ \nabla h_{5,3} &= -2\cos x_{g7}, & \nabla h_{6,3} &= 2\cos x_{g7}(x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7}),\\ \nabla h_{5,4} &= -2\sin x_{g7}, & \nabla h_{6,4} &= 2\sin x_{g7}(x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7}),\\ \nabla h_{5,7} &= 2(x_{g3}\sin x_{g7} - x_{g4}\cos x_{g7}), & \nabla h_{6,7} &= 2(x_{g3}\cos x_{g7} + x_{g4}\sin x_{g7})\\ & & &\quad \times (-x_{g3}\sin x_{g7} + x_{g4}\cos x_{g7}).\end{aligned}$$

With the addition of the two constraints (2.62), construction of the constrained controller proceeds similarly as the construction of the GPAW-compensated controller in Section 2.5, resulting in a controller of similar description as (2.27). Theorem 2.5.3 can be modified to show that if there exists a $T \in \mathbb{R}$ such that $\hat{H}(x_g(T)) \geq 0$, then $\hat{H}(x_g(t)) \geq 0$ holds for all $t \geq T$.

As will be shown in Section 4.1, the GPAW-compensated controller (2.27) (and hence the constrained controller with additional constraints (2.62)) can be realized as (4.11)

$$\begin{aligned} \dot{x}_g &= \Phi x^*(x_g, y, r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g), \end{aligned} \tag{2.63}$$

where $\Phi \in \mathbb{R}^{11\times 11}$ is defined by a decomposition of the parameter $\Gamma = \Phi\Phi^{\mathrm{T}}$ [124, Theorem 7.2.7, p. 406], and x^* is the unique solution to the convex quadratic program (4.12)

$$\begin{aligned} &\min_{x\in\mathbb{R}^{11}} \|\Phi^{-1} f_c(x_g, y, r) - x\|^2, \\ &\text{subject to} \qquad N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}}(x_g)\Phi x \leq 0. \end{aligned} \tag{2.64}$$

In the implementation, we use the matrix square root for Φ, i.e. $\Phi = \sqrt{\Gamma}$. See Section 4.1.1 for guidelines to initialize the quadratic program solver. In the preceding, the functions f_c and g_c are defined in (2.58), and the definition of the active constraint set $\mathcal{I}_{\text{sat}}$ needs to be modified to $\mathcal{I}_{\text{sat}} := \mathcal{I}_{\text{sat}}(x_g) = \{i \in \mathcal{I}_6 \mid h_i(x_g) \geq 0\}$ to include the additional constraints (2.62). To solve (2.64), note that the matrix $N_{\mathcal{I}_{\text{sat}}}(x_g)$ can be constructed from $\nabla h_i(x_g)$ for $i \in \mathcal{I}_6$. The gradients $\nabla h_i(x_g)$ for $i \in \mathcal{I}_4$ are (see (2.28))

$$\begin{aligned} \nabla h_1(x_g) &= \nabla g_{c1}(x_g), & \nabla h_2(x_g) &= \nabla g_{c2}(x_g), \\ \nabla h_3(x_g) &= -\nabla g_{c1}(x_g), & \nabla h_4(x_g) &= -\nabla g_{c2}(x_g), \end{aligned}$$

and $\nabla g_{c1}(x_g), \nabla g_{c2}(x_g)$ can be evaluated as shown in Section 2.8.2. The closed-loop system comprising the plant (2.46) and constrained controller (2.63), (2.58), with $u = u_g$ will be called the *constraint controlled system* and denoted by $\Sigma_{\hat{H}}$.

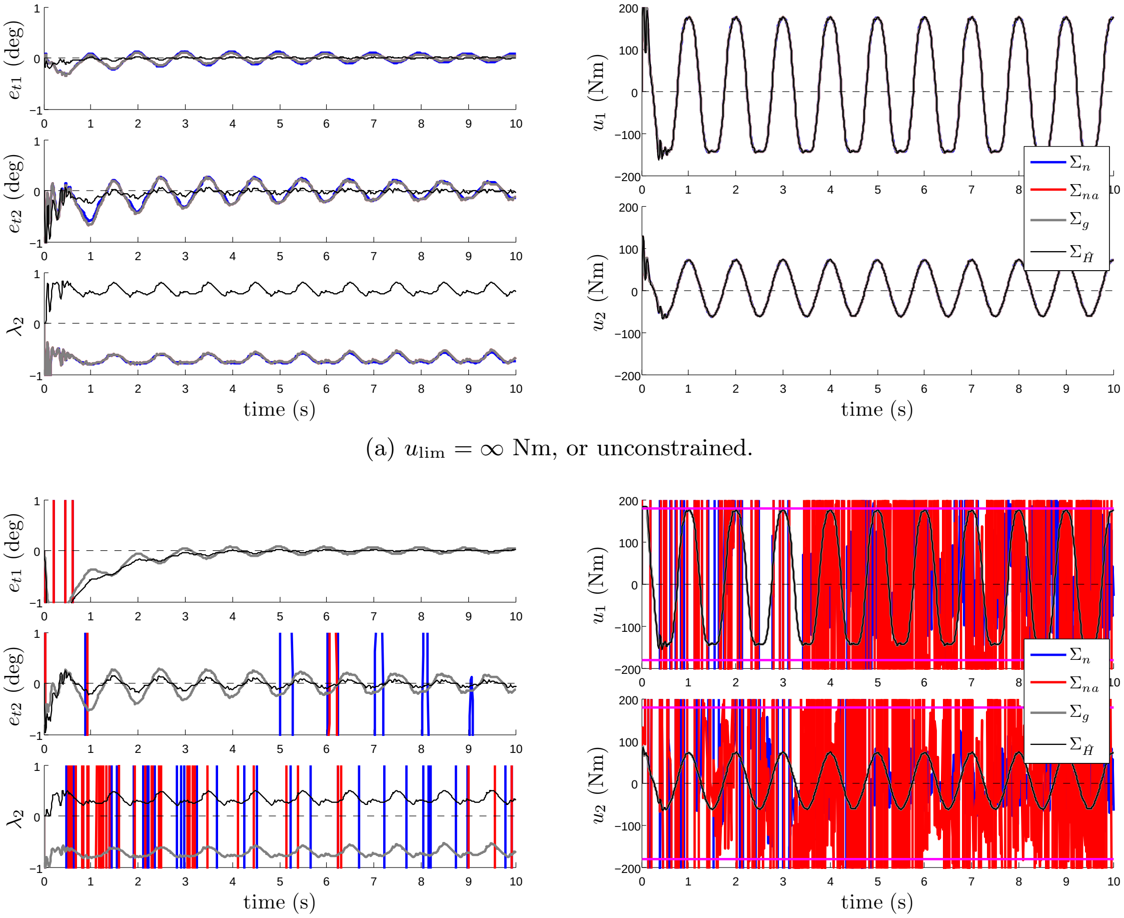
2.8.4 Numerical Results

First, observe that the four systems Σ_u, Σ_{ua}, Σ_n, and Σ_{na} have been fully defined by the controller parameters in (2.51) and approximation parameter b (= 1000). For the *GPAW-compensated system* Σ_g (see Section 2.8.2), it was found empirically that the GPAW parameter $\Gamma := \text{diag}(\Theta, 10^{-4} I_7)$, where Θ is defined in (2.51), yields a stabilizing GPAW-compensated controller (2.27), (2.58). For the *constraint controlled system* $\Sigma_{\hat{H}}$ (see Section 2.8.3), the same GPAW parameter induces a stabilizing constrained controller (2.63) with parameter $\Phi := \sqrt{\Gamma}$, that enforces positive semidefiniteness of the inertia matrix estimate $\hat{H}(x_g)$ (2.61). Using the reference trajectory (2.52), the four systems Σ_n, Σ_{na}, Σ_g, $\Sigma_{\hat{H}}$, are simulated for seven saturation levels,

$$u_{\max,i} = -u_{\min,i} = u_{\lim} \in \{\infty, 180, 150, 120, 90, 60, 30\} \text{ Nm}, \qquad \forall i \in \{1, 2\}.$$

Notice that when $u_{\lim} = \infty$ Nm, systems Σ_n and Σ_{na} correspond to systems Σ_u and Σ_{ua} respectively. The initial conditions used for all four systems are

$$(x_t(0), \dot{x}_t(0)) = (0, 0), \qquad \hat{a}_0 = 0, \qquad x_{caug}(0) = z_{in}(0), \qquad x_{c0} = (\hat{a}_0, x_{caug}(0)),$$



(a) $u_{\text{lim}} = \infty$ Nm, or unconstrained.

(b) $u_{\text{lim}} = 180$ Nm, or 100% steady-state control effectiveness.

Figure 2-9: Time responses for unconstrained and mild saturation cases.

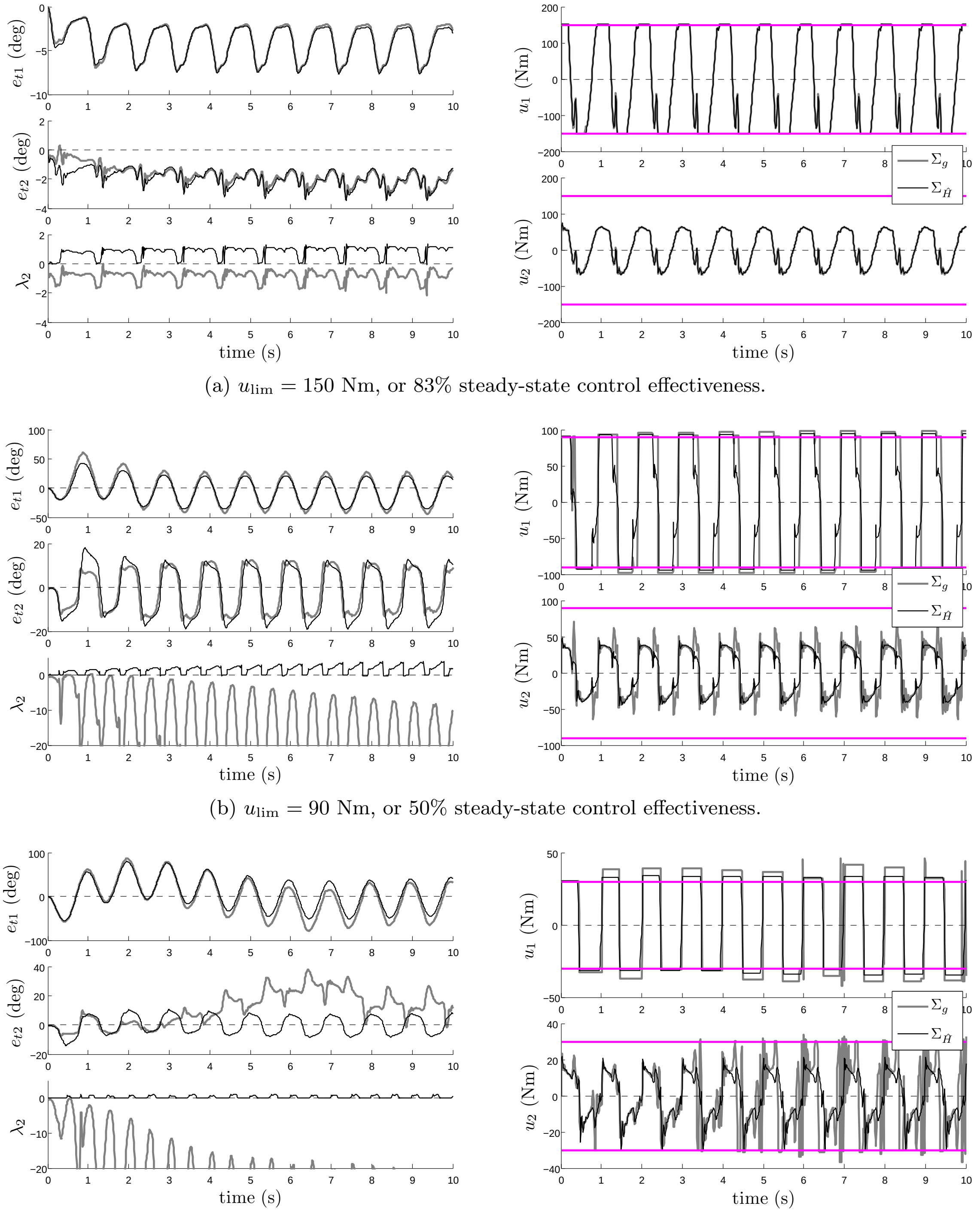
where z_{in} is defined in (2.55). The time responses for the unconstrained and mildly constrained cases, i.e. $u_{\text{lim}} \in \{\infty, 180\}$ Nm, are shown in Fig. 2-9, while the time responses for the severely constrained cases, i.e. $u_{\text{lim}} \in \{150, 90, 30\}$ Nm, are shown in Fig. 2-10.

In Figs. 2-9 and 2-10, the signal $\lambda_2 := \lambda_2(\hat{H})$ is the smaller eigenvalue of the inertia matrix estimate[29] $\hat{H}$ defined by (2.59). The remaining signals are the tracking errors $e_t = [e_{t1}, e_{t2}]^{\text{T}}$ (see (2.49)) and control signals $u = [u_1, u_2]^{\text{T}}$.

Observe from the unconstrained response in Fig. 2-9(a) that after transients, the control signals are less than 180 Nm in absolute value, i.e. $|u_1(t)| \leq 180$ Nm and $|u_2(t)| \leq 180$ Nm for all $t \geq 1$ s. We call the 180 Nm saturation level the 100% *steady-state control effectiveness* level. Then saturation levels $\{150, 120, 90, 60, 30\}$ Nm correspond to $\{83, 67, 50, 33, 17\}\%$ steady-state control effectiveness respectively.

Consider the unconstrained case $u_{\text{lim}} = \infty$ Nm in Fig. 2-9(a), where it can be seen that all four systems Σ_n, Σ_{na}, Σ_g, $\Sigma_{\hat{H}}$ are stable. The responses of systems Σ_g and Σ_{na} are numerically *identical*, while those of Σ_{na} and Σ_n are distinct but close, i.e. there is a small difference between the time responses of systems Σ_{na} and Σ_n. These show Σ_{na} to be a good

[29] For the nominal system Σ_n, the inertia matrix estimate $\hat{H}$ is defined by the joint angle x_2 rather than the joint angle *estimate* (see Section 2.8.3). For systems Σ_{na} and Σ_g, their inertia matrix estimates are defined as for system $\Sigma_{\hat{H}}$.



(a) $u_{\text{lim}} = 150$ Nm, or 83% steady-state control effectiveness.

(b) $u_{\text{lim}} = 90$ Nm, or 50% steady-state control effectiveness.

(c) $u_{\text{lim}} = 30$ Nm, or 17% steady-state control effectiveness.

Figure 2-10: Time responses for severe saturation cases. The responses of systems Σ_n and Σ_{na} are unstable (see Fig. 2-9(b) for an example), and excluded to prevent clutter.

approximation of Σ_n, and Σ_g recovers the response of the nominal approximate system Σ_{na} when the saturation limits are not triggered. Observe that system $\Sigma_{\hat{H}}$ exhibits superior tracking performance when compared to systems Σ_n, Σ_{na}, and Σ_g. Moreover, signal λ_2 for system $\Sigma_{\hat{H}}$ is always non-negative, which shows the inertia matrix estimate $\hat{H}$ to be always positive semidefinite. In contrast, the inertia matrix estimate $\hat{H}$ for systems Σ_n, Σ_{na}, and Σ_g are not always positive semidefinite. Since for system $\Sigma_{\hat{H}}$, the nominal system response (those of either Σ_n or Σ_{na}) is not recovered when saturation constraints are not triggered, it is *not* an anti-windup compensated system. This is expected because the constraints to enforce positive semidefiniteness of $\hat{H}$ may be triggered irrespective of any control saturation. Non-negativity of λ_2 (and hence positive semidefiniteness of $\hat{H}$) is also seen in Figs. 2-9(b) and 2-10. These show the effectiveness of the constrained controller (2.63) in enforcing the additional constraints (2.62).

In Fig. 2-9(b), we see that systems Σ_n and Σ_{na} became unstable. This is also the case for the remaining saturated cases in Fig. 2-10. To prevent clutter, the responses of systems Σ_n and Σ_{na} are omitted in Fig. 2-10.

For the saturated cases $u_{\lim} \in \{180, 150, 90, 30\}$ Nm in Figs. 2-9(b) and 2-10, we see that systems Σ_g and $\Sigma_{\hat{H}}$ are stable. Moreover, the saturation constraints $-u_{\lim} = u_{\min} \leq u \leq u_{\max} = u_{\lim}$ holds at least approximately. The violations of the saturation constraints in Figs. 2-10(b) and 2-10(c) can be attributed to numerical errors caused by finite discretization of time-steps in the simulations. Notice that these two cases correspond to 50% and 17% steady-state control effectiveness, which will be considered to be severely constrained by most standards.

To see the graceful performance degradation[30] induced by GPAW compensation, define the peak *steady-state* tracking errors

$$\max|e_{t1}| := \max_{t\in[5,10]} |e_{t1}(t)|, \qquad \max|e_{t2}| := \max_{t\in[5,10]} |e_{t2}(t)|.$$

In Fig. 2-11, the peak steady-state tracking errors for system Σ_g are shown against the saturation levels $u_{\lim} \in \{180, 150, 120, 90, 60, 30\}$ Nm. We see that as the saturation constraints become more severe, i.e. as $u_{\lim}$ *decrease*, the peak steady-state tracking errors increase *gradually*. While $\max|e_{t2}|$ does not increase monotonically with decrease of $u_{\lim}$, the fact that $\max|e_{t1}|$ does, implies that a performance measure of the form $\mu \max|e_{t1}| + |\max e_{t2}|$ will increase monotonically with decrease of $u_{\lim}$ when $\mu > 0$ is sufficiently large. Clearly, GPAW compensation induces graceful performance degradation.

2.9 Relations with Existing Methods

Here, we discuss similarities the GPAW scheme shares with existing anti-windup schemes and control methodologies. The first observation is that GPAW compensation requires the online solution to a combinatorial optimization subproblem.[31] As such, it has similarities with *model predictive control* (MPC) [6, 82], where an optimization problem is solved online. The GPAW scheme can be seen as a specialized form of MPC subject only to saturation constraints. Connections between anti-windup schemes and MPC are studied in [130].

[30] The term "graceful performance degradation" has been used in the anti-windup literature, e.g. [12, 24]. We interpret this as a *gradual* performance degradation as the severity of the saturation constraints *increases*.

[31] It will be shown in Section 4.1 that GPAW-compensated controllers can be equivalently defined by the online solution to a convex quadratic program or a projection onto a convex polyhedral cone problem.

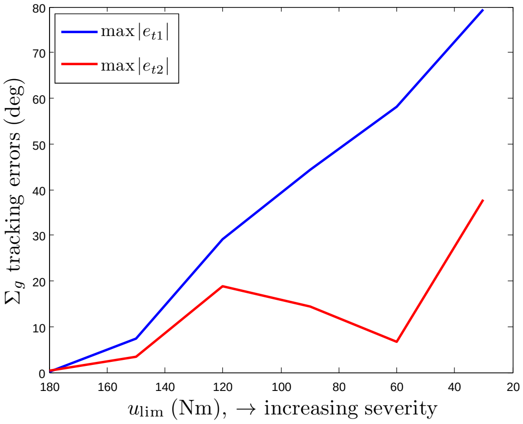


Figure 2-11: Tracking performance degrades gradually with severity of saturation constraints.

Closed-form expressions for GPAW-compensated controllers in Appendices A and B, namely (A.5) and (B.5), show that they are *switched* or hybrid controllers in general. Hence they have similarities with the switching anti-windup scheme of [131] and sliding mode anti-windup schemes of [132–134]. We note that the focus of these anti-windup schemes is presently for saturated LTI plants driven by LTI controllers.

Observe that GPAW compensation modifies the controller state exclusively through the projection operator. In contrast, the *conditioning technique* (see Section 1.4.1) modifies the reference input exclusively. Both attempt to achieve *controller state-output consistency*. The conditioning technique does this by an operation analogous to *back-calculation* [53], which can be thought of as a projection operation. If we think of the controller state or output as a pseudo "position", then GPAW compensation is projecting onto the "velocity space", whereas the back-calculation method attempts to project onto the "position space".

Minor similarities between the GPAW scheme and the *optimal directionality compensator* (see Section 1.4.5), *reference governor* (see Section 1.4.6) exist only in the sense that they all solve an optimization problem online.

2.10 Chapter Summary

We described the conditional integration method well-known for PID controllers, and motivated the need for a projection operator to extend it for general MIMO nonlinear controllers. The projection operator is obtained by extending Rosen's gradient projection method for nonlinear programming to continuous time. The GPAW-compensated controller was constructed using the projection operator, and defined by the online solution to a combinatorial optimization subproblem that always admits a feasible solution. By restricting to "strictly proper" nominal controllers, the GPAW scheme achieves *controller state-output consistency*, a unique property among anti-windup schemes. For nominal controllers that are not "strictly proper", they can be approximated arbitrarily well so that GPAW compensation can be applied to the approximate controllers. We showed that the projection operator is passive and has L_2-gain less than one. These properties can be used to derive passivity and small-gain based stability results. Pertinent features of GPAW compensation are demonstrated on a two-link robot driven by an adaptive sliding mode controller. These also show the possibility of using the projection operator for general constrained control.

Chapter 3

Input Constrained Planar LTI Systems

In Chapter 2, the construction of the GPAW-compensated controller was presented together with some fundamental properties, the most significant of which is the *controller state-output consistency* property (Theorem 2.5.3) that will be invoked in this chapter. Thus far, no stability results have been presented, which motivates the study of the simplest possible feedback system, namely, an input-constrained first order LTI plant driven by a first order LTI controller, where the objective is to regulate the system state about the origin. This case is particularly insightful because the closed-loop system is a planar dynamical system whose governing vector field is easily visualized, and is highly tractable because there is a large body of relevant work, e.g. [39, Chapter 2, pp. 31 – 75], [37, Chapter 2, pp. 35 – 86], [135, Chapter 2, pp. 51 – 77], [136, Chapter V & VI, pp. 253 – 418], [137, Chapter 3, 53 – 87], [138, Chapter 16, pp. 389 – 403], [139]. Related literature on input-constrained planar systems include [140–146].

After presenting the generalities in Section 3.1, we address the existence and uniqueness of solutions to the GPAW-compensated system. Due to *discontinuities* of the governing vector field of the GPAW-compensated system on the saturation constraint boundaries, classical existence and uniqueness results based on Lipschitz continuity of vector fields [37, 39, 46, 135, 137, 138] do not apply directly. We show that the GPAW-compensated system is in fact a *projected dynamical system* (PDS) [107–110] in Section 3.2. Observe that PDS is a significant line of independent research that has attracted economists, physicists, and mathematicians, among others. The link to PDS thus enables cross utilization of ideas and methods, as demonstrated in [147–151], [152, Section 4.1.5, pp. 81 – 84]. Using results from the PDS literature, existence and uniqueness of solutions to the GPAW-compensated system can thus be easily established, as shown in Section 3.3. In Section 3.4, equilibria of the systems are characterized, leading to the study of the associated region of attraction (ROA).

The ROA is a crucial property of control systems, the size of which limits the utility of the systems. In this sense, we consider the size of the ROA as a robustness property. Since it is widely accepted as a rule that the performance of a control system can be improved by trading off its robustness [153, Section 9.1, pp. 349 – 352], we consider an anti-windup scheme to be valid only if it can enhance performance *without reducing the system's ROA*. The first question to be addressed is whether the GPAW scheme satisfy such a criterion, and is shown to be affirmative in Section 3.5. Numerical results not only affirms the theoretical predictions, but shows the need to address *asymmetric* saturation constraints for general input-constrained systems, as discussed in Section 3.6.

In Section 3.7, we illustrate some qualitative weaknesses of some results in the anti-windup literature, and propose a new paradigm to address the anti-windup problem, in which results *relative* to the uncompensated system are sought. This new framework is reflected in the statement of the general anti-windup problem (Problem 1 of Section 1.3). In Section 3.8, we discuss how the obtained results on GPAW compensation for this simple system relate to the general anti-windup problem.

The final section (Section 3.9) shows that the solution of the GPAW-compensated system can *bounce*[1] on each saturation constraint boundary at most once. This is a consequence of the controller state-output consistency property unique to the GPAW scheme. It also suggests the possible existence of some optimality properties, which we present as some conjectures for future work.

3.1 Preliminaries

Let the first order single-input-single-output (SISO) input-constrained LTI plant (1.1), (1.2), be described by

$$\begin{aligned} \dot{x} &= ax + b\,\mathrm{sat}(u), \\ \mathrm{sat}(u) &= \max\{\min\{u, u_{\max}\}, u_{\min}\}, \end{aligned} \tag{3.1}$$

where $x, u \in \mathbb{R}$ are the state and control input respectively, and $a, b, u_{\min}, u_{\max} \in \mathbb{R}$ are constant plant parameters with $u_{\min}, u_{\max}$ satisfying $u_{\min} < 0 < u_{\max}$. Let the first order SISO LTI nominal controller (2.26) be[2]

$$\begin{aligned} \dot{x}_c &= \tilde{c}x_c + \tilde{d}x, \\ u &= \tilde{e}x_c, \end{aligned}$$

where $x_c, u \in \mathbb{R}$ are the controller state and output respectively, $x \in \mathbb{R}$ is the measurement of the plant state, and $\tilde{c}, \tilde{d}, \tilde{e} \in \mathbb{R}$ are the controller gains chosen to *globally* stabilize the origin of the *unconstrained* system, i.e. when $u_{\max} = -u_{\min} = \infty$. A simple transformation of the preceding nominal controller yields the equivalent realization

$$\dot{u} = cx + du, \tag{3.2}$$

with $c := \tilde{d}\tilde{e}$ and $d := \tilde{c}$. Applying the GPAW scheme to the preceding transformed nominal controller yields the GPAW-compensated controller[3]

$$\dot{u} = \begin{cases} 0, & \text{if } u \geq u_{\max} \wedge cx + du > 0, \\ 0, & \text{if } u \leq u_{\min} \wedge cx + du < 0, \\ cx + du, & \text{otherwise}, \end{cases} \tag{3.3}$$

which is similar[4] to the anti-windup compensated PID controller (2.1) obtained by the conditional integration method (see Remark 3.1). Observe that the first order GPAW-

[1] The notion of a solution *bounce* is made precise in Definition 3.3.

[2] See Section 2.6 if the nominal controller is of more general structure.

[3] This is obtained from the closed form expressions (A.7) in Appendix A, by replacing $(x_g, u_g, y, A_c, B_{cy}, B_{cr}, c_c)$ with $(u, u, x, d, c, 0, 1)$ respectively. Note also that $\wedge$ denotes the logical AND operator.

[4] This is expected because the GPAW scheme is a generalization of the conditional integration method.

compensated controller (3.3) is independent of the GPAW tuning parameter Γ in (2.30), which is true for all first-order controllers (see Remark B.1 in Appendix B).

Remark 3.1. Observe that controller (2.1) would be identical to (3.3) if (e, K_p, K_i, K_d) in (2.1) is replaced with $(cx + du, 0, 1, 0)$ respectively. □

The *nominal* constrained closed-loop system, Σ_n, is described by (3.1) and (3.2),

$$\Sigma_n\colon \begin{cases} \dot{x} = ax + b\,\mathrm{sat}(u), \\ \dot{u} = cx + du, \end{cases} \qquad \text{or} \qquad \Sigma_n\colon \dot{z} = f_n(z), \tag{3.4}$$

while the GPAW-compensated closed-loop system, Σ_g, is described by (3.1) and (3.3),

$$\Sigma_g\colon \begin{cases} \dot{x} = ax + b\,\mathrm{sat}(u), \\ \dot{u} = \begin{cases} 0, & \text{if } u \geq u_{\max} \wedge cx + du > 0, \\ 0, & \text{if } u \leq u_{\min} \wedge cx + du < 0, \\ cx + du, & \text{otherwise,} \end{cases} \end{cases} \qquad \text{or} \qquad \Sigma_g\colon \dot{z} = f_g(z). \tag{3.5}$$

Each of these planar systems can be expressed in the form $\dot{z} = f(z)$ with $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ being the governing vector field. The representing functions (vector fields) for systems Σ_n and Σ_g are denoted by f_n and f_g respectively, as indicated above.

The following assumption ensures that the origin of the *unconstrained* system, i.e. Σ_n with $u_{\max} = -u_{\min} = \infty$, is globally exponentially stable.

Assumption 3.1. For any plant parameters (a, d), the controller parameters (c, d) satisfy

$$a + d < 0, \tag{3.6}$$

$$ad - bc > 0, \tag{3.7}$$

and $bc \neq 0$. □

The characteristic equation of the *unconstrained* system can be verified to be $s^2 - (a + d)s + (ad - bc) = 0$, so that Assumption 3.1 ensures global exponential stability of the origin for the *unconstrained* system, as well as *local* exponential stability for both the nominal system Σ_n and the GPAW-compensated system Σ_g. The condition $bc \neq 0$ ensures (c, d) can be chosen to satisfy (3.6) and (3.7), and that Σ_n, Σ_g are *feedback* systems.

We will need the following sets[5]

$$\begin{aligned} K &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid u_{\min} < \bar{u} < u_{\max}\}, & \bar{K} &:= K \cup \partial K_+ \cup \partial K_-, \\ K_+ &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} > u_{\max}\}, & K_- &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} < u_{\min}\}, \\ \partial K_+ &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} = u_{\max}\}, & \partial K_- &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} = u_{\min}\}, \end{aligned}$$

$$\begin{aligned} \partial K_{+div} &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} > u_{\max}, c\bar{x} + d\bar{u} = 0\}, \\ K_{+in} &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} > u_{\max}, c\bar{x} + d\bar{u} < 0\}, \\ K_{+out} &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} > u_{\max}, c\bar{x} + d\bar{u} > 0\}, \\ \partial K_{+in} &:= \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} = u_{\max}, c\bar{x} + du_{\max} < 0\}, \end{aligned} \tag{3.8}$$

[5] In this chapter, K denotes the *interior* (an open set) of the unsaturated region in the state space of the *closed-loop system*. The closure of K, denoted by $\bar{K}$, is the unsaturated region. In other chapters, K denotes the unsaturated region in the *controller state space*.

$$\begin{aligned}
\partial K_{+out} &:= \{(\bar{x},\bar{u}) \in \mathbb{R}^2 \mid \bar{u} = u_{\max}, c\bar{x} + du_{\max} > 0\},\\
\partial K_{-div} &:= \{(\bar{x},\bar{u}) \in \mathbb{R}^2 \mid \bar{u} < u_{\min}, c\bar{x} + d\bar{u} = 0\},\\
K_{-in} &:= \{(\bar{x},\bar{u}) \in \mathbb{R}^2 \mid \bar{u} < u_{\min}, c\bar{x} + d\bar{u} > 0\},\\
K_{-out} &:= \{(\bar{x},\bar{u}) \in \mathbb{R}^2 \mid \bar{u} < u_{\min}, c\bar{x} + d\bar{u} < 0\},\\
\partial K_{-in} &:= \{(\bar{x},\bar{u}) \in \mathbb{R}^2 \mid \bar{u} = u_{\min}, c\bar{x} + du_{\min} > 0\},\\
\partial K_{-out} &:= \{(\bar{x},\bar{u}) \in \mathbb{R}^2 \mid \bar{u} = u_{\min}, c\bar{x} + du_{\min} < 0\},
\end{aligned}$$

and the points

$$z_+ := (-\tfrac{d}{c}u_{\max}, u_{\max}), \qquad z_- := (-\tfrac{d}{c}u_{\min}, u_{\min}),$$

which will be collectively referred by (3.8). Observe that the points z_+ and z_- are limit points [154, Definition 2.18, p. 32] of the line segments ∂K_{+div} and ∂K_{-div} respectively. These sets and associated vector fields are illustrated in Fig. 3-1 for an open-loop *unstable* plant, and in Fig. 3-2 for an open-loop *stable* plant.

Observe that $K_+ = K_{+in} \cup K_{+out} \cup K_{+div}$ and $\partial K_+ = \partial K_{+in} \cup \partial K_{+out} \cup \{z_+\}$, with analogous counterparts for K_- and ∂K_-. Furthermore, on ∂K_{+in} and ∂K_{-in}, vector fields of systems Σ_n and Σ_g (f_n and f_g respectively) point into K. On ∂K_{+out}, f_n points into K_+ and f_g points into ∂K_+. On ∂K_{-out}, f_n points into K_- and f_g points into ∂K_-.

By inspection of the vector fields f_n (3.4) and f_g (3.5) *from their definitions*, we have the following, which will be used in the proofs in this chapter.[6]

Fact 3.1.1 (Coincidence of Vector Fields)**.** *The vector fields f_n and f_g coincide in*

$$K \cup K_{+in} \cup K_{-in} \cup \partial K_{+in} \cup \partial K_{-in} \cup \partial K_{+div} \cup \partial K_{-div} \cup \{z_+, z_-\}.$$

In other words, they coincide in $\mathbb{R}^2 \setminus (K_{+out} \cup K_{-out} \cup \partial K_{+out} \cup \partial K_{-out})$.

Fact 3.1.2 (Solution Entry to Unsaturated Region)**.** *Any solution of systems Σ_n or Σ_g can pass from K_+ to K if and only if it intersects the line segment ∂K_{+in}, and analogously with respect to K_- and ∂K_{-in}.*

Fact 3.1.3 (Solution Exit from Unsaturated Region)**.** *Any solution of system Σ_n can pass from K to K_+ if and only if it intersects the line segment ∂K_{+out}, and analogously with respect to K_- and ∂K_{-out}.*

Also, recall Theorem 2.5.3, which implies that the solution of the GPAW-compensated system stays in the unsaturated region $\bar{K}$ once it reaches $\bar{K}$, i.e. $\bar{K}$ is a *positively invariant set* [135, p. 47] for system Σ_g.

3.2 GPAW-Compensated System as a Projected Dynamical System

Two of the most fundamental properties required for a meaningful study of dynamic systems are the existence and uniqueness of their solutions. As evident from the definition of

[6] These *facts* can be readily obtained by inspection of the vector fields. Moreover, recognizing that the results in this chapter have limited applicability, we name the results *claims* and *propositions* for what will conventionally be called *lemmas* and *theorems* respectively. These results are mainly to show the attractive features of the GPAW scheme, and gain insights to GPAW-compensated systems.

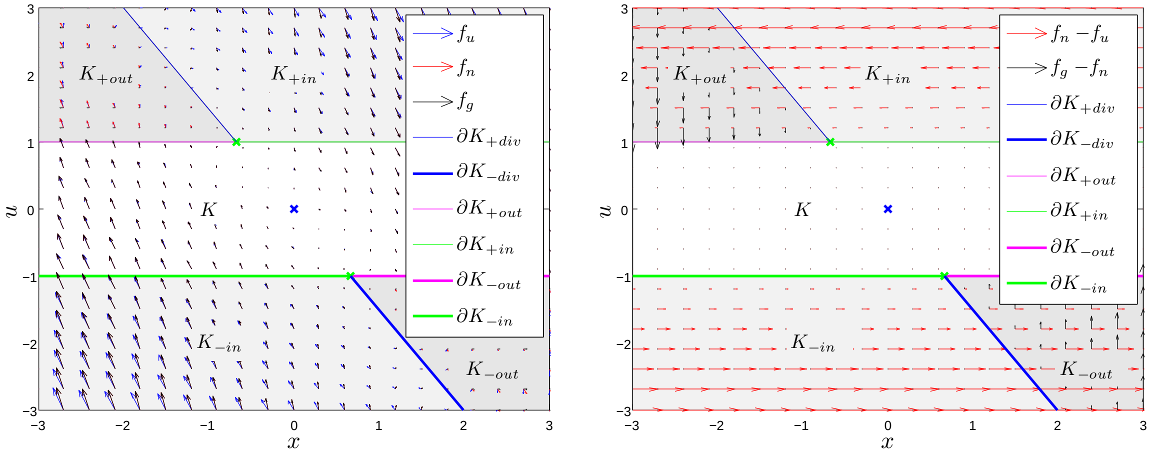


Figure 3-1: Closed-loop vector fields (f_n, f_g) of systems Σ_n, Σ_g and the *unconstrained* system (Σ_u, f_u), associated with an open-loop *unstable* system (plant and controller parameters: $a = 1$, $b = 1$, $c = -3$, $d = -2$, $u_{\max} = -u_{\min} = 1$). Vector fields of systems Σ_n, Σ_g and Σ_u (f_n, f_g, f_u) are shown on the left, while the vector field differences ($f_n - f_u$, $f_g - f_n$) are shown on the right.

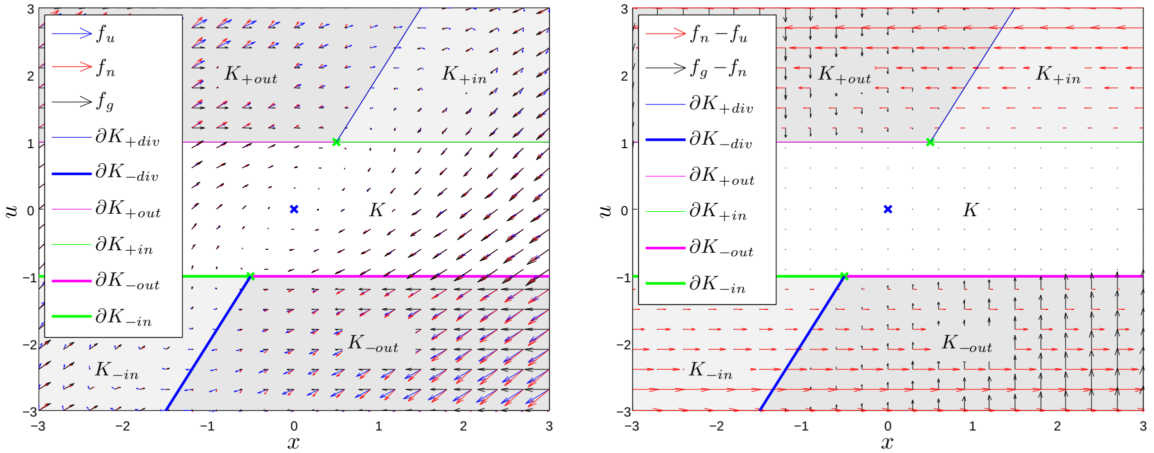


Figure 3-2: Closed-loop vector fields (f_n, f_g) of systems Σ_n, Σ_g and the *unconstrained* system (Σ_u, f_u) associated with an open-loop *stable* system (plant and controller parameters: $a = -1$, $b = 1$, $c = -1$, $d = 0.5$, $u_{\max} = -u_{\min} = 1$). Vector fields of systems Σ_n, Σ_g and Σ_u (f_n, f_g, f_u) are shown on the left, while the vector field differences ($f_n - f_u$, $f_g - f_n$) are shown on the right.

the GPAW-compensated controller (3.3) (see also Figs. 3-1 and 3-2), the vector field of the GPAW-compensated system f_g is *discontinuous* on the saturation constraint boundaries ∂K_{+out} ($\subset \partial K_+$) and ∂K_{-out} ($\subset \partial K_-$). Classical results on the existence and uniqueness of solutions [37, 39, 46, 135, 137, 138] rely on Lipschitz continuity [37, p. 87] of the governing vector fields, and hence do not apply to GPAW-compensated systems. While results in the theory of *differential equations with discontinuous right-hand sides* [120] can be used to assert such properties, we will use results from the *projected dynamical system* (PDS) literature [107–110] to assert the existence and uniqueness of solutions to the GPAW-compensated system Σ_g. Note that PDS is a significant line of independent research that has attracted the attention of economists, physicists, and mathematicians, among others. The link between control theory and PDS is known, e.g. [147–151], [152, Section 4.1.5, pp. 81 – 84], but little explored. Such links allow cross utilization of ideas and methods between different research fields, and is strategic in nature. We show here that the GPAW-compensated system Σ_g is indeed a PDS.

Observe that the unsaturated region $\bar{K}$ is a closed convex set (in fact, a closed convex polyhedron) with interior K and boundary $\partial K_+ \cup \partial K_-$. Let $P\colon \mathbb{R}^2 \to \bar{K}$ be the projection map defined for all $\tilde{z} \in \mathbb{R}^2$ by [107]

$$P(\tilde{z}) = \arg\min_{z \in \bar{K}} \|\tilde{z} - z\|.$$

It can be seen that $P((x,u)) = (x, \mathrm{sat}(u))$ for any $(x,u) \in \mathbb{R}^2$. Next, for any $\tilde{z} \in \bar{K}$, $v \in \mathbb{R}^2$, define the projection of vector v at $\tilde{z}$ by [107, 108]

$$\pi(\tilde{z}, v) = \lim_{\delta \downarrow 0} \frac{P(\tilde{z} + \delta v) - \tilde{z}}{\delta}.$$

Note that the limit is one-sided in the above definition [108]. The second order PDS is described by an ODE of the form [107]

$$\dot{z} = \pi(z, f(z)), \qquad z(0) \in \bar{K},$$

for some vector field $f\colon \mathbb{R}^2 \to \mathbb{R}^2$.

With the vector field f_n of Σ_n written explicitly as (see (3.4))

$$f_n(x,u) = \begin{bmatrix} ax + bu \\ cx + du \end{bmatrix}, \qquad \forall (x,u) \in \bar{K},$$

we have the following, the corollary of which is the desired result.

Claim 3.2.1. *For all $(x,u) \in \bar{K}$, the vector field f_g of the GPAW-compensated system Σ_g satisfies*

$$f_g(x,u) = \pi((x,u), f_n(x,u)).$$

Proof. If $(x,u) \in K$, the result follows from [108, Lemma 2.1(i)] and Fact 3.1.1. Next, consider a boundary point, $(x,u) \in \partial K_{+in} \cup \{z_+\}$. On this segment, we have $u = u_{\max}$ and $cx + du_{\max} \le 0$ from definition of the set $\partial K_{+in} \cup \{z_+\}$ (3.8). Since $\mathrm{sat}(u_{\max} + \delta\beta) = u_{\max} + \delta\beta$ for $\beta \le 0$ and a sufficiently small $\delta > 0$, we have

$$P((x,u) + \delta f_n(x,u)) = \begin{bmatrix} x + \delta(ax + bu) \\ \mathrm{sat}(u + \delta(cx + du)) \end{bmatrix} = \begin{bmatrix} x + \delta(ax + bu) \\ u + \delta(cx + du) \end{bmatrix},$$

so that

$$\pi((x,u),f_n(x,u)) = \lim_{\delta\downarrow 0}\frac{P((x,u)+\delta f_n(x,u))-(x,u)}{\delta} = \begin{bmatrix} ax+bu \\ cx+du \end{bmatrix} = f_n(x,u) = f_g(x,u),$$

for all $(x,u) \in \partial K_{+in} \cup \{z_+\}$, where the final equality follows from Fact 3.1.1.

Finally, consider a boundary point $(x,u) \in \partial K_{+out}$. On this segment, we have $u = u_{\max}$ and $cx + du_{\max} > 0$ from the definition of ∂K_{+out} (3.8). Since $\mathrm{sat}(u_{\max} + \delta\beta) = u_{\max}$ for $\beta > 0$ and a sufficiently small $\delta > 0$, we have

$$P((x,u)+\delta f_n(x,u)) = \begin{bmatrix} x+\delta(ax+bu) \\ \mathrm{sat}(u+\delta(cx+du)) \end{bmatrix} = \begin{bmatrix} x+\delta(ax+bu) \\ u \end{bmatrix},$$

so that

$$\pi((x,u),f_n(x,u)) = \lim_{\delta\downarrow 0}\frac{P((x,u)+\delta f_n(x,u))-(x,u)}{\delta} = \begin{bmatrix} ax+bu \\ 0 \end{bmatrix} = f_g(x,u),$$

for all $(x,u) \in \partial K_{+out}$, where the final equality follows from the definition of f_g (3.5) on ∂K_{+out}. The above established the claim for all points on $\bar{K} \setminus \partial K_-$. The verification on the boundary ∂K_- is similar to that for ∂K_+. ■

Corollary 3.2.2 (GPAW-Compensated System as a Projected Dynamical System)**.** *The GPAW-compensated system Σ_g is a projected dynamical system [107] governed by*

$$\dot{z} = f_g(z) = \pi(z, f_n(z)),$$

where $z = (x,u)$.

Corollary 3.2.2 will be used in the next section to assert the existence and uniqueness of solutions to system Σ_g. See [107–110] for a detailed development of PDS, and [147–151], [152, Section 4.1.5, pp. 81 – 84] for known relations with some classes of control systems.

3.3 Existence and Uniqueness of Solutions

Here, we assert the existence and uniqueness of solutions to both the nominal system and GPAW-compensated system.

Claim 3.3.1 (Existence and Uniqueness of Solutions to Nominal System)**.** *The nominal system Σ_n has a unique solution for all initial conditions $(x(0),u(0)) \in \mathbb{R}^2$ and all $t \geq 0$.*

Proof. For all $z := (x,u) \in \mathbb{R}^2$, the vector field f_n (3.4) can be written as

$$f_n(z) = Az + \begin{bmatrix} b \\ 0 \end{bmatrix}\mathrm{sat}(u), \qquad A = \begin{bmatrix} a & 0 \\ c & d \end{bmatrix}.$$

It can be verified [37, Example 3.2, pp. 91 – 92] that the saturation function is globally Lipschitz with unity Lipschitz constant, i.e. $|\mathrm{sat}(\alpha)-\mathrm{sat}(\beta)| \leq |\alpha-\beta|$. Then global Lipschitz continuity of f_n for all $t \in \mathbb{R}$ follows from

$$\begin{aligned}\|f_n(z) - f_n(\tilde{z})\| &= \|A(z - \tilde{z}) + [b, 0]^{\mathrm{T}}(\mathrm{sat}(u) - \mathrm{sat}(\tilde{u}))\|,\\ &\le \|A(z - \tilde{z})\| + \|[b, 0]^{\mathrm{T}}(\mathrm{sat}(u) - \mathrm{sat}(\tilde{u}))\|,\\ &= \|A(z - \tilde{z})\| + |b||\mathrm{sat}(u) - \mathrm{sat}(\tilde{u})|,\\ &\le \|A\|\|z - \tilde{z}\| + |b||u - \tilde{u}|,\\ &\le (\|A\| + |b|)\|z - \tilde{z}\|, \end{aligned} \tag{3.9}$$

for all $z := (x, u) \in \mathbb{R}^2$, $\tilde{z} := (\tilde{x}, \tilde{u}) \in \mathbb{R}^2$. By [37, Theorem 3.2, p. 93], Σ_n has a unique solution defined for all $t \ge 0$, for all $(x(0), u(0)) \in \mathbb{R}^2$. ■

We will need the following assumption used to assert the existence and uniqueness of solutions to PDS.

Assumption 3.2 ([107, Assumption 1])**.** There exists $B < \infty$ such that the vector field $f_n \colon \mathbb{R}^k \to \mathbb{R}^k$ satisfies the following conditions

$$\|f_n(z)\| \le B(1 + \|z\|), \qquad \forall z \in \bar{K}, \tag{3.10}$$

$$\langle f_n(z) - f_n(\tilde{z}), z - \tilde{z}\rangle \le B\|z - \tilde{z}\|^2, \qquad \forall z, \tilde{z} \in \bar{K}, \tag{3.11}$$

where $\langle x, y\rangle = x^{\mathrm{T}}y = y^{\mathrm{T}}x$ denotes the dot product of vectors x and y. □

The following result is stated without proof in the remark following [107, Assumption 1].

Claim 3.3.2. *If f_n is Lipschitz in $\bar{K} \subset \mathbb{R}^k$, then Assumption 3.2 holds.*

Proof. Since f_n is Lipschitz in $\bar{K}$, there exists an $L < \infty$ such that $\|f_n(z) - f_n(\tilde{z})\| \le L\|z - \tilde{z}\|$ for all $z, \tilde{z} \in \bar{K}$. To show that (3.10) holds, observe that

$$\begin{aligned}\|f_n(z)\| &= \|f_n(z) - f_n(\tilde{z}) + f_n(\tilde{z})\|,\\ &\le \|f_n(z) - f_n(\tilde{z})\| + \|f_n(\tilde{z})\|,\\ &\le L\|z - \tilde{z}\| + \|f_n(\tilde{z})\|,\\ &\le L\|z\| + L\|-\tilde{z}\| + \|f_n(\tilde{z})\|,\\ &= L\|z\| + L\|\tilde{z}\| + \|f_n(\tilde{z})\|,\end{aligned}$$

for all $z, \tilde{z} \in \bar{K}$. Fix any $\tilde{z} \in \bar{K}$ and define $\alpha := L\|\tilde{z}\| + \|f_n(\tilde{z})\|$ $(< \infty)$ and $B := \max\{L, \alpha\}$ $(< \infty)$, so that the preceding inequality becomes

$$\|f_n(z)\| \le L\|z\| + \alpha \le B(1 + \|z\|), \qquad \forall z \in \bar{K},$$

which proves (3.10).

By the Cauchy-Schwarz inequality [154, Theorem 1.35, pp. 15 – 16], we have

$$\langle f_n(z) - f_n(\tilde{z}), z - \tilde{z}\rangle \le \|f_n(z) - f_n(\tilde{z})\|\|z - \tilde{z}\| \le L\|z - \tilde{z}\|^2 \le B\|z - \tilde{z}\|^2, \qquad \forall z, \tilde{z} \in \bar{K},$$

which proves (3.11). ■

Remark 3.2. Both Assumption 3.2 and Claim 3.3.2 are stated for general vector fields f_n and regions $\bar{K}$ in $\mathbb{R}^k$. They will be specialized to vector fields and regions in $\mathbb{R}^2$ in the sequel. □

The following is the main result of this section.

Proposition 3.3.3 (Existence and Uniqueness of Solutions to GPAW-Compensated System)**.** *The GPAW-compensated system Σ_g has a unique solution for all initial conditions $(x(0), u(0)) \in \mathbb{R}^2$ and all $t \geq 0$.*

Proof. Since $f_n \colon \mathbb{R}^2 \to \mathbb{R}^2$ is globally Lipschitz (see (3.9)), it is Lipschitz in $\bar{K} \subset \mathbb{R}^2$, so that Assumption 3.2 holds due to Claim 3.3.2. Since Σ_g is a PDS (see Corollary 3.2.2 and [107, Equation (7)]), it follows from Assumption 3.2 and [107, Theorem 2] that Σ_g has a unique solution contained in $\bar{K}$ and defined for all $t \geq 0$ whenever the initial condition satisfies $(x(0), u(0)) \in \bar{K}$ (also recall Theorem 2.5.3). To assert the existence and uniqueness of solutions for all initial conditions $(x(0), u(0)) \in \mathbb{R}^2$, it is sufficient to establish this outside $\bar{K}$, and if the solution enters $\bar{K}$, there will be a unique continuation in $\bar{K}$ for all future times from this result.

Consider the region $K_+ = K_{+in} \cup K_{+out} \cup \partial K_{+div}$. The proof for the region K_- is similar. For any $z_1, z_2 \in K_+$, there are three possible cases. Firstly, in the region $\hat{K}_{+out} := K_{+out} \cup \partial K_{+div}$, we get from the definition of f_g (3.5) and $\hat{K}_{+out}$ (3.8), that $f_g(z) = f_g(x, u) = (ax + bu_{\max}, 0)$. Clearly, for any $z_1 := (x_1, u_1) \in \hat{K}_{+out}$, $z_2 := (x_2, u_2) \in \hat{K}_{+out}$, we have $\|f_g(z_1) - f_g(z_2)\| = |a(x_1 - x_2)| \leq L_{out}\|z_1 - z_2\|$ where $L_{out} := |a| < \infty$. Secondly, from Fact 3.1.1, f_g and f_n coincide in $\hat{K}_{+in} := K_{+in} \cup \partial K_{+div}$, so that f_g is also Lipschitz in $\hat{K}_{+in}$. For any $z_1, z_2 \in \hat{K}_{+in}$, we have $\|f_g(z_1) - f_g(z_2)\| \leq L_{in}\|z_1 - z_2\|$ where $L_{in} := \|A\| + |b| < \infty$ (see (3.9)). The last case corresponds to z_1 and z_2 being in *different* regions, $\hat{K}_{+in}$ and $\hat{K}_{+out}$. This can happen only if z_1 and z_2 do not *both* lie on the common line $\partial K_{+div} = \hat{K}_{+in} \cap \hat{K}_{+out}$, so that the straight line connecting z_1 and z_2 cannot be parallel to ∂K_{+div}. Without loss of generality, let $z_1 \in \hat{K}_{+in}$, $z_2 \in \hat{K}_{+out}$, and $\tilde{z}$ be the unique intersection point of ∂K_{+div} and the straight line connecting z_1 to z_2. Then $\tilde{z} \in \partial K_{+div}$ is such that $\tilde{z} \in \hat{K}_{+in} \cap \hat{K}_{+out}$, $\|z_1 - \tilde{z}\| \leq \|z_1 - z_2\|$, and $\|z_2 - \tilde{z}\| \leq \|z_1 - z_2\|$. Hence

$$\begin{aligned}
\|f_g(z_1) - f_g(z_2)\| &= \|f_g(z_1) - f_g(\tilde{z}) + f_g(\tilde{z}) - f_g(z_2)\|, \\
&\leq \|f_g(z_1) - f_g(\tilde{z})\| + \|f_g(z_2) - f_g(\tilde{z})\|, \\
&\leq L_{in}\|z_1 - \tilde{z}\| + L_{out}\|z_2 - \tilde{z}\|, \\
&\leq (L_{in} + L_{out})\|z_1 - z_2\|,
\end{aligned}$$

which, together with the first two cases, shows that f_g is Lipschitz in K_+. By [135, Theorem 3.1, pp. 18 – 19], Σ_g has a unique solution contained in K_+ whenever $(x(0), u(0)) \in K_+$. If the solution stays in K_+ for all $t \geq 0$, the claim holds. Otherwise, by [135, Theorem 2.1, p. 17], the solution can be continued to the boundary $\partial K_+ \subset \bar{K}$ of K_+. In this case, the first part of the proof shows that there is a unique continuation in $\bar{K}$ for all $t \geq 0$. ■

Remark 3.3. Care is due when interpreting the existence and uniqueness result of Proposition 3.3.3. Let $\phi_n(t, z_0)$ be the unique solution of system Σ_n starting from $z_0 \in \mathbb{R}^2$ at time $t = 0$. For system Σ_n, existence and uniqueness of solutions imply that no two different paths intersect [135, p. 38], and that

$$\phi_n(-t, \phi_n(t, z_0)) = z_0, \qquad \forall t \in \mathbb{R}, \forall z_0 \in \mathbb{R}^2.$$

That is, proceeding forwards and then backwards in time by the same amount, the solution always reaches its starting point. This is not true for system Σ_g whenever the solution intersects ∂K_{+out} or ∂K_{-out}. Inspection of the vector field f_g (3.5) reveals that when

Assumption 3.1 holds, all *forward* solutions either stay in ∂K_{+out} or ∂K_{-out} for all future times, or they eventually reach the points z_+ or z_-. Traversing *backwards in time* from any point of ∂K_{+out} or ∂K_{-out}, the solution stays on these segments indefinitely. That is, ∂K_{+out} and ∂K_{-out} are *negatively invariant sets* [135, p. 47] for system Σ_g. If a *forward* solution of Σ_g intersects ∂K_{+out} or ∂K_{-out} starting from some *interior* point $z_0 \in K$, then traversing backwards in time, the solution will never reach z_0.

Existence and uniqueness of solutions of system Σ_g means that if two distinct trajectories, $\phi_g(t, z_1)$, $\phi_g(t, z_2)$, intersect at some time, then they will be identical for all future times, i.e. if $\phi_g(T_1, z_1) = \phi_g(T_2, z_2)$ for some $T_1, T_2 \in \mathbb{R}$, then $\phi_g(t + T_1, z_1) = \phi_g(t + T_2, z_2)$ for all $t \geq 0$. Specifically, they can never diverge into two distinct trajectories. □

3.4 Existence of Multiple Equilibria

In this section, we characterize all equilibria of systems Σ_n and Σ_g. Of primary interest is the origin, stated next.

Claim 3.4.1 (Equilibrium Point at the Origin)**.** *The origin $z_{eq0} := (0, 0)$, is the* only *equilibrium point of systems Σ_n and Σ_g in K, and it must be either a stable node or stable focus.*

Proof. In K, the vector fields f_n and f_g coincide (see Fact 3.1.1), and can be written as $f_n(z) = f_g(z) = \tilde{A}z$, where $\tilde{A} = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. From (3.7), the matrix $\tilde{A}$ is invertible and hence, its null space is $\{z_{eq0}\}$, which shows z_{eq0} to be the *only* equilibrium point in K. Due to Assumption 3.1, z_{eq0} must be either a stable node or a stable focus [39, Section 2.2.1, pp. 32 – 35]. ■

Additional equilibria of the nominal system Σ_n are characterized below.

Claim 3.4.2 (Additional Equilibria of Nominal System)**.** *Apart from the origin z_{eq0}, the nominal system Σ_n admits two additional isolated equilibrium points defined by*

$$z_{eq+} := (-\tfrac{b}{a}u_{\max}, \tfrac{bc}{ad}u_{\max}), \qquad z_{eq-} := (-\tfrac{b}{a}u_{\min}, \tfrac{bc}{ad}u_{\min}),$$

only when

(i) the open-loop system is unstable ($a > 0$); or

(ii) the open-loop system is strictly *stable ($a < 0$) and controller parameter satisfies $d \in (0, -a)$.*

Moreover, if z_{eq+} and z_{eq-} are equilibria of Σ_n, they are saddle points and lie strictly in K_+ and K_- respectively, i.e. $z_{eq+}, z_{eq-} \notin (\partial K_+ \cup \partial K_-)$.

Remark 3.4. When z_{eq+} and z_{eq-} are equilibria of Σ_n, it can be verified that they must lie in ∂K_{+div} and ∂K_{-div} respectively. □

Proof. All equilibria of Σ_n are determined from the condition $f_n(z_{eq}) = 0$. With $z_{eq} := (x_{eq}, u_{eq})$, it can be verified from the conditions

$$f_n(z_{eq}) = \begin{bmatrix} ax_{eq} + b\,\mathrm{sat}(u_{eq}) \\ cx_{eq} + du_{eq} \end{bmatrix} = 0, \qquad bc \neq 0,$$

($bc \neq 0$ from Assumption 3.1) that whenever the open-loop system is marginally stable, i.e. $a = 0$, we have[7] $0 \cdot x_{eq} + b\,\mathrm{sat}(u_{eq}) = 0 \Rightarrow u_{eq} = 0 \Rightarrow cx_{eq} + d \cdot 0 = 0 \Rightarrow x_{eq} = 0$, and there can be no equilibria apart from z_{eq0}. Similarly, whenever $d = 0$, the preceding conditions imply there can be no additional equilibria apart from z_{eq0}. Together, these give $ad \neq 0$, and z_{eq+} and z_{eq-} are well-defined. A simple computation shows that apart from z_{eq0}, the additional equilibria are z_{eq+} and z_{eq-} as defined, provided $z_{eq+} \in K_+ \cup \partial K_+$ and $z_{eq-} \in K_- \cup \partial K_-$, which hold if and only if $ad \neq 0$ and $\frac{bc}{ad} \geq 1$. From (3.7), $\frac{bc}{ad} \geq 1$ holds if and only if $ad < 0$, which results in the *strict* condition $\frac{bc}{ad} > 1$. Therefore, if z_{eq+} and z_{eq-} are indeed equilibria of Σ_n, they must lie in K_+ and K_- respectively, i.e. they cannot lie on ∂K_+ or ∂K_-. If the open-loop system is unstable, i.e. $a > 0$, then from (3.6), we must have $d < -a < 0$, which implies $ad < 0$ and Σ_n indeed has z_{eq+} and z_{eq-} as equilibria. If the open-loop system is strictly stable, i.e. $a < 0$, then $ad < 0$ and (3.6) hold if and only if $d \in (0, -a)$. It remains to show that z_{eq+} and z_{eq-} must be *saddle points* [39, Section 2.2.1, pp. 32 – 35] whenever they are equilibria of Σ_n.

The Jacobian of f_n at the isolated equilibrium points $z_{eq+} \in K_+$ and $z_{eq-} \in K_-$ are identical and given by

$$\frac{\partial f_n}{\partial z}(z_{eq+}) = \frac{\partial f_n}{\partial z}(z_{eq-}) = A = \begin{bmatrix} a & 0 \\ c & d \end{bmatrix}.$$

Since its eigenvalues are a, d, and $ad < 0$, the equilibria z_{eq+} and z_{eq-} must be saddle points. ■

The following characterizes additional equilibria of the GPAW-compensated system Σ_g.

Claim 3.4.3 (Additional Equilibria of GPAW-Compensated System)**.** *Apart from the origin z_{eq0}, the GPAW-compensated system Σ_g admits additional equilibria only when*

(i) the open-loop system is unstable ($a > 0$). Then additional equilibria are all points on the two finite line segments defined by

$$Z_{eq+} = \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{x} = -\tfrac{b}{a}u_{\max}, u_{\max} \leq \bar{u} \leq \tfrac{bc}{ad}u_{\max}\} \subset (K_+ \cup \partial K_+),$$
$$Z_{eq-} = \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{x} = -\tfrac{b}{a}u_{\min}, \tfrac{bc}{ad}u_{\min} \leq \bar{u} \leq u_{\min}\} \subset (K_- \cup \partial K_-);$$

(ii) the open-loop system is strictly *stable ($a < 0$) and controller parameter satisfies $d \in (0, -a)$. Then additional equilibria are all points on the two* infinite *line segments defined by*

$$Z_{eq+} = \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{x} = -\tfrac{b}{a}u_{\max}, \bar{u} \geq \tfrac{bc}{ad}u_{\max}\} \subset K_+,$$
$$Z_{eq-} = \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{x} = -\tfrac{b}{a}u_{\min}, \bar{u} \leq \tfrac{bc}{ad}u_{\min}\} \subset K_-.$$

Remark 3.5. Observe that Σ_n admits additional equilibria if and only if Σ_g admits additional equilibria. Moreover, observe that z_{eq+} and z_{eq-} are endpoints of the line segments Z_{eq+} and Z_{eq-} respectively. □

Proof. All equilibria of Σ_g are determined from the condition $f_g(z_{eq}) = 0$. By Claim 3.4.1, any additional equilibria $z_{eq} := (x_{eq}, u_{eq})$ must satisfy either of the two conditions $u_{eq} \geq u_{\max}$ or $u_{eq} \leq u_{\min}$. Computation using the definition of f_g (3.5) shows that apart from

[7] Note that $\Rightarrow$ denotes logical implication.

z_{eq0}, all points in the sets

$$\tilde{Z}_{eq+} = \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{x} = -\tfrac{b}{a}u_{\max}, \bar{u} \geq u_{\max}, d\bar{u} \geq \tfrac{bc}{a}u_{\max}\},$$
$$\tilde{Z}_{eq-} = \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{x} = -\tfrac{b}{a}u_{\min}, \bar{u} \leq u_{\min}, d\bar{u} \leq \tfrac{bc}{a}u_{\min}\},$$

are also equilibria of Σ_g, provided these sets are well-defined and non-empty. It can be verified from the conditions $ax_{eq} + b\,\mathrm{sat}(u_{eq}) = 0$ (from $f_g(z_{eq}) = 0$) and $bc \neq 0$ (of Assumption 3.1), that whenever the open-loop system is marginally stable, i.e. $a = 0$, we have $0 \cdot x_{eq} + b\,\mathrm{sat}(u_{eq}) = 0 \Rightarrow u_{eq} = 0 \Rightarrow (u_{eq} \ngeq u_{\max}, u_{eq} \nleq u_{\min})$, and there can be no equilibria apart from z_{eq0}. Hence $a \neq 0$ and the sets $\tilde{Z}_{eq+}$ and $\tilde{Z}_{eq-}$ are well-defined. Considering the conditions $u \geq u_{\max}$ and $du \geq \frac{bc}{a}u_{\max}$ (and their analogous counterparts), these sets are non-empty if and only if (a) $d > 0$; (b) $d = 0$ and $\frac{bc}{a} \leq 0$; or (c) $d < 0$ and $\frac{bc}{ad} \geq 1$.

Consider case (a). From (3.6), this case ($d > 0$) is possible only when $a < 0$, i.e. the open-loop system is strictly stable. To satisfy (3.6) and $d > 0$, we must restrict $d \in (0, -a)$. Hence $ad < 0$ and (3.7) implies $\frac{bc}{ad} > 1$. The above sets $\tilde{Z}_{eq+}$ and $\tilde{Z}_{eq-}$ then simplifies to Z_{eq+} and Z_{eq-} respectively for case (ii).

Now consider case (b). With $d = 0$, conditions (3.6) and (3.7) reduces to $a < 0$ and $bc < 0$ respectively, which implies $\frac{bc}{a} > 0$. Therefore, Assumption 3.1 ensures that this case (in particular, $\frac{bc}{a} \leq 0$) cannot occur.

Finally, consider case (c). From (3.7), this case (in particular, $\frac{bc}{ad} \geq 1$) is possible only when $ad < 0$, which in turn implies $\frac{bc}{ad} > 1$ holds with *strict* inequality. The condition $ad < 0$ for this case (in particular, $d < 0$) implies $a > 0$, i.e. the open-loop system is unstable. It is easily verified that the above sets $\tilde{Z}_{eq+}$ and $\tilde{Z}_{eq-}$ then simplifies to Z_{eq+} and Z_{eq-} respectively for case (i). ■

Remark 3.6. Observe that the presence of additional equilibria precludes the possibility of the origin being a globally asymptotically stable equilibrium point for both systems Σ_n and Σ_g. However, note that (a, b, c, d) are given fixed parameters in the anti-windup context. □

In summary, z_{eq0} is an isolated stable equilibrium point of systems Σ_n and Σ_g for all $a, b, c, d \in \mathbb{R}$ satisfying Assumption 3.1, and it is the *only* equilibrium point in the interior of the unsaturated region K. When the open-loop system is marginally stable, or strictly stable with $d \leq 0$, there cannot be additional equilibria. When the open-loop system is unstable, or strictly stable with $d \in (0, -a)$, Σ_n has two more isolated saddle equilibrium points z_{eq+} and z_{eq-}, and Σ_g has a continuum of equilibria Z_{eq+} and Z_{eq-}.

3.5 Region of Attraction

The *region of attraction* (ROA) is a crucial property of control systems, the size of which limits the utility of the systems. In this sense, we consider the size of the ROA as a robustness property. Anti-windup schemes aim to enhance performance only in the presence of control saturation. Since it is widely accepted as a rule that the performance of a control system can be improved by trading off its robustness [153, Section 9.1, pp. 349 – 352], we consider an anti-windup scheme to be valid only if it can enhance performance *without reducing the system's ROA*.[8] We show in this section that GPAW compensation can only maintain/enlarge the ROA of system Σ_n. In other words, the ROA of the nominal system Σ_n is *contained within* the ROA of the GPAW-compensated system Σ_g.

[8] This is the reason for condition (i) in Problem 1 of Section 1.3.

While there may exist multiple equilibria for systems Σ_n and Σ_g (see Claims 3.4.2 and 3.4.3), we are primarily interested in the ROA of the equilibrium point at the origin, z_{eq0}. A distinguishing feature is that the results herein refers to the *exact ROA*, in contrast to *ROA estimates* that is found in a significant portion of the literature on anti-windup compensation. For clarity of presentation, we present the result in two parts, where the ROA containment is shown for the unsaturated region $\bar{K}$ and saturated region $\mathbb{R}^2 \setminus \bar{K}$ separately. Some numerical examples in Section 3.5.4 will illustrate typical ROAs and show that the said ROA containment can hold strictly for some systems. In the sequel, we will state and prove results only for one side of the state space, namely, with respect to the positive saturation region $K_+ \cup \partial K_+$. The analogous results with respect to the negative saturation region $K_- \cup \partial K_-$ can be readily extended, and will not be expressly stated.

Let $\phi_n(t, z_0)$ and $\phi_g(t, z_0)$ be the unique solutions of systems Σ_n and Σ_g respectively, both starting at initial state z_0 at time $t = 0$. The unique solutions exist due to Claim 3.3.1 and Proposition 3.3.3. The ROA of the origin $z_{eq0} := (0, 0)$ for systems Σ_n and Σ_g are then defined by [37, p. 314]

$$R_n := \{\bar{z} \in \mathbb{R}^2 \mid \phi_n(t, \bar{z}) \to z_{eq0} \text{ as } t \to \infty\}, \qquad R_g := \{\bar{z} \in \mathbb{R}^2 \mid \phi_g(t, \bar{z}) \to z_{eq0} \text{ as } t \to \infty\},$$

respectively. Recall the notion of *transverse sections* and ω *limit sets*.

Definition 3.1 (Transverse Section [39, p. 46])**.** A *transverse section* σ to a vector field $f \colon \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous, connected arc in $\mathbb{R}^2$ such that the dot product of the unit normal to σ and f is not zero and does not change sign on σ. □

In other words, the vector field has no equilibrium points on σ and is never tangent to σ [39, p. 46]. It is clear from the definitions of ∂K_{+in} and ∂K_{-in} (3.8) that both of these line segments are transverse sections to f_n and f_g. Moreover, ∂K_{+out} and ∂K_{-out} are also transverse sections to f_n.

Definition 3.2 (ω Limit Set [39, Definition 2.11, p. 44])**.** A point $z \in \mathbb{R}^2$ is said to be an ω *limit point* of a trajectory $\phi(t, z_0)$ if there exists a sequences of times $\{t_n\}, n \in \{1, 2, \ldots, \infty\}$, such that $t_n \uparrow \infty$ as $n \to \infty$ for which $\lim_{n\to\infty} \phi(t_n, z_0) = z$. The set of all ω limit points of a trajectory is called the ω *limit set* of the trajectory. □

For convenience, let the straight line connecting two points $\alpha, \beta \in \mathbb{R}^2$ be denoted by $l(\alpha, \beta)$ $(= l(\beta, \alpha))$ and defined by

$$l(\alpha, \beta) := \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \theta\alpha + (1 - \theta)\beta, \forall\theta \in (0, 1)\}.$$

Observe that $l(\alpha, \beta)$ does not contain the endpoints α, β, except for the degenerate case of *identical* endpoints, in which case, $l(\alpha, \alpha) = \{\alpha\}$. Next, the ROA containment in the unsaturated and saturated regions are shown separately, which combines to yield the desired result in Section 3.5.3 on page 92.

3.5.1 ROA Containment in Unsaturated Region

What follows is a series of intermediate claims to arrive at the main result of this section, Proposition 3.5.7. The proofs of the intermediate claims are available in Section 3.11.3. Let the straight lines connecting the origin to the points z_+ and z_- be

$$\sigma_+ := l(z_{eq0}, z_+) \cup \{z_+\}, \qquad \sigma_- := l(z_{eq0}, z_-) \cup \{z_-\}, \tag{3.12}$$

respectively. Consider a point $z_0 \in \partial K_{+in}$ with the property that $z_0 \in R_n$ and $\phi_n(t, z_0) \notin K_+$ for all $t \geq 0$. In other words, z_0 is in the ROA of system Σ_n and the solution starting from z_0 stays in $\bar{K} \cup K_-$ for all $t \geq 0$. As a consequence of Fact 3.1.3, $\phi_n(t, z_0)$ can never intersect ∂K_{+out} for all $t \geq 0$. Let

$$t_{int} := \inf\{\bar{t} \in (0, \infty) \mid \phi_n(\bar{t}, z_0) \in \sigma_+\}.$$

That is, t_{int} is the first time instant that the solution starting from z_0 at $t = 0$ intersects σ_+, or ∞ if it does not intersect σ_+. Define the path $\eta(z_0) \subset \mathbb{R}^2$ by

$$\eta(z_0) := \begin{cases} \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_n(t, z_0), \forall t \in [0, t_{int}]\} \cup l(\phi_n(t_{int}, z_0), z_+) \\ \qquad\qquad\qquad\qquad\qquad \cup \{z_+\} \cup l(z_0, z_+), & \text{if } t_{int} < \infty, \\ \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_n(t, z_0), \forall t \geq 0\} \cup \{z_{eq0}\} \cup \sigma_+ \cup l(z_0, z_+), & \text{otherwise,} \end{cases}$$

which can be verified to be closed and connected. Observe that $\eta(z_0)$ traces the path along the solution $\phi_n(t, z_0)$ until it intersects σ_+ or reach the origin, proceeds along σ_+ towards z_+, then along ∂K_{+in} until it reaches its starting point z_0. Let the *open, bounded* region enclosed by $\eta(z_0)$ be $D(z_0)$, and its closure be $\bar{D}(z_0)$. The region $D(z_0)$ is illustrated in Fig. 3-3.

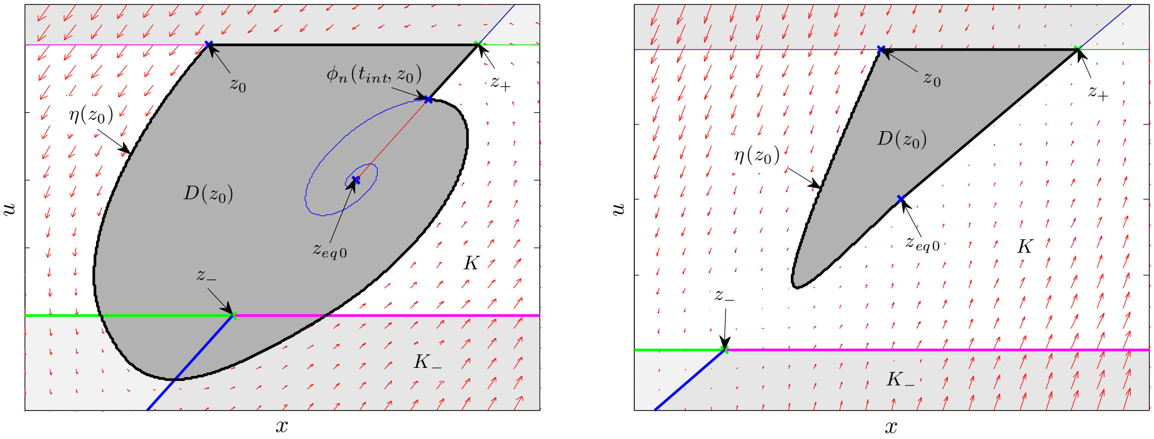


Figure 3-3: Closed path $\eta(z_0)$ encloses region $D(z_0) \subset \bar{K} \cup K_-$. A case where the solution enters K_- and also intersects σ_+ is shown on the left, while a case where the solution never enters K_- and never intersects σ_+ is shown on the right.

The following result states that $\bar{D}(z_0)$ is a *positively invariant set* [135, p. 47] for system Σ_n, and it must contain the origin z_{eq0}.

Claim 3.5.1 (Invariance of $\bar{D}(z_0)$)**.** *If there exists a point $z_0 \in \partial K_{+in}$ such that $z_0 \in R_n$ and $\phi_n(t, z_0) \in \bar{K} \cup K_-$ for all $t \geq 0$, then $\bar{D}(z_0) \subset \bar{K} \cup K_-$ is a positively invariant set for system Σ_n, and it must contain z_{eq0}, i.e. $z_{eq0} \in \bar{D}(z_0)$.*

Remark 3.7. Claim 3.5.1 states that under the assumptions, it is not possible for $\phi_n(t, z_0)$ to intersect σ_+ without having $\eta(z_0)$ enclose z_{eq0}, a case not illustrated in Fig. 3-3. □

Claim 3.5.2. *If there exists a point $z_0 \in \partial K_{+in}$ such that $z_0 \in R_n$ and $\phi_n(t, z_0) \in \bar{K}$ for all $t \geq 0$, then all points in $\bar{D}(z_0) \subset \bar{K}$ also lie in the ROA of system Σ_n, i.e. $\bar{D}(z_0) \subset R_n$.*

Remark 3.8. Specifically, the conclusion implies $z_+ \in \bar{D}(z_0) \subset R_n$. □

The points $\tilde{z}_+ \in \partial K_+$, $\tilde{z}_- \in \partial K_-$, and line segments $\xi_+ \subset \partial K_+$, $\xi_- \subset \partial K_-$, defined by

$$\begin{aligned} \tilde{z}_+ &:= (-\tfrac{b}{a}u_{\max}, u_{\max}), \qquad & \xi_+ &:= l(\tilde{z}_+, z_+), \\ \tilde{z}_- &:= (-\tfrac{b}{a}u_{\min}, u_{\min}), \qquad & \xi_- &:= l(\tilde{z}_-, z_-), \end{aligned}$$

will be needed in the subsequent development.

Claim 3.5.3. *If the open-loop system is marginally or strictly stable, i.e. $a \leq 0$, then f_g points towards z_+ on ∂K_{+out}, i.e. $f_g(z) = \alpha(z_+ - z)$ for all $z \in \partial K_{+out}$ and some $\alpha := \alpha(z) > 0$. If the open-loop system is unstable, i.e. $a > 0$, then f_g points towards z_+ on ξ_+, $f_g(\tilde{z}_+) = 0$, and f_g points away from z_+ on $\partial K_{+out} \setminus (\xi_+ \cup \{\tilde{z}_+\})$.*

Remark 3.9. It is clear that when $a > 0$, we have $\tilde{z}_+ \in Z_{eq+}$ where Z_{eq+} is the set of equilibria defined in item (i) of Claim 3.4.3. □

Claim 3.5.4. *If the open-loop system is unstable, i.e. $a > 0$, and $z_0 \in \partial K_{+out} \cap R_n$, then $z_0 \in \xi_+$.*

The above results are summarized below. It shows how the solution of the GPAW-compensated system Σ_g must behave on ∂K_{+out} when the initial state is in the ROA of the nominal system Σ_n.

Claim 3.5.5. *If there exists a $z_0 \in \partial K_{+out} \cap R_n$, then for every $z \in l(z_0, z_+) \cup \{z_0\}$, there exists a $T(z) \in (0, \infty)$ such that the solution of system Σ_g satisfies $\phi_g(T(z), z) = z_+$ and $\phi_g(t, z) \in \partial K_{+out}$ for all $t \in [0, T(z))$.*

Remark 3.10. Observe that under the assumptions, the solution $\phi_g(t, z_0)$ of the GPAW-compensated system *slides* along the line segment ∂K_{+out} (or ξ_+ as appropriate) to reach z_+. Note that Theorem 2.5.3 corroborates this observation. □

Next, we will show that a solution of Σ_n converging to the origin can intersect ∂K_{+out} or ∂K_{-out} only in a specific way, namely that subsequent intersection points, if any, must steadily approach z_+ or z_-.

Claim 3.5.6. *If $z_0 \in \partial K_{+out} \cap R_n$ and there exists a $T \in (0, \infty)$ such that $\phi_n(T, z_0) \in \partial K_{+out}$, then $\phi_n(T, z_0) \in l(z_0, z_+)$.*

The following is the main result of this section. The proof amounts to using the solution of Σ_n to bound the solution of Σ_g.

Proposition 3.5.7 (ROA Containment in Unsaturated Region)**.** *The part of the ROA of the origin of system Σ_n contained in $\bar{K}$, is itself contained within the ROA of the origin of system Σ_g, i.e. $(R_n \cap \bar{K}) \subset R_g$.*

Remark 3.11. The distinction between the solutions of systems Σ_n and Σ_g, namely $\phi_n(t, z)$ and $\phi_g(t, z)$, and their ROAs, R_n and R_g, should be kept clear when examining the proof below. □

Proof. The following argument will be used repeatedly in the present proof. If for some $z \in \bar{K}$, we have $\phi_n(t, z) \in \bar{K}$ for all $t \geq 0$, then Fact 3.1.3 implies that $\phi_n(t, z)$ cannot intersect ∂K_{+out} or ∂K_{-out}, i.e. $\phi_n(t, z) \in \bar{K} \setminus (\partial K_{+out} \cup \partial K_{-out})$ for all $t \geq 0$. Fact 3.1.1 shows that f_n and f_g coincide in $\bar{K} \setminus (\partial K_{+out} \cup \partial K_{-out})$, which implies $\phi_g(t, z) = \phi_n(t, z)$ for all $t \geq 0$. If in addition, we have $\lim_{t\to\infty} \phi_n(t, z) = z_{eq0}$, then $\lim_{t\to\infty} \phi_g(t, z) = \lim_{t\to\infty} \phi_n(t, z) = z_{eq0}$.

In summary, if $\phi_n(t, z) \in \bar{K}$ for all $t \geq 0$ and $z \in R_n$, then $z \in R_g$. For ease of reference, we call this the *coincidence argument*.

We need to show that if $z_0 \in R_n \cap \bar{K}$, then $z_0 \in R_g$. Let $z_0 \in R_n \cap \bar{K}$, so that $\phi_n(0, z_0) = z_0 \in \bar{K}$ and $\lim_{t\to\infty} \phi_n(t, z_0) = z_{eq0}$. Consider the case where $\phi_n(t, z_0)$ stays in $\bar{K}$ for all $t \geq 0$. It follows from the *coincidence argument* that $z_0 \in R_g$.

Now, we let the solution $\phi_n(t, z_0)$ enter K_+ and consider all possible continuations. Due to Fact 3.1.3, $\phi_n(t, z_0)$ must intersect ∂K_{+out} at least once. If $\phi_n(t, z_0)$ intersects ∂K_{+out} multiple times, it can only intersect it for finitely many times. Otherwise, there is an infinite sequence of times $\{t_m\}, m \in \{1, 2, \ldots, \infty\}$ such that $t_m \uparrow \infty$ as $m \to \infty$ for which $\phi_n(t_m, z_0) \in \partial K_{+out}$. Since $z_0 \in R_n$, it follows that $\phi_n(t_m, z_0) \in \partial K_{+out} \cap R_n$ for every m. As a consequence of Claim 3.5.6, we have $\lim_{m\to\infty} \phi_n(t_m, z_0) = z_+$, which shows that z_+ is an ω limit point of $\phi_n(t, z_0)$. But this is impossible because $\lim_{t\to\infty} \phi_n(t, z_0) = z_{eq0} \neq z_+$. Similarly, if $\phi_n(t, z_0)$ intersects ∂K_{-out} multiple times, it can only intersect it for finitely many times.

Hence, let T_1 and T_2 be the first and last times for which $\phi_n(t, z_0)$ intersects ∂K_{+out}, and let T_3 be the (*only*) time after T_2 that $\phi_n(t, z_0)$ intersects ∂K_{+in}. Then we have $0 \leq T_1 \leq T_2 < T_3 < \infty$, and

$$\phi_n(T_1, z_0), \phi_n(T_2, z_0) \in \partial K_{+out}, \qquad \phi_n(t, z_0) \in K_+, \forall t \in (T_2, T_3), \qquad \phi_n(T_3, z_0) \in \partial K_{+in},$$

with behavior after T_3 to be specified. Let $z_1 := \phi_n(T_1, z_0)$, $z_2 := \phi_n(T_2, z_0)$, and $z_3 := \phi_n(T_3, z_0)$. Since $z_0 \in R_n$, we have

$$z_1, z_2 \in \partial K_{+out} \cap R_n, \qquad z_3 \in \partial K_{+in} \cap R_n.$$

It is clear that $\phi_g(t, z_0) = \phi_n(t, z_0)$ for all $t \in [0, T_1]$. By Claim 3.5.5, there exist a $\tilde{T}_1 < \infty$ such that

$$\phi_g(T_1 + \tilde{T}_1, z_0) = \phi_g(\tilde{T}_1, \phi_g(T_1, z_0)) = \phi_g(\tilde{T}_1, \phi_n(T_1, z_0)) = \phi_g(\tilde{T}_1, z_1) = z_+. \tag{3.13}$$

Because $\phi_n(t, z_0)$ cannot intersect ∂K_{+out} for all $t > T_2$, the only possible continuations from time T_3 ($> T_2$) onwards are

(i) $\phi_n(t, z_0)$ stays in $\bar{K}$ for all $t \geq T_3$, or
(ii) $\phi_n(t, z_0)$ enters K_- at some finite time.

Consider case (i), which implies $\bar{D}(z_3) \subset \bar{K}$ (see Fig. 3-3). Claim 3.5.2 yields $z_+ \in \bar{D}(z_3) \subset R_n$ (see also Remark 3.8), and Claim 3.5.1 shows that $\bar{D}(z_3)$ is a positively invariant set for system Σ_n. Then we have $\phi_n(t, z_+) \in \bar{D}(z_3) \subset \bar{K}$ for all $t \geq 0$. It follows from the *coincidence argument* that $z_+ \in R_g$. Because $\phi_g(t, z_+) = \phi_g(t, \phi_g(T_1 + \tilde{T}_1, z_0))$ for all $t \geq 0$ (see (3.13)), we have $z_0 \in R_g$, as desired.

Now, consider case (ii). Due to Fact 3.1.3, $\phi_n(t, z_0)$ must intersect ∂K_{-out} at least once. From the above discussion, $\phi_n(t, z_0)$ can intersect ∂K_{-out} only finitely many times. Let T_4 be the first time (after T_3) and T_5 be the last time for which $\phi_n(t, z_0)$ intersects ∂K_{-out}, and let T_6 be the (*only*) time after T_5 that $\phi_n(t, z_0)$ intersects ∂K_{-in}. Then $T_3 < T_4 \leq T_5 < T_6 < \infty$, and

$$\phi_n(T_4, z_0), \phi_n(T_5, z_0) \in \partial K_{-out}, \qquad \phi_n(t, z_0) \in K_-, \forall t \in (T_5, T_6), \qquad \phi_n(T_6, z_0) \in \partial K_{-in}.$$

Let $z_4 := \phi_n(T_4, z_0)$, $z_5 := \phi_n(T_5, z_0)$ and $z_6 := \phi_n(T_6, z_0)$. Since $z_0 \in R_n$, we have

$$z_4, z_5 \in \partial K_{-out} \cap R_n, \qquad z_6 \in \partial K_{-in} \cap R_n.$$

Now, the only possible continuation after T_6 is for $\phi_n(t, z_0) \in \bar{K}$ for all $t \geq T_6$. Recall the definition of $\eta(z)$ and $\bar{D}(z)$ for some $z \in \partial K_{+in} \cap R_n$, as illustrated in Fig. 3-3. It is clear that $z_+ \in \bar{D}(z_3)$. Claim 3.5.1 shows that $\bar{D}(z_3)$ (with a portion in K_-) is a positively invariant set for system Σ_n, so that $\phi_n(t, z_+) \in \bar{D}(z_3)$ for all $t \geq 0$. Recall also, that $\phi_g(T_1 + \tilde{T}_1, z_0) = z_+$ (3.13) and we want to show that $z_+ \in R_g$ (which implies $z_0 \in R_g$).

There are two possible ways for the solution $\phi_n(t, z_+)$ to continue. Either $\phi_n(t, z_+)$ stays in $\bar{D}(z_3) \cap \bar{K}$ for all $t \geq 0$, or it enters $\bar{D}(z_3) \cap K_-$ at some finite time. If $\phi_n(t, z_+) \in \bar{D}(z_3) \cap \bar{K}$ for all $t \geq 0$, then as in the proof of Claim 3.5.2 (on page 104), Bendixson's Criterion [37, Lemma 2.2, pp. 67] and the absence of saddle points in $\bar{D}(z_3) \cap \bar{K}$ means that $\{z_{eq0}\}$ is the ω limit set of $\phi_n(t, z_+)$ and hence $z_+ \in R_n$. By the *coincidence argument*, we have $z_+ \in R_g$. It follows from $\phi_g(t, z_+) = \phi_g(t, \phi_g(T_1 + \tilde{T}_1, z_0))$ for all $t \geq 0$ (see (3.13)), that $z_0 \in R_g$.

Finally, consider when $\phi_n(t, z_+)$ enters $\bar{D}(z_3) \cap K_-$ at some finite time. By Fact 3.1.3, $\phi_n(t, z_+)$ must intersect ∂K_{-out} at least once. Let $\tilde{T}_2 < \infty$ be such that $\phi_n(\tilde{T}_2, z_+) \in \partial K_{-out}$ and $\phi_n(t, z_+) \in K$ for all $t \in (0, \tilde{T}_2)$, and let $\tilde{z}_2 := \phi_n(\tilde{T}_2, z_+) \in \partial K_{-out}$. Because the boundary of $\bar{D}(z_3)$ intersects ∂K_{-out} at z_4 and $\tilde{z}_2 \in \bar{D}(z_3) \cap \partial K_{-out}$, we have that $\tilde{z}_2 \in l(z_4, z_-)$. Since $z_4 \in \partial K_{-out} \cap R_n$, we have by (the analogous counterpart to) Claim 3.5.5 that there exists a $\tilde{T}_3 < \infty$ such that $\phi_g(\tilde{T}_3, \tilde{z}_2) = z_-$. Since $z_6 \in \partial K_{-in} \cap R_n$, it follows from (the analogous counterparts to) Claims 3.5.2 and 3.5.1 that $z_- \in \bar{D}(z_6) \subset R_n$, $\bar{D}(z_6)$ is a positively invariant set, and $\phi_n(t, z_-) \in \bar{D}(z_6) \subset \bar{K}$ for all $t \geq 0$. The *coincidence argument* then yields $z_- \in R_g$. Since $\phi_n(t, z_+) \in K \cup \{z_+\}$ for all $t \in [0, \tilde{T}_2)$, Fact 3.1.1 implies that $\phi_g(t, z_+) = \phi_n(t, z_+)$ for all $t \in [0, \tilde{T}_2]$. We can trace back the path to z_0 by observing that

$$\begin{aligned}\phi_g(t, z_-) &= \phi_g(t, \phi_g(\tilde{T}_3, \tilde{z}_2)) = \phi_g(t + \tilde{T}_3, \tilde{z}_2) = \phi_g(t + \tilde{T}_3, \phi_n(\tilde{T}_2, z_+)),\\ &= \phi_g(t + \tilde{T}_3, \phi_g(\tilde{T}_2, z_+)) = \phi_g(t + \tilde{T}_3 + \tilde{T}_2, z_+) = \phi_g(t + \tilde{T}_3 + \tilde{T}_2, \phi_g(T_1 + \tilde{T}_1, z_0)),\end{aligned}$$

for all $t \geq 0$. Since $z_- \in R_g$, we have $z_0 \in R_g$, as desired.

In similar manner, it can be shown that if $z_0 \in R_n \cap \bar{K}$ and the solution $\phi_n(t, z_0)$ enters K_- first, then $z_0 \in R_g$. ■

3.5.2 ROA Containment in Saturated Region

In this section, we show that the ROA containment also holds in the saturated region. What follows is a series of intermediate claims to arrive at the main result of this section, Proposition 3.5.12. The proofs of the intermediate claims are available in Section 3.11.4.

Define the line segments

$$\begin{aligned}\sigma_{+div} &:= \partial K_{+div} \cap \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} < \tfrac{bc}{ad} u_{\max}\}, \qquad \tilde{\sigma}_{+div} := \partial K_{+div} \setminus \sigma_{+div},\\ \sigma_{-div} &:= \partial K_{-div} \cap \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} > \tfrac{bc}{ad} u_{\min}\}, \qquad \tilde{\sigma}_{-div} := \partial K_{-div} \setminus \sigma_{-div}.\end{aligned} \tag{3.14}$$

It can be verified that these line segments are related to the saddle equilibrium points z_{eq+} and z_{eq-} of Σ_n in Claim 3.4.2 (and hence to the sets of equilibria Z_{eq+} and Z_{eq-} of Σ_g in Claim 3.4.3 as mentioned in Remark 3.5) by

$$\sigma_{+div} = l(z_+, z_{eq+}), \qquad \sigma_{-div} = l(z_-, z_{eq-}), \qquad z_{eq+} \in \tilde{\sigma}_{+div}, \qquad z_{eq-} \in \tilde{\sigma}_{-div},$$

whenever $ad < 0$. For the next result, recall the definition of transverse sections, Definition 3.1.

Claim 3.5.8 (Transverse Sections in Saturated Region)**.** *If the open-loop system is*

(i) marginally stable ($a = 0$), or strictly stable with a stable controller ($a < 0$ and $d \leq 0$), then ∂K_{+div} is a transverse section to f_n;

(ii) strictly stable with an unstable controller ($a < 0$ and $d \in (0, -a)$), or unstable ($a > 0$), then σ_{+div} ($\subset \partial K_{+div}$) is a transverse section to f_n.

Claim 3.5.9. *If the open-loop system is*

(i) strictly stable with an unstable controller ($a < 0$ and $d \in (0, -a)$); or

(ii) unstable ($a > 0$);

and $z_0 \in R_n$, then $z_0 \notin \tilde{\sigma}_{+div}$.

The next result states that when started from the saturated region K_{+out} within the ROA of the *nominal* system, i.e. $z_0 \in K_{+out} \cap R_n$, the solution of the GPAW-compensated system will always enter the unsaturated region.

Claim 3.5.10. *If $z_0 \in K_{+out} \cap R_n$, then there exists a $T_n \in (0, \infty)$ such that*

$$\phi_n(T_n, z_0) \in \partial K_{+in}, \qquad \phi_n(t, z_0) \in K_+, \forall t \in [0, T_n). \tag{3.15}$$

Moreover, there exists a $T_g < T_n$ such that the solution of the GPAW-compensated system satisfy

$$\phi_g(T_g, z_0) \in l(z_+, \phi_n(T_n, z_0)) \subset \partial K_{+in}.$$

Remark 3.12. A weaker version of Claim 3.5.10 (where the conclusion is that a $T_g \leq T_n$ exists such that $\phi_g(T_g, z_0) \in l(z_+, \phi_n(T_n, z_0)) \cup \{\phi_n(T_n, z_0)\}$) suffices for the purpose of proving Proposition 3.5.12. The proof (in pages 109 – 111) would have been shorter, as the condition $u_g(t) < u_n(t)$ for all $t \in (0, T]$ would be unnecessary. We present this marginally stronger result to confirm the intuitively reasonable conclusion. □

The following construction of $E(z_0)$ is analogous to the construction of $D(z_0)$ in Fig. 3-3. Consider a point $z_0 \in \partial K_{+in} \cap R_g$ in the ROA of the *GPAW-compensated system.* Theorem 2.5.3 shows that $\phi_g(t, z_0) \in \bar{K}$ for all $t \geq 0$. Recall the definition of σ_+ (see (3.12)) and let

$$t_{int} := \inf\{\bar{t} \in (0, \infty) \mid \phi_g(\bar{t}, z_0) \in \sigma_+\}.$$

In other words, t_{int} is the first time instant that the solution of the *GPAW-compensated system* $\phi_g(t, z_0)$ intersects σ_+, or ∞ if it does not intersect σ_+. Define the path $\gamma(z_0) \subset \mathbb{R}^2$ by

$$\gamma(z_0) := \begin{cases} \gamma_{int\phi_g}(z_0) \cup l(\phi_g(t_{int}, z_0), z_+) \cup \{z_+\} \cup l(z_0, z_+), & \text{if } t_{int} < \infty, \\ \gamma_{0\phi_g}(z_0) \cup \{z_{eq0}\} \cup \sigma_+ \cup l(z_0, z_+), & \text{otherwise,} \end{cases} \tag{3.16}$$

where

$$\gamma_{int\phi_g}(z_0) := \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_g(t, z_0), \forall t \in [0, t_{int}]\},$$
$$\gamma_{0\phi_g}(z_0) := \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_g(t, z_0), \forall t \geq 0\},$$

which can be verified to be closed and connected. Observe that $\gamma(z_0)$ traces the path along the solution $\phi_g(t, z_0)$ until it intersects σ_+ or reach the origin, proceeds along σ_+ towards z_+, then along ∂K_{+in} until it reaches its starting point z_0. Let the *open, bounded* region enclosed by $\gamma(z_0)$ be $E(z_0)$, and its closure be $\bar{E}(z_0)$. The region $E(z_0)$ is illustrated in Fig. 3-4.

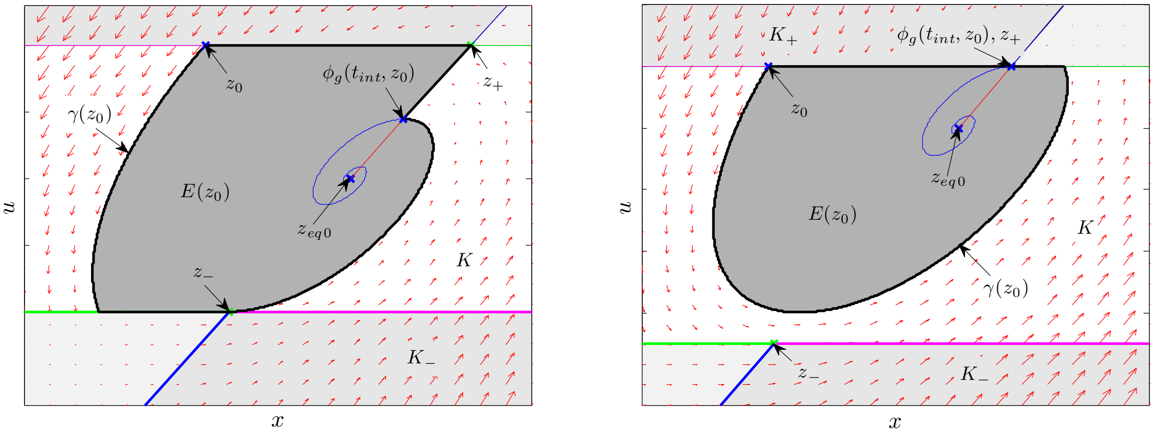


Figure 3-4: Closed path $\gamma(z_0)$ encloses region $E(z_0) \subset \bar{K}$. A case where the solution intersects ∂K_{-out}, then intersects σ_+, is shown on the left. A case where the solution intersects ∂K_{+out}, then intersects σ_+ at z_+, is shown on the right.

Remark 3.13. Observe that $z_+ \in \sigma_+$. If $t_{int} < \infty$ and $\phi_g(t_{int}, z_0) = z_+ \in \sigma_+$, then $\gamma(z_0)$ reduces to $\gamma(z_0) = \gamma_{int\phi_g}(z_0) \cup l(z_0, z_+)$, since $l(\phi_g(t_{int}, z_0), z_+) = l(z_+, z_+) = \{z_+\}$ and $z_+ \in \gamma_{int\phi_g}(z_0)$. This case is shown on the right plot of Fig. 3-4. A case analogous to the preceding is also possible for the path $\eta(z_0)$ in Fig. 3-3. However, it happens only for a *single* trajectory (if it exists). The peculiar nature of $\phi_g(t, z_0)$ implied by controller state-output consistency (Theorem 2.5.3) means that this case happens whenever $\phi_g(t, z_0)$ intersects ∂K_{+out} and slides along it to reach z_+. □

The following result in analogous to Claims 3.5.1 and 3.5.2 combined, with respect to $\bar{E}(z_0)$.

Claim 3.5.11 (Invariance of $\bar{E}(z_0) \subset R_g$). *If $z_0 \in \partial K_{+in} \cap R_g$, then $\bar{E}(z_0) \subset \bar{K}$ is a positively invariant set for system Σ_g. Moreover, $\bar{E}(z_0)$ is contained in the ROA of system Σ_g, and it must contain z_{eq0}, i.e. $z_{eq0} \in \bar{E}(z_0) \subset R_g$.*

The following is the main result of this section.

Proposition 3.5.12 (ROA Containment in Saturated Region)**.** *The part of the ROA of the origin of system Σ_n contained in $\mathbb{R}^2 \setminus \bar{K}$, is itself contained within the ROA of the origin of system Σ_g, i.e. $(R_n \cap (\mathbb{R}^2 \setminus \bar{K})) \subset R_g$.*

Proof. We need to show that if $z_0 \in R_n \cap (\mathbb{R}^2 \setminus \bar{K})$, then $z_0 \in R_g$. First, observe that $\mathbb{R}^2 \setminus \bar{K} = K_+ \cup K_-$, and $K_+ = K_{+out} \cup K_{+in} \cup \partial K_{+div}$. We will show that if $z_0 \in R_n \cap K_+$, then $z_0 \in R_g$. The proof where $z_0 \in R_n \cap K_-$ is similar. Let $z_0 \in R_n \cap K_+$. Since $z_0 \in R_n$ and $z_{eq0} \in K$, Fact 3.1.2 shows that $\phi_n(t, z_0)$ must intersect ∂K_{+in} at least once. Let T be the first time instant that $\phi_n(t, z_0)$ intersects ∂K_{+in}, so that $\phi_n(T, z_0) \in \partial K_{+in}$ and $\phi_n(t, z_0) \in K_+$ for all $t \in [0, T)$.

Consider when $z_0 \in R_n \cap (K_{+in} \cup \partial K_{+div}) \subset R_n \cap K_+$. We claim that $\phi_n(t, z_0)$ must be contained in $K_{+in} \cup \partial K_{+div}$ (and hence cannot enter K_{+out}) for all $t \in [0, T)$. Otherwise, $\phi_n(t, z_0)$ must intersect ∂K_{+div} at some finite time $\tilde{T} \in (0, T)$ and then pass into K_{+out}. Claims 3.5.8 and 3.5.9 show that $\phi_n(t, z_0)$ must pass through ∂K_{+div} or σ_{+div}, which are transverse sections. By similar reasoning as in the proof of Claim 3.5.10 (on pages 109 – 111), $\phi_n(t, z_0)$ can never return to K_{+in} during the interval $[\tilde{T}, T]$. In that case, $\phi_n(t, z_0)$ can never intersect ∂K_{+in} at $t = T$, which is a contradiction that establishes the immediate claim.

Since $\phi_n(t, z_0) \in K_{+in} \cup \partial K_{+div}$ for all $t \in [0, T)$, Fact 3.1.1 yields $\phi_g(t, z_0) = \phi_n(t, z_0)$ for all $t \in [0, T]$. Since $z_0 \in R_n$, we have $\phi_n(T, z_0) = \phi_g(T, z_0) \in R_n \cap \partial K_{+in} \subset R_n \cap \bar{K}$. Proposition 3.5.7 then shows that $\phi_g(T, z_0) \in R_g$, so that $z_0 \in R_g$, as desired.

Next, consider when $z_0 \in R_n \cap K_{+out} \subset R_n \cap K_+$. Claim 3.5.10 shows that there exists a $T_g \in (0, T)$ such that $\phi_g(T_g, z_0) \in l(z_+, \phi_n(T, z_0)) \subset \partial K_{+in}$. Since $z_0 \in R_n$ and $\phi_n(T, z_0) \in \partial K_{+in}$, we have $\phi_n(T, z_0) \in R_n \cap \partial K_{+in} \subset R_n \cap \bar{K}$. Proposition 3.5.7 then shows that $\phi_n(T, z_0) \in R_g$. Observing that $\phi_n(T, z_0) \in \partial K_{+in} \cap R_g$, Claim 3.5.11 shows that $l(z_+, \phi_n(T, z_0)) \subset \bar{E}(\phi_n(T, z_0)) \subset R_g$. Then $\phi_g(T_g, z_0) \in l(z_+, \phi_n(T, z_0)) \subset R_g$ implies $z_0 \in R_g$, as desired.

Finally, by observing that $(R_n \cap (K_{+in} \cup \partial K_{+div})) \cup (R_n \cap K_{+out}) = R_n \cap K_+$, the conclusion follows. ■

3.5.3 Main Result

The following is the main result of this chapter, which shows that the GPAW scheme can only maintain/enlarge the ROA of the uncompensated system. This shows that the GPAW scheme satisfies condition (i) of the general anti-windup problem (Problem 1 in Section 1.3).

Proposition 3.5.13 (ROA Containment). *The ROA of the origin of system Σ_n is contained within the ROA of the origin of system Σ_g, i.e. $R_n \subset R_g$.*

Proof. Propositions 3.5.7 and 3.5.12 gives

$$(R_n \cap \bar{K}) \subset R_g, \qquad (R_n \cap (\mathbb{R}^2 \setminus \bar{K})) \subset R_g,$$

respectively. The conclusion follows by taking the respective unions of both the left and right sides in the preceding, which gives

$$R_n = ((R_n \cap \bar{K}) \cup (R_n \cap (\mathbb{R}^2 \setminus \bar{K}))) \subset (R_g \cup R_g) = R_g. \qquad ■$$

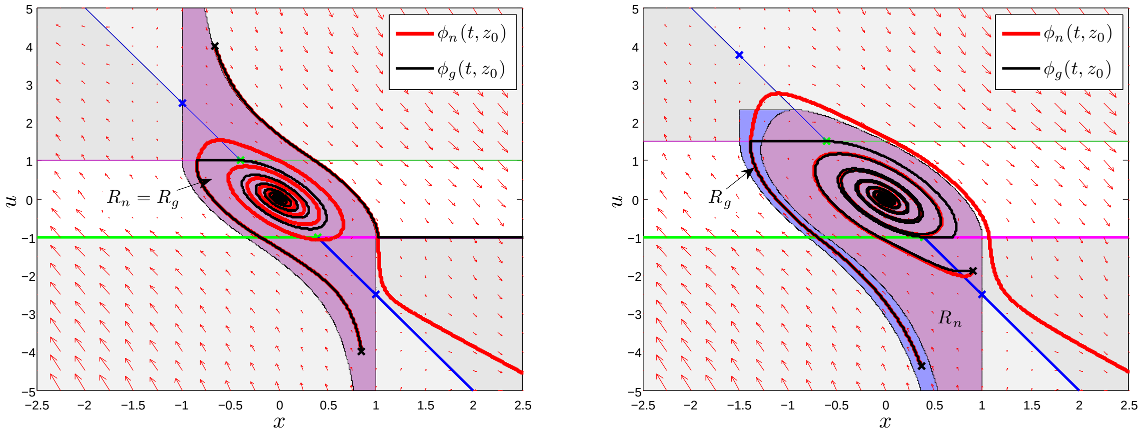
Remark 3.14. Observe that Proposition 3.5.13 is a strong result. It implies that for *every* Lyapunov function $V_n(z)$ that certifies an ROA $\tilde{R}_n \subset R_n$ for the nominal system Σ_n, there exists a Lyapunov function $V_g(z)$ (possibly $V_g(z) \equiv V_n(z)$) that certifies an ROA $\tilde{R}_g$ for the GPAW-compensated system Σ_g satisfying $\tilde{R}_n \subset \tilde{R}_g \subset R_g$. This is in contrast to conventional Lyapunov analysis that seeks a single *non-unique* Lyapunov function. In particular, this result implies that if there exists a Lyapunov function that certifies global asymptotic stability for the origin of the nominal system, i.e. $\tilde{R}_n = \mathbb{R}^2$, then there exists a Lyapunov function that also certifies global asymptotic stability for the origin of the GPAW-compensated system, i.e. $(\tilde{R}_g \supset \tilde{R}_n = \mathbb{R}^2) \Rightarrow (\tilde{R}_g = \mathbb{R}^2)$. See Corollary 3.7.2 for an example application. □

Remark 3.15. While Proposition 3.5.13 does not provide an estimate of the ROA, it can be *estimated* by numerous known methods, e.g. see [155–166]. We consider the estimation of

ROAs as a separate problem that need not be associated with the anti-windup problem, in contrast to [14]. □

3.5.4 Numerical Results

Here, we present some numerical results on the *exact* ROAs of systems Σ_n and Σ_g. As will be shown in Section 3.6, these results also indicate that *asymmetric* saturation constraints should be considered for any anti-windup scheme to be practically useful. The ROAs in these figures (Fig. 3-5, Fig. 3-6, and Fig. 3-7) are to be interpreted as *open* sets, since ROAs must be open [37, Lemma 8.1, p. 314].



(a) symmetric constraints, $u_{\max} = -u_{\min}$.

(b) asymmetric constraints, $u_{\max} > -u_{\min}$.

Figure 3-5: Region of attraction (ROA) containment for system with open-loop unstable plant (parameters: $a = 1$, $b = 1$, $c = -3$, $d = -1.2$). The vector field f_n is shown in the background, light purple regions represent R_n ($\subset R_g$), and light blue regions represent $R_g \setminus R_n$. In (a), the saturation limits are *symmetric* ($u_{\max} = -u_{\min} = 1$), resulting in $R_n = R_g$. Two pairs of solutions starting at $z_0 = (0.85, -4) \in R_n \cap R_g$ and $z_0 = (-0.66, 4) \notin R_n \cup R_g$ are included. In (b), the ROA containment $R_n \subset R_g$ of Proposition 3.5.13 holds *strictly*. The system is identical with the one in (a), except with *asymmetric* saturation limits ($u_{\max} = 1.5 > -u_{\min} = 1$). Two pairs of solutions starting from $z_0 = (0.9, -1.9) \in R_n \cap R_g$ and $z_0 = (0.37, -4.37) \in R_g \setminus R_n$ are included.

Fig. 3-5(a) shows the case where $R_n = R_g$ for a system with an open-loop unstable plant, together with two pairs of representative solutions, when the saturation constraints are symmetric, i.e. $u_{\max} = -u_{\min}$. When the same system is subjected to *asymmetric* saturation constraints, the ROAs are illustrated in Fig. 3-5(b). Clearly, the set containment $R_n \subset R_g$ result of Proposition 3.5.13 holds strictly. In Fig. 3-6, the ROAs are illustrated for a system with an open-loop stable plant and an unstable controller ($a < 0 < d < -a$). Again, the set containment $R_n \subset R_g$ is strict.

Remark 3.16. Note that the case of asymmetric saturation constraints is not pathological. Even with actuators having symmetric saturation constraints, it arises when regulating about an equilibrium point not lying in $\{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid \bar{u} = 0\}$, and the system state is transformed such that the resulting equilibrium lies at the origin. □

Remark 3.17. Observe from Fig. 3-6 that if we force the initial controller state to satisfy $u(0) = \text{sat}(u(0))$, e.g. by initializing the controller state to be $u(0) = \text{sat}(u_0)$ for any given

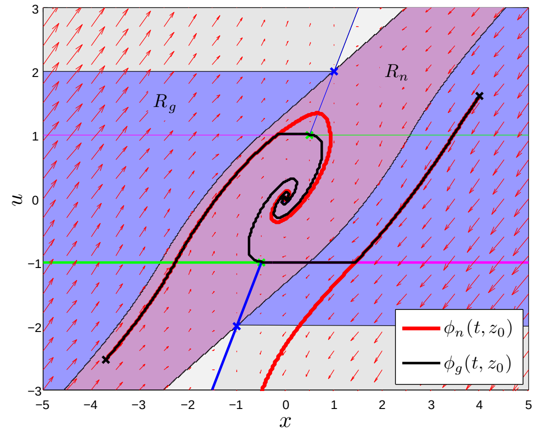


Figure 3-6: Region of attraction (ROA) containment for system with open-loop stable plant (parameters: $a = -1$, $b = 1$, $c = -1$, $d = 0.5$, $u_{\max} = -u_{\min} = 1$), which shows the ROA containment $R_n \subset R_g$ of Proposition 3.5.13 can hold *strictly*. The vector field f_n is shown in the background, light purple regions represent R_n ($\subset R_g$), and light blue regions represent $R_g \setminus R_n$. Two pairs of solutions starting from $z_0 = (-3.7, -2.54) \in R_n \cap R_g$ and $z_0 = (4, 1.6) \in R_g \setminus R_n$ are included.

nominal initialization u_0, then the effective ROA

$$R_{ge} := \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid z_0 = (\bar{x}, \text{sat}(\bar{u})), \phi_g(t, z_0) \to z_{eq0} \text{ as } t \to \infty\},$$

would be the entire state space, i.e. $R_{ge} = \mathbb{R}^2$, achieving global asymptotic stability for this case. □

Assumption 3.1 restricts the possible qualitative characteristics of the nominal controller. When the open-loop plant is unstable, i.e. $a > 0$, condition (3.6) restricts the nominal controller to be strictly stable, i.e. $d < -a < 0$. This case is covered in Fig. 3-5. When the open-loop plant is marginally or strictly stable, i.e. $a \leq 0$, the nominal controller can be either stable or unstable as long as $d < -a$ (due to condition (3.6)). The case of unstable nominal controllers, i.e. $0 < d < -a$ is covered in Fig. 3-6. The case where both the open-loop plant and nominal controller are marginally or strictly stable will be discussed in Section 3.7.

3.6 Illustration of the Need to Consider Asymmetric Saturation Constraints

The main purpose of the numerical results in Section 3.5.4 is to validate the ROA containment result $R_n \subset R_g$ of Proposition 3.5.13. Here, we discuss further implications of these results, and show the need to consider *asymmetric* saturation constraints for analysis of *general input-constrained systems* (see also Remark 3.16).

First, consider the following intuitively appealing statement.

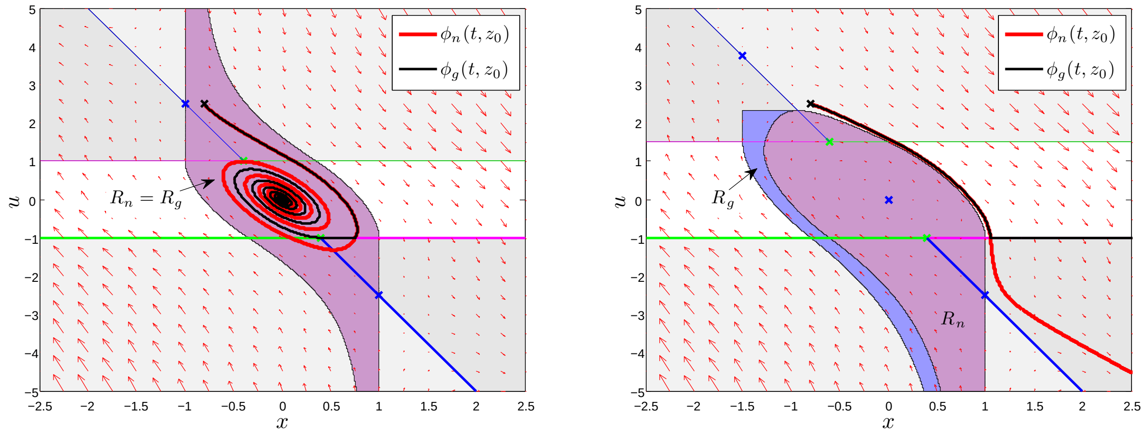
Statement 3.1 (Relaxed Constraints Imply ROA Enlargement)**.** *Consider a closed-loop autonomous system Σ_{aut} defined by the feedback interconnection of a saturated open-loop system and some controller. Let R_{aut1} be the ROA of some equilibrium point z_{eq} of system Σ_{aut} corresponding to some saturation limits $u_{\min1,i}, u_{\max1,i}$ (see* (1.2)*) satisfying $u_{\min1,i} <$*

$u_{\max1,i}$ *for all* $i \in \{1, 2, \dots, m\}$, *where* m *is the dimension of the control input. If some* $u_{\min2,i}, u_{\max2,i}$ *are such that*

$$u_{\min2,i} \leq u_{\min1,i} < u_{\max1,i} \leq u_{\max2,i}, \qquad \forall i \in \{1, 2, \dots, m\},$$

then the ROA R_{aut2} *of* z_{eq} *of system* Σ_{aut} *corresponding to these* relaxed *saturation limits,* $u_{\min2,i}, u_{\max2,i}$, *must contain* R_{aut1}, *i.e.* $R_{aut1} \subset R_{aut2}$.

Statement 3.1 suggests that *relaxing* the saturation constraints can only enlarge the associated ROA, and appeals to the intuitive notion that “things cannot get worse if we lessen the contributing factors that led to the original performance/stability problems”. If Statement 3.1 is true, then analysis of saturated feedback systems reduces to the analysis with the worst case saturation constraints (under perturbations) and provides a significant amount of simplifications. While there is evidence to suggest the truth of Statement 3.1 when the open-loop plant is not unstable (provided the controller is appropriately designed), the ROA results in Fig. 3-7 shows Statement 3.1 to be *false* in general, in particular when the open-loop plant is unstable. The ROAs shown in Fig. 3-7 are identical to those in Fig. 3-5, but with only a single pair of solutions starting from the point $z_0 = (-0.8, 2.5)$. As before, Fig. 3-7(b) shows the ROAs associated with a system *identical* to that in Fig. 3-7(a), except with *relaxed* but *asymmetric* saturation limits $u_{\max} \neq -u_{\min}$. It can be seen from Fig. 3-7(a) that with the original (symmetric) saturation limits, the point z_0 lies in both the ROAs of systems Σ_n and Σ_g. With the *relaxed* saturation limits, the point z_0 lies *outside* the ROAs of both systems Σ_n and Σ_g. Hence any sufficiently “tight” ROA estimate[9] obtained with symmetric saturation limits becomes invalid with asymmetric saturation limits, even when they are *relaxed*.



(a) symmetric constraints, $u_{\max} = -u_{\min}$.

(b) asymmetric constraints, $u_{\max} > -u_{\min}$.

Figure 3-7: Illustrating the need to consider asymmetric saturation constraints. The ROAs shown are identical to those in Fig. 3-5. In the left plot where the system is with symmetric saturation constraints ($u_{\max} = -u_{\min} = 1$), the point $z_0 = (-0.8, 2.5)$ lies in both ROAs, i.e. $z_0 \in R_n \cap R_g$. In the right plot representing an *identical* system with *relaxed* but *asymmetric* saturation constraints ($u_{\max} = 1.5 > -u_{\min} = 1$), the same point lies outside both ROAs, i.e. $z_0 \notin R_n \cup R_g$.

[9] By a “tight” ROA estimate, we mean one that is close to the exact ROA.

The vast majority of literature on input-constrained systems, including the anti-windup literature, consider only symmetric saturation constraints. One exception is [47], which is motivated by the fact that results obtained by the same authors for the symmetric saturation case do not carry over to the asymmetric saturation case. These results indicate a dire need to consider *asymmetric* saturation constraints for practical utility.

3.7 A Paradigm Shift in Anti-windup Compensation

Here, we propose a new way of addressing the general anti-windup problem, which has been reflected in the statement of Problem 1 (in Section 1.3) as conditions (i) and (iii). To aid in the subsequent discussion, we present the next result, which states that the *nominal uncompensated system* achieves global asymptotic stability (GAS) and local exponential stability (LES) when both the open-loop plant and nominal controller are marginally or strictly stable.

Claim 3.7.1 (Global Asymptotic Stability of Nominal System)**.** *If in addition to Assumption 3.1, both the open-loop plant and nominal controller are marginally or strictly stable* ($a \leq 0$ *and* $d \leq 0$)*, then the origin* z_{eq0} *of the nominal system* Σ_n *is globally asymptotically stable and locally exponentially stable.*

Remark 3.18. This is the main reason why this case is not considered in Section 3.5.4. □

Proof. The proof follows [38, Example 3.14, pp. 74 – 75] closely. First, the nominal system Σ_n (3.4) is governed by the ODEs

$$\begin{aligned} \dot{x} &= ax + b\,\mathrm{sat}(u), \\ \dot{u} &= cx + du, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} \ddot{u} &= c\dot{x} + d\dot{u} = c(ax + b\,\mathrm{sat}(u)) + d\dot{u} = a(\dot{u} - du) + bc\,\mathrm{sat}(u) + d\dot{u}, \\ &= (a+d)\dot{u} - adu + bc\,\mathrm{sat}(u). \end{aligned} \tag{3.17}$$

Consider the continuously differentiable function

$$V(u, \dot{u}) = \tfrac{1}{2}\dot{u}^2 + \tilde{V}(u), \qquad \tilde{V}(u) := \int_0^u (ad\tau - bc\,\mathrm{sat}(\tau))\,d\tau.$$

We will show that $V(u, \dot{u})$ is positive definite when $ad \geq 0$, which is implied by the assumption $a \leq 0$ and $d \leq 0$. Clearly, it is sufficient to show that $\tilde{V}(u)$ is positive definite. When $u_{\min} \leq u \leq u_{\max}$, we have $\mathrm{sat}(u) = u$, and

$$\tilde{V}(u) = \int_0^u (ad - bc)\tau\,d\tau = \tfrac{1}{2}(ad - bc)u^2,$$

so that from (3.7), $\tilde{V}(u) > 0$ for all $u \in [u_{\min}, u_{\max}] \setminus \{0\}$. Next, consider when $u =$

$\tilde{u} + u_{\max} > u_{\max}$ for some $\tilde{u} > 0$. Direct computation yields

$$\begin{aligned}
\tilde{V}(u) &= \tfrac{1}{2}adu^2 - \int_0^{u_{\max}} bc\tau \, d\tau - \int_{u_{\max}}^{u} bcu_{\max} \, d\tau, \\
&= \tfrac{1}{2}adu^2 - \tfrac{1}{2}bcu_{\max}^2 - bc\tilde{u}u_{\max}, \\
&= \tfrac{1}{2}ad(\tilde{u}^2 + 2\tilde{u}u_{\max} + u_{\max}^2) - \tfrac{1}{2}bcu_{\max}^2 - bc\tilde{u}u_{\max}, \\
&= \tfrac{1}{2}ad\tilde{u}^2 + (ad - bc)\tilde{u}u_{\max} + \tfrac{1}{2}(ad - bc)u_{\max}^2.
\end{aligned}$$

Clearly, when $ad \geq 0$, (3.7) implies $\tilde{V}(u) > 0$ for all $u > u_{\max}$. The case when $u < u_{\min}$ can be shown similarly. Hence $V(u, \dot{u})$ is positive definite. The above expressions also show that $V(u, \dot{u})$ is radially unbounded.

Taking the time derivative of $V(u, \dot{u})$ and using (3.17) yields

$$\begin{aligned}
\dot{V}(u, \dot{u}) &= \dot{u}\ddot{u} + (adu - bc\,\mathrm{sat}(u))\dot{u} = \dot{u}((a + d)\dot{u} - adu + bc\,\mathrm{sat}(u)) + (adu - bc\,\mathrm{sat}(u))\dot{u}, \\
&= (a + d)\dot{u}^2.
\end{aligned}$$

Condition (3.6) then shows $\dot{V}(u, \dot{u})$ to be negative semidefinite, i.e. $\dot{V}(u, \dot{u}) \leq 0$.

To complete the proof for global asymptotic stability, it is sufficient to show that $\dot{V}(u, \dot{u}) \equiv 0$ implies $\dot{u} \equiv 0$ and $u \equiv 0$. The first condition is obtained immediately due to (3.6) (specifically, $a + d \neq 0$). When $\dot{u} \equiv 0$, (3.17) reduces to

$$\ddot{u} = -adu + bc\,\mathrm{sat}(u),$$

so that by (3.7), $\ddot{u} \neq 0$ as long as $u \neq 0$. Hence only the trivial solution $u \equiv 0$, $\dot{u} \equiv 0$ can stay identically in the set $S = \{(u, \dot{u}) \in \mathbb{R}^2 \mid \dot{V}(u, \dot{u}) = 0\}$. By [37, Corollary 4.2, p. 129], the origin of Σ_n is globally asymptotically stable. Local exponential stability of the origin follows immediately from Assumption 3.1. ■

Remark 3.19. Observe that (3.6) precludes $ad \geq 0$ being satisfied when either the open-loop plant or nominal controller is unstable, i.e. $a > 0$ or $d > 0$. □

Corollary 3.7.2 (Global Asymptotic Stability of GPAW-Compensated System)**.** *If in addition to Assumption 3.1, both the open-loop plant and nominal controller are marginally or strictly stable ($a \leq 0$ and $d \leq 0$), then the origin z_{eq0} of the GPAW-compensated system Σ_g is globally asymptotically stable and locally exponentially stable.*

Proof. Claim 3.7.1 shows that the origin z_{eq0} is globally asymptotically stable for system Σ_n, which implies $R_n = \mathbb{R}^2$. Proposition 3.5.13 then yields $R_g \supset R_n = \mathbb{R}^2$, which implies $R_g = \mathbb{R}^2$ and the origin z_{eq0} is globally asymptotically stable for system Σ_g. Local exponential stability of the origin follows immediately from Assumption 3.1. ■

Numerous results in the anti-windup literature are of the form of Corollary 3.7.2, i.e. under some assumptions and applying some anti-windup method, some stability properties are achieved. Such results *sound* impressive, and may indeed give some confidence in the application of the particular anti-windup method. However, we argue that it may not reveal any advantages of the anti-windup method. First, observe that for any meaningful anti-windup problem, local stability must be assumed. Otherwise, the anti-windup problem is ill-posed. Any results asserting local stability are only restating the assumption. Observe from Claim 3.7.1 that the *uncompensated* nominal system achieves GAS. In other words, GAS is achieved *without any anti-windup compensation.* While Corollary 3.7.2 asserts GAS, it tells *nothing* of any advantages gained by adopting the particular anti-windup method.

In contrast, the ROA containment result of Proposition 3.5.13 truly reflects an advantage of the GPAW scheme, namely, that the ROA of the system will always be maintained/enlarged by its application. As such, we propose this new paradigm to address the anti-windup problem, i.e. results on the anti-windup compensated system *relative* to the uncompensated system. This is reflected as conditions (i) and (iii) of Problem 1 (in Section 1.3).

3.8 Relation to the General Anti-windup Problem

Here, we discuss how the obtained results for the GPAW scheme relate to the general anti-windup problem (Problem 1 of Section 1.3) for the nominal system (3.4). First, observe that condition (i) of Problem 1 is fulfilled by the GPAW scheme due to Proposition 3.5.13. Condition (ii) is fulfilled by construction (see Remark 2.17 and item (ii) of the concluding observations in Section 2.5 on page 52). Numerous performance metrics can be specified to show that application of GPAW compensation on system (3.4) will maintain/improve performance, hence fulfilling condition (iii) of Problem 1. However, this is non-trivial (see Remark 3.20), and similar arguments as in the proofs of Propositions 3.5.7 and 3.5.12 would be involved in general.

For an informal discussion, consider the L_2 norm [41, p. 2] of the state[10]

$$J = \|z\|_2 = \|(x,u)\|_2 := \left(\int_0^\infty V(x(t),u(t))\,dt\right)^{\frac{1}{2}}, \qquad V(z) := V(x,u) := x^2 + u^2, \quad (3.18)$$

with a lower J indicating greater performance. Let the performance attained by systems Σ_g and Σ_n be

$$J_g = \left(\int_0^\infty V(x_g(t),u_g(t))\,dt\right)^{\frac{1}{2}}, \qquad J_n = \left(\int_0^\infty V(x_n(t),u_n(t))\,dt\right)^{\frac{1}{2}},$$

respectively, where $(x_g(t),u_g(t)) = \phi_g(t,z_0)$ and $(x_n(t),u_n(t)) = \phi_n(t,z_0)$ are their respective solutions. The function $V(x,u)$ is the square of the magnitude of the state from the origin. It can be seen from Figs. 3-5, 3-6, and 3-7, that when started from the same point z_0, the state of the GPAW-compensated system Σ_g for these cases are in some sense smaller in magnitude than the state of the nominal system Σ_n, so that $J_g \leq J_n$ likely holds. We leave the actual proofs as future work (see Section 7.1.8).

Remark 3.20. It can be shown that for $V(z)$ as defined in (3.18),

$$\frac{\partial V(z)}{\partial z} f_g(z) \leq \frac{\partial V(z)}{\partial z} f_n(z), \qquad \forall z \in \mathbb{R}^2,$$

holds, where f_g and f_n are the vector fields of the GPAW-compensated system Σ_g (3.5) and nominal system Σ_n (3.4) respectively. We note that this is *not* a sufficient condition to

[10] Note that the use of the L_2 performance measure differs from that in [14]. Moreover, the finite L_2 gain performance measure in [14] is not meaningful for autonomous systems that are not driven by any exogenous inputs.

show that $V(x_g(t), u_g(t)) \leq V(x_n(t), u_n(t))$ for all $t \geq 0$. However, *if*

$$\frac{\partial V(\phi_g(t,z_0))}{\partial z} f_g(\phi_g(t,z_0)) \leq \frac{\partial V(\phi_n(t,z_0))}{\partial z} f_n(\phi_n(t,z_0)), \qquad \forall t \geq 0, \forall z_0 \in \mathbb{R}^2,$$

holds, then $V(\phi_g(t,z_0)) = V(x_g(t), u_g(t)) \leq V(x_n(t), u_n(t)) = V(\phi_n(t,z_0))$ for all $t \geq 0$ follows easily, which implies $J_g \leq J_n$. The need to evaluation the function $V(z)$ along two *different* trajectories renders it a non-trivial task. □

In summary, when GPAW compensation is applied to the nominal system (3.4), (i) nominal performance is recovered whenever no controls saturate; and (ii) stability is never compromised. Observe that these are achieved for *all* plant and controller parameters $(a, b, c, d, u_{\max}, u_{\min})$ with $u_{\min} < u_{\max}$ and (a, b, c, d) satisfying the standard anti-windup assumption (Assumption 3.1).

3.9 Solution Bounce Property and Some Conjectures

Here, we show that the solution of the GPAW-compensated system Σ_g can *bounce* (defined more concretely in Definition 3.3 below) on the saturation constraint boundaries ∂K_+ and ∂K_- at most once. This indicates the possible existence of some optimality properties, which we state as a conjecture. Further optimality conjectures are also presented, the verification of which we leave as future work (see Section 7.1.8). First, we define a *solution bounce*.

Definition 3.3 (Solution Bounce)**.** A solution $\phi(t, z_0)$ is said to *bounce* on the saturation constraint boundary ∂K_+ (3.8) if there exists an $\epsilon > 0$ and an interval $[t_1, t_2]$, $t_1 \leq t_2$, such that

$$\begin{aligned} &\phi(t,z_0) \in (\partial K_+ \cup K_+), && \forall t \in [t_1, t_2], \\ &\phi(t,z_0) \in K, && \forall t \in [t_1 - \epsilon, t_1) \cup (t_2, t_2 + \epsilon], \end{aligned}$$

with an analogous counterpart for the saturation constraint boundary ∂K_-. □

In other words, a *solution bounce* is an event where the solution starting from the *interior* of the unsaturated region $\bar{K}$ enters the region $\mathbb{R}^2 \setminus K$ (which contains the saturated region $K_+ \cup K_-$ and constraint boundaries $\partial K_+ \cup \partial K_-$) for some finite interval (or single time instant when $t_1 = t_2$), and returns to the interior K thereafter. Observe that Definition 3.3 excludes those instances where the solution *starts* from $\mathbb{R}^2 \setminus K$ and enters K. If $t_1 = t_2$ in Definition 3.3, the solution intersects the saturation constraint boundary only at a single point in time.

The following gives an upper bound on the number of solution bounces for the GPAW-compensated system Σ_g. This property depends critically on the controller state-output consistency property (Theorem 2.5.3) unique to the GPAW scheme.

Proposition 3.9.1 (Upper Bound on Number of Solution Bounces for Σ_g)**.** *For all $z_0 \in R_g$, the solution of the GPAW-compensated system $\phi_g(t, z_0)$ can have at most one bounce each on ∂K_+ and ∂K_-.*

Proof. We will prove the case when the solution bounces on ∂K_+. The proof when the solution bounces on ∂K_- is similar. Assume for the sake of contradiction that for some

$z_0 \in R_g$, the solution $\phi_g(t, z_0)$ has two or more bounces on ∂K_+. Then there exists $\epsilon_1 > 0$, $\epsilon_2 > 0$, and two intervals $[t_1, t_2]$, $[t_3, t_4]$ such that

$$\begin{aligned}
&\phi_g(t, z_0) \in (\partial K_+ \cup K_+), && \forall t \in [t_1, t_2] \cup [t_3, t_4],\\
&\phi_g(t, z_0) \in K, && \forall t \in [t_1 - \epsilon_1, t_1) \cup (t_2, t_2 + \epsilon_1],\\
&\phi_g(t, z_0) \in K, && \forall t \in [t_3 - \epsilon_2, t_3) \cup (t_4, t_4 + \epsilon_2].
\end{aligned}$$

Observing that $K \cap (\partial K_+ \cup K_+) = \emptyset$ (see (3.8)), the preceding implies either $t_2 < t_3$ or $t_4 < t_1$. Without loss of generality, assume that $t_2 < t_3$. From the definitions of f_g (3.5) and ∂K_{+out} (3.8), the point where the solution leaves ∂K_+ must lie in $\partial K_+ \setminus \partial K_{+out} = (\partial K_{+in} \cup \{z_+\})$, i.e. $\phi_g(t_2, z_0), \phi_g(t_4, z_0) \in \partial K_{+in} \cup \{z_+\}$. Theorem 2.5.3 shows that $\phi_g(t, z_0) \notin K_+$ for all $t \in [t_1, t_2] \cup [t_3, t_4]$, which, together with Fact 3.1.2 shows that the solution cannot leave $\partial K_+ \cup K_+$ through ∂K_{+in}. Hence $\phi_g(t_2, z_0) = \phi_g(t_4, z_0) = z_+$. The solution in the interval $[t_2, t_4]$ then forms a closed orbit, $\tilde{\gamma} = \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_g(t, z_0), \forall t \in [t_2, t_4]\}$. By the uniqueness of solutions (see Proposition 3.3.3 and Remark 3.3), we have $\phi_g(t, z_0) \in \tilde{\gamma}$ for all $t \geq t_4$ and $\lim_{t\to\infty} \phi_g(t, z_0) \neq z_{eq0}$. Hence $z_0 \notin R_g$, a contradiction. ■

Remark 3.21. It is clear from the proof that if the solution $\phi_g(t, z_0)$ intersects ∂K_+ at a single point in time, then it must intersect it at z_+. This is also clear from the definitions of ∂K_{+out}, ∂K_{+in}, and z_+ in (3.8). □

Proposition 3.9.1 suggests an intuitively clear optimality property of the GPAW-compensated system, namely that for system Σ_n, the GPAW scheme achieves the *minimal* number of solution bounces. In other words, no other anti-windup schemes can do better. We state this below, whose proof we leave as future work (see Section 7.1.8).

Conjecture 3.1 (Optimality in Number of Solution Bounces)**.** *The GPAW scheme applied to system* Σ_n (3.4) *is an optimal anti-windup scheme achieving the least number of solution bounces on* ∂K_+ *and* ∂K_- *for all initial conditions in the ROA of the* uncompensated *system* Σ_n*, i.e. for all* $z_0 \in R_n$.

Remark 3.22. We have stated the conjecture in terms of initial conditions in the ROA of the *nominal* system to provide a fair basis of comparison. Different anti-windup schemes may achieve different ROAs, but they all must contain R_n according to our criterion for a valid anti-windup scheme. Let R_{aw1}, R_{aw2} and $\phi_{aw1}(t, z_0), \phi_{aw2}(t, z_0)$ be the ROAs and solutions corresponding to two different anti-windup compensated (closed-loop) systems. If $\partial K_+ \cap R_{aw1} = \emptyset$ and $\partial K_+ \cap R_{aw2} \neq \emptyset$, then $\phi_{aw1}(t, z_0)$ may not bounce (it may tend to infinity without returning to K), while $\phi_{aw2}(t, z_0)$ will bounce, if started at a point within $R_{aw2} \setminus R_{aw1}$ sufficiently close to ∂K_+. However, this would not be a fair comparison, and hence the restriction to initial conditions within $R_n \subset (R_{aw1} \cap R_{aw2})$. □

To see that Conjecture 3.1 is at least plausible, assume that $z_0 \in K \cap R_n$ is such that the nominal solution $\phi_n(t, z_0)$ will intersect ∂K_+ at some time. Let $\phi_{aw}(t, z_0)$ be the solution of some anti-windup compensated system, which must satisfy $\phi_{aw}(t, z_0) = \phi_n(t, z_0)$ for all t such that $\phi_n(t, z_0) \in K$ (to recover nominal performance in the absence of saturation). Hence $\phi_{aw}(t, z_0)$ must bounce on ∂K_+ at least once, which is the maximum achieved by GPAW compensation due to Proposition 3.9.1. This is sufficient to prove Conjecture 3.1 if the anti-windup compensated system is second order. To prove the conjecture, we need to show this for anti-windup compensated systems of arbitrary (finite) order, e.g. when using dynamic anti-windup schemes, which we leave as future work (see Section 7.1.8).

The next conjecture is motivated by Proposition 3.5.13, Fig. 3-6, and Remark 3.17.

Conjecture 3.2 (Optimality in ROA)**.** *Consider the GPAW scheme applied to system* Σ_n (3.4) *with the controller state initialization* $u(0) = \text{sat}(u_0)$*, where* u_0 *is some nominal initialization. The effective ROA of this GPAW-compensated system defined by*

$$R_{gi} := \{(\bar{x}, \bar{u}) \in \mathbb{R}^2 \mid z_0 = (\bar{x}, \text{sat}(\bar{u})), \phi_g(t, z_0) \to z_{eq0} \textit{ as } t \to \infty\},$$

is the largest possible ROA for any anti-windup scheme. In other words, the GPAW scheme with this initialization is an optimal anti-windup scheme achieving the largest ROA.

The conjecture states that if R_{aw} is the achieved ROA for some anti-windup scheme, then $R_{aw} \subset R_{gi}$. To see that this is at least plausible, let Σ_{aw} and R_{aw} be the anti-windup compensated system and ROA obtained by some anti-windup scheme. Now, let Σ_{gi} and R_{gi} be the system and ROA obtained by *applying GPAW compensation on* Σ_{aw} (with the controller state initialization $u(0) = \text{sat}(u_0)$). Proposition 3.5.13 then yields $R_{aw} \subset R_{gi}$. Now, observe that any anti-windup scheme can modify the nominal vector field f_n only in the saturated region $K_+ \cup K_- \cup \partial K_+ \cup \partial K_-$ due to the necessity of recovering nominal performance in the absence of saturation. With the controller state initialization $u(0) = \text{sat}(u_0)$, *all solutions of* Σ_{gi} *starts within the unsaturated region.* Hence only the part of the vector field f_g within the unsaturated region $\bar{K}$ is of concern. Moreover, the GPAW scheme overrides any modification of the nominal vector field on the saturation constraint boundaries $\partial K_+ \cup \partial K_-$, cancelling any effect of the prior anti-windup scheme. Note that if some anti-windup scheme is applied on the GPAW-compensated controller and driven by the signal $\text{sat}(u) - u$, then it will be *disabled.* This is due to $u(0) = \text{sat}(u_0)$ and Theorem 2.5.3, yielding $\text{sat}(u) - u \equiv 0$ (see also Remark 2.21).

The preceding is a plausibility argument for comparing the GPAW scheme against another anti-windup method. To prove the conjecture, we first need to show that the controller state initialization $u(0) = \text{sat}(u_0)$ forcing the state to the unsaturated region, will not induce instability, i.e. if $(x_0, u_0) \in R_g$, then $(x_0, \text{sat}(u_0)) \in R_g$. The comparison against dynamic anti-windup schemes is also necessary to prove Conjecture 3.2. This will involve defining some ways to compare ROAs in different dimensions. For example, in $\mathbb{R}^n$, we can define the ROA of the second order GPAW-compensated system as $R_{gi} \times \mathbb{R}^{n-2}$ for the purpose of ROA comparison.

3.10 Chapter Summary

We applied the GPAW scheme to a first order input-constrained LTI plant driven by a first order LTI controller, where the objective is to regulate the system state about the origin. Existence and uniqueness of solutions to the GPAW-compensated system are assured using results from the projected dynamical systems literature, and equilibria are characterized. The main result of this chapter is that GPAW compensation applied to this simple system can only maintain/enlarge the system's region of attraction. Numerical results indicate a need to consider *asymmetric* saturation constraints for *general input-constrained systems.* The weaknesses of some qualitative results on anti-windup methods are illustrated, which motivated a new paradigm for addressing the anti-windup problem. We discuss how the results in this chapter relate to the general anti-windup problem (Problem 1 of Section 1.3), and presented the solution bounce property together with some conjectures.

The results in this chapter, while limited in applicability, reveals some attractive features of the GPAW scheme when restricted to this simple system. These results are strong, and are valid for all plant and controller parameters satisfying the standard anti-windup assumption of unconstrained nominal stability. In the remainder of this dissertation, we develop tools to enable the extension of these results to more general plants and/or controllers.

3.11 Chapter Supplement

Here, we collect some supplementary material for this chapter. The *strict* Comparison Lemma presented in Section 3.11.2 may be of general interest.

3.11.1 Translating Some Logical Statements

In some of the proofs in this chapter, e.g. Claim 3.5.4, we need to assert the truth of statements of the form

$$\text{“if } z \in \alpha \text{ and } z \in \beta, \text{ then } z \in \gamma\text{”.} \tag{3.19}$$

Here, we show explicitly that this statement is equivalent to

$$\text{“if } z \in \alpha \setminus \gamma, \text{ then } z \notin \beta\text{”.} \tag{3.20}$$

Note that $\neg$, $\wedge$, $\vee$, $\Rightarrow$, and $\Leftrightarrow$, represent logical *negation* (NOT operator), *conjunction* (AND operator), *disjunction* (OR operator), *implication*, and *equivalence* respectively. Let

$$A \Leftrightarrow (z \in \alpha), \qquad B \Leftrightarrow (z \in \beta), \qquad C \Leftrightarrow (z \in \gamma),$$

so that the original statement (3.19) is equivalent to $(A \wedge B) \Rightarrow C$. Using the equivalence $(A \Rightarrow B) \Leftrightarrow (\neg A \vee B)$ [167, Fig. 7.11, p. 210], the preceding can be rewritten as

$$\begin{aligned}(A \wedge B) \Rightarrow C \Leftrightarrow \neg(A \wedge B) \vee C \Leftrightarrow \neg A \vee \neg B \vee C \Leftrightarrow \neg A \vee C \vee \neg B,\\ \Leftrightarrow \neg(A \wedge \neg C) \vee \neg B \Leftrightarrow (A \wedge \neg C) \Rightarrow \neg B.\end{aligned}$$

In other words, the original statement is equivalent to

$$\text{“if } z \in \alpha \text{ and } z \notin \gamma, \text{ then } z \notin \beta\text{”,}$$

or more compactly as (3.20).

Moreover, observe that we can always replace A by more complex statements to get an analogous equivalence relation. For example, if $A \Leftrightarrow (D \vee E) \wedge F$, then

$$((D \vee E) \wedge F \wedge B \Rightarrow C) \Leftrightarrow ((D \vee E) \wedge F \wedge \neg C \Rightarrow \neg B).$$

In fact, the more complex form is encountered more often.

3.11.2 A Variant of the Comparison Lemma

Here, we present a variant of the Comparison Lemma [37, Lemma 3.4, pp. 102 – 103], where the conclusion results in a *strict* inequality. It is a direct consequence of uniqueness of solutions of the *scalar* differential equation, with an application of the original Comparison Lemma.

Lemma 3.11.1 (Strict Comparison Lemma)**.** *Consider the scalar differential equation*

$$\dot{u} = f(t, u), \qquad u(t_0) = u_0, \tag{3.21}$$

where $f(t, u)$ is continuous in t and locally Lipschitz in u, for all $t \geq t_0$ and all $u \in J \subset \mathbb{R}$, and where J is a connected *interval. Let $[t_0, T)$ (T could be infinity) be the maximal interval of existence of the solution $u(t)$, and suppose $u(t) \in J$ for all $t \in [t_0, T)$. Let $v(t)$ be a continuous function whose upper right-hand derivative $D^+v(t)$ satisfies the differential inequality*

$$D^+v(t) \leq f(t, v(t)), \qquad v(t_0) < v_0, \tag{3.22}$$

where $v(t) \in J$ for all $t \in [t_0, T)$. Then $v(t) < u(t)$ holds for all $t \in [t_0, T)$.

Remark 3.23. Observe that the fundamental qualitative difference with [37, Lemma 3.4, pp. 102 – 103] is the *strict* inequality of the initial condition $v(t_0) < u_0$, and the conclusion $v(t) < u(t)$ for all $t \in [t_0, T)$. The requirement of J being *connected* is purely technical, as seen in the proof. See also [37, Appendix C.2, pp. 659 – 660] for a definition of the upper right-hand derivative $D^+v(t)$. □

Proof. Consider the initial value problem

$$\dot{w} = f(t, w), \qquad w(t_0) = v(t_0) < u_0. \tag{3.23}$$

With the assumptions, [37, Theorem 3.1, pp. 88 – 89] implies existence and uniqueness of solutions of (3.21) and (3.23). Let $[t_0, T_w)$ be the maximal interval of existence of the solution $w(t)$ such that $w(t) \in J$ for all $t \in [t_0, T_w)$. Define $\tilde{T} := \min\{T, T_w\}$.

We claim that $w(t) \neq u(t)$ for all $t \in [t_0, \tilde{T})$ (due to $w(t_0) \neq u(t_0)$). Otherwise, there exists a $\hat{T} \in [t_0, \tilde{T})$ such that $w(\hat{T}) = u(\hat{T})$. By solving (3.21) and (3.23) *backwards in time* from $t = \hat{T}$ to $t = t_0$, we obtain $w(t_0) = u(t_0)$ due to uniqueness of solutions. This contradicts $w(t_0) \neq u(t_0)$ and establishes the claim.

Since $w(t_0) < u(t_0)$, and $w(t) \neq u(t)$ for all $t \in [t_0, \tilde{T})$, continuity of both $w(t)$ and $u(t)$ shows that $w(t) < u(t)$ holds with *strict* inequality for all $t \in [t_0, \tilde{T})$. The Comparison Lemma [37, Lemma 3.4, pp. 102 – 103] applied to (3.23) and the differential inequality (3.22) yields $v(t) \leq w(t)$ for all $t \in [t_0, \tilde{T})$. Then we have $v(t) \leq w(t) < u(t)$ for all $t \in [t_0, \tilde{T})$. This, together with the connectivity of J and the condition $u(t), v(t) \in J$ for all $t \in [t_0, T)$ implies $w(t) \in J$ for all $t \in [t_0, T)$, i.e. $T_w \geq T$ and $\tilde{T} = T$. Hence the conclusion $v(t) < u(t)$ holds with *strict* inequality for all $t \in [t_0, T)$. ■

3.11.3 Proofs of Intermediate Results for Section 3.5.1

Proof of Claim 3.5.1 (Invariance of $\bar{D}(z_0)$)

Let

$$\tilde{\sigma}_+ := \begin{cases} l(\phi_n(t_{int}, z_0), z_+) \cup \{z_+\}, & \text{if } t_{int} < \infty, \\ \sigma_+, & \text{otherwise.} \end{cases}$$

We first show that $\tilde{\sigma}_+$ is a transverse section to f_n, and that f_n always points into $\bar{D}(z_0)$ on $\tilde{\sigma}_+$. Let $\alpha \in \{-1, +1\}$ be chosen such that $\langle \alpha \tilde{T} z_+, z_0 - z_+ \rangle > 0$, where $\tilde{T} z_+ := (u_{\max}, \frac{d}{c} u_{\max})$ is orthogonal to z_+ (see (3.8)). Then $\alpha \tilde{T} \frac{z_+}{\|z_+\|}$ is the unit normal of $\tilde{\sigma}_+$ that points into $\bar{D}(z_0)$. Hence $\tilde{\sigma}_+$ is a transverse section to f_n, and f_n points into $\bar{D}(z_0)$ on $\tilde{\sigma}_+$ if and only if $\langle \alpha \tilde{T} z_+, f_n(z) \rangle > 0$ holds with *strict* inequality for all $z \in \tilde{\sigma}_+$.

Since $z_0 \in \partial K_{+in}$, we have from the definition of ∂K_{+in} (3.8) that $z_0 = (x_0, u_{\max})$ for some x_0 that satisfies $cx_0 + du_{\max} < 0$. Then $z_0 - z_+ = (x_0 + \frac{d}{c}u_{\max}, 0)$. Due to $cx_0 + du_{\max} < 0$, the condition

$$\langle \alpha\tilde{T}z_+, z_0 - z_+\rangle = \alpha\langle (u_{\max}, \tfrac{d}{c}u_{\max}), (x_0 + \tfrac{d}{c}u_{\max}, 0)\rangle = \tfrac{\alpha}{c}u_{\max}(cx_0 + du_{\max}) > 0,$$

can hold only if $\alpha = -\operatorname{sgn}(c)$. From the definition of $\tilde{\sigma}_+$, any $z \in \tilde{\sigma}_+$ has the form $z = (-\theta\frac{d}{c}u_{\max}, \theta u_{\max})$ for some $\theta \in (0, 1]$, so that $f_n(z) = ((b - \frac{ad}{c})\theta u_{\max}, 0)$ on $\tilde{\sigma}_+$ (see (3.4)). Using the definition of f_n on $\tilde{\sigma}_+$, we have

$$\begin{aligned}\langle \alpha\tilde{T}z_+, f_n(z)\rangle &= \alpha\langle (u_{\max}, \tfrac{d}{c}u_{\max}), ((b - \tfrac{ad}{c})\theta u_{\max}, 0)\rangle, \\ &= -\operatorname{sgn}(c)(b - \tfrac{ad}{c})\theta u_{\max}^2 = \tfrac{ad-bc}{|c|}\theta u_{\max}^2.\end{aligned}$$

Since $\theta > 0$ for any $z \in \tilde{\sigma}_+$, we have from (3.7) that $\langle \alpha\tilde{T}z_+, f_n(z)\rangle > 0$, which shows that $\tilde{\sigma}_+$ is a transverse section to f_n and that f_n always points into $\bar{D}(z_0)$ on $\tilde{\sigma}_+$.

It is clear that $l(z_0, z_+) \subset \partial K_{+in}$ is also a transverse section to f_n, and that f_n always points into $\bar{D}(z_0)$ on $l(z_0, z_+)$. Both of these results show that any solution originating in $\bar{D}(z_0)$ cannot exit $\bar{D}(z_0)$ through the line segments $\tilde{\sigma}_+$ or $l(z_0, z_+)$. Furthermore, since the solution is unique and no two different paths can intersect [135, pp. 38], the region $\bar{D}(z_0)$ enclosed by $\eta(z_0)$ must be a *positively invariant set* [135, pp. 47] for system Σ_n. The assumption $\phi_n(t, z_0) \in \bar{K} \cup K_-$ for all $t \geq 0$ implies $\eta(z_0) \subset \bar{K} \cup K_-$, and hence $\bar{D}(z_0) \subset \bar{K} \cup K_-$.

Finally, from the assumption $z_0 \in R_n$, we have $\phi_n(t, z_0) \to z_{eq0}$ as $t \to \infty$. Since $\bar{D}(z_0)$ is a positively invariant set and $z_0 \in \bar{D}(z_0)$, we have $\phi_n(t, z_0) \in \bar{D}(z_0)$ for all $t \geq 0$. The conclusion $z_{eq0} \in \bar{D}(z_0)$ then follows from the fact that $\bar{D}(z_0)$ is *closed* and hence contains all its limit points. ■

Proof of Claim 3.5.2

Since $\bar{K} \subset (\bar{K} \cup K_-)$, the hypotheses of Claim 3.5.1 are satisfied. Claim 3.5.1 shows that $\bar{D}(z_0)$ is a positively invariant set. The condition $\phi_n(t, z_0) \in \bar{K}$ for all $t \geq 0$ implies $\bar{D}(z_0) \subset \bar{K}$. It was shown in [136, §VI.2, pp. 353 – 363], [135, Theorem 1.3, p. 55] that for *planar* dynamic systems with only a countable number of equilibria and with unique solutions, the ω limit set of any trajectory contained in any bounded region can only be of three types: equilibrium points, closed orbits, or *heteroclinic/homoclinic orbits* [168, p. 45], which are unions of saddle points and the trajectories connecting them. It follows from Claims 3.4.1 and 3.4.2 that the origin z_{eq0} is the *only* equilibrium point of Σ_n in $\bar{K}$, which must be a stable node or stable focus. Hence the ω limit set of any trajectory contained in $\bar{D}(z_0) \subset \bar{K}$ cannot be heteroclinic/homoclinic orbits. By Bendixson's Criterion [37, Lemma 2.2, p. 67] and (3.6), the simply connected[11] region $\bar{D}(z_0)$ contains no closed orbits. As a result, the ω limit sets must consist of equilibrium points only, and it must be z_{eq0} since it is the only equilibrium point in $\bar{K}$. The conclusion follows by observing that $\bar{D}(z_0)$ is a positively invariant set, and any trajectory starting in it must converge to the ω limit set $\{z_{eq0}\}$ due to [37, Lemma 4.1, p. 127]. ■

[11] Recall that a *simply connected* set X is such that every closed curve in X can be continuously contracted into a point without leaving X [125, Section 4.3-6, p. 90].

Proof of Claim 3.5.3

From the definition of ∂K_{+out} (3.8), any $z \in \partial K_{+out}$ has the form $z = (x_0, u_{\max})$ for some x_0 satisfying $cx_0 + du_{\max} > 0$. For any $z \in \partial K_{+out}$, we have from (3.5),

$$f_g(z) = (ax_0 + bu_{\max}, 0), \qquad z_+ - z = (-(x_0 + \tfrac{d}{c}u_{\max}), 0),$$

where $z = (x_0, u_{\max})$ and $cx_0 + du_{\max} > 0$. The condition $f_g(z) = \alpha(z_+ - z)$ is clearly equivalent to $ax_0 + bu_{\max} = -\frac{\alpha}{c}(cx_0 + du_{\max})$. Since $cx_0 + du_{\max} > 0$, it follows that $f_g(z) = \alpha(z_+ - z)$ can hold with $\alpha > 0$ if and only if

$$c(ax_0 + bu_{\max}) < 0. \tag{3.24}$$

If $a = 0$, (3.7) reduces to $bc < 0$ and (3.24) follows. If $a < 0$, we have from (3.7) and $cx_0 + du_{\max} > 0$ that

$$c(ax_0 + bu_{\max}) < acx_0 + adu_{\max} = a(cx_0 + du_{\max}) < 0,$$

and (3.24) holds. This proves the first statement of the claim.

Next, consider the case $a > 0$. Then (3.24) is equivalent to $cx_0 < -\frac{bc}{a}u_{\max}$, and $cx_0 + du_{\max} > 0$ is equivalent to $cx_0 > -du_{\max}$. Hence $f_g(z)$ points towards z_+ on some $z = (x_0, u_{\max}) \in \partial K_{+out}$ if and only if x_0 satisfies

$$-du_{\max} < cx_0 < -\tfrac{bc}{a}u_{\max}. \tag{3.25}$$

It can be verified that $-du_{\max} < -\frac{bc}{a}u_{\max}$ due to (3.7). The above condition (3.25) can be decomposed and rewritten as

$$\begin{aligned} -\tfrac{d}{c}u_{\max} < x_0 < -\tfrac{b}{a}u_{\max}, &\qquad \text{if } c > 0,\\ -\tfrac{b}{a}u_{\max} < x_0 < -\tfrac{d}{c}u_{\max}, &\qquad \text{otherwise,} \end{aligned}$$

so that (3.25) is equivalent to $x_0 = (-\theta\frac{d}{c} - (1-\theta)\frac{b}{a})u_{\max}$ for some $\theta \in (0, 1)$. In other words, $f_g(z)$ points towards z_+ if and only if $z \in \xi_+$. The fact that $f_g(\tilde{z}_+) = 0$ can be verified by substitution, and the last statement of the claim follows. ■

Proof of Claim 3.5.4

We will show that[12] if $a > 0$ and $z_0 \in \partial K_{+out} \setminus \xi_+$, then $z_0 \notin R_n$. If $z_0 \in R_n$, then $\phi_n(t, z_0) \to z_{eq0}$ as $t \to \infty$. Since $z_{eq0} \in K$, it is sufficient to show that if $a > 0$ and $z_0 \in \partial K_{+out} \setminus \xi_+$, then $\phi_n(t, z_0) \notin K$ for all $t \geq 0$. Let $z_0 = (x_0, u_{\max}) \in \partial K_{+out}$ so that $cx_0 + du_{\max} > 0$. At the point z_0, we have $f_n(z_0) = (ax_0 + bu_{\max}, cx_0 + du_{\max})$ (see (3.4)). It follows that $\dot{u}(0) = cx_0 + du_{\max} > 0$ at time $t = 0$, and $u(t)$ must increase (and hence $\text{sat}(u(t)) = u_{\max}$) at least for some non-zero interval. The initial value problem to be considered is

$$\begin{aligned} \dot{x} &= ax + bu_{\max}, &\qquad x(0) &= x_0,\\ \dot{u} &= cx + du, &\qquad u(0) &= u_{\max}, \end{aligned}$$

[12] See Section 3.11.1 on page 102 for clarifications.

whose solution will coincide with the solution of Σ_n, i.e. $\phi_n(t, z_0)$, as long as it remains outside K. We will show that $u(t) \geq u_{\max}$ for all $t \geq 0$, so that $\phi_n(t, z_0) \notin K$ for all $t \geq 0$.

If $c > 0$, we have $-\frac{d}{c}u_{\max} < -\frac{b}{a}u_{\max}$ from (3.7). If $z_0 = (x_0, u_{\max}) \in \partial K_{+out} \setminus \xi_+$, then x_0 satisfies $x_0 \geq -\frac{b}{a}u_{\max}$, and hence $\dot{x}(0) = ax_0 + bu_{\max} \geq 0$. Moreover, because $a > 0$, $x(t)$ is non-decreasing at least until $u(t) < u_{\max}$. Hence $x(t) \geq x_0$ and $cx(t) \geq cx_0$ during this interval.

If $c < 0$, then $-\frac{d}{c}u_{\max} > -\frac{b}{a}u_{\max}$ from (3.7). If $z_0 = (x_0, u_{\max}) \in \partial K_{+out} \setminus \xi_+$, then x_0 satisfies $x_0 \leq -\frac{b}{a}u_{\max}$, and hence $\dot{x}(0) = ax_0 + bu_{\max} \leq 0$. Moreover, because $a > 0$, $x(t)$ is non-increasing at least until $u(t) < u_{\max}$. Hence $x(t) \leq x_0$ and $cx(t) \geq cx_0$ during this interval.

In either case, we have

$$\dot{u} = cx + du \geq cx_0 + du, \qquad u(0) = u_{\max},$$

as the differential inequality governing $u(t)$. To apply the Comparison Lemma [37, Lemma 3.4, pp. 102 – 103], define $v := -u$, so that

$$\dot{v} = -\dot{u} \leq -cx_0 - du = dv - cx_0, \qquad v(0) = -u_{\max}.$$

Applying the Comparison Lemma [37, Lemma 3.4, pp. 102 – 103] to the above differential inequality yields $v(t) \leq -u_{\max}e^{dt} - \frac{c}{d}x_0(e^{dt} - 1)$, and hence

$$u(t) = -v(t) \geq u_{\max}e^{dt} + \tfrac{c}{d}x_0(e^{dt} - 1), \qquad \forall t \geq 0.$$

Since $a > 0$, it follows from (3.6) that $d < -a < 0$ and hence $(e^{dt} - 1) \leq 0$ for all $t \geq 0$. Because $cx_0 + du_{\max} > 0$, we have $\frac{c}{d}x_0 < -u_{\max}$ and $\frac{c}{d}x_0(e^{dt} - 1) \geq -u_{\max}(e^{dt} - 1)$. With these, the above inequality becomes

$$u(t) \geq u_{\max}e^{dt} + \tfrac{c}{d}x_0(e^{dt} - 1) \geq u_{\max}e^{dt} - u_{\max}(e^{dt} - 1) = u_{\max}, \qquad \forall t \geq 0,$$

as desired. ■

Proof of Claim 3.5.5

If $a \leq 0$, the result is a direct consequence of Claim 3.5.3 and the fact that $\partial K_{+out} \cup \{z_+\}$ contains no equilibrium points of Σ_g. If $a > 0$, then the result follows from Claim 3.5.4 and Claim 3.5.3, and the fact that $\xi_+ \cup \{z_+\}$ contains no equilibrium points of Σ_g. ■

Proof of Claim 3.5.6

We will show that[13] if $\phi_n(T, z_0) \notin l(z_0, z_+)$, then $z_0 \notin R_n$. Let $z_1 := \phi_n(T, z_0)$ and assume $z_1 \in \partial K_{+out} \setminus l(z_0, z_+)$. If $z_1 = z_0$, then the solution forms a closed orbit, and due to uniqueness of solutions (see Claim 3.3.1), $\phi_n(t, z_0)$ will stay on the orbit for all $t \geq 0$ and never approach z_{eq0}. Hence $z_0 \notin R_n$. Otherwise, we have $z_1 \in \partial K_{+out} \setminus (l(z_0, z_+) \cup \{z_0\})$. Let the closed bounded region enclosed by the closed path

$$\tilde{\eta}(z_0) := \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_n(t, z_0), \forall t \in [0, T]\} \cup l(z_0, z_1),$$

[13] See Section 3.11.1 on page 102 for clarifications.

be $\tilde{D}(z_0)$. Note that $\phi_n(t, z_0)$ must necessarily intersect ∂K_{+in} and enter K before it can intersect ∂K_{+out} at time T due to Fact 3.1.3 (and Fact 3.1.2). It can be seen that $l(z_0, z_1) \subset \partial K_{+out}$ is a transverse section to f_n, with f_n pointing *out* of $\tilde{D}(z_0)$ on $l(z_0, z_1)$. Hence $\tilde{D}(z_0)$ is a *negatively invariant set* [135, p. 47] of system Σ_n. If $z_{eq0} \in \tilde{D}(z_0)$, then there is no way for $\phi_n(t, z_0)$ to reach z_{eq0}, which will prove the claim. We will show that z_{eq0} must be contained in $\tilde{D}(z_0)$ using *index theory* [39, Section 2.4, pp. 49 – 51], [136, §V.8, pp. 300 – 305]. Noting that the *index* [39, Definition 2.16, p. 49] of a closed orbit is +1 [136, p. 301], it can be shown that the index of the closed path $\tilde{\eta}(z_0)$, formed by a section of a trajectory and a transverse section, is also +1 [136, pp. 301 – 302]. The indices of a node, focus and saddle are +1, +1, and −1 respectively [136, p. 301]. Since the index of $\tilde{\eta}(z_0)$ is the sum of all indices of equilibria enclosed by $\tilde{\eta}(z_0)$ [136, p. 301], and system Σ_n has only one node or focus at the origin with possibly two additional saddle points (see Claims 3.4.1 and 3.4.2), the only way for $\tilde{\eta}(z_0)$ to have an index of +1 is for it to enclose the origin z_{eq0} alone. That is, $z_{eq0} \in \tilde{D}(z_0)$. ■

Remark 3.24. The above proof is most evident by visualizing the vector field f_n on the path $\tilde{\eta}(z_0)$. □

3.11.4 Proofs of Intermediate Results for Section 3.5.2

Proof of Claim 3.5.8 (Transverse Sections in Saturated Region)

We need to show that the dot product of the unit normals to ∂K_{+div} (respectively, σ_{+div}) and f_n does not vanish and never changes sign on these line segments. For any $z \in K_+$, we have $f_n(z) = f_n(x, u) = (ax + bu_{\max}, cx + du)$ (see (3.4)), which is valid on σ_{+div} and ∂K_{+div} because $\sigma_{+div} \subset \partial K_{+div} \subset K_+$. Let $\tilde{T}z_+ := (u_{\max}, \frac{d}{c}u_{\max})$. For case (i) (respectively, (ii)), it can be verified that $\tilde{T}\frac{z_+}{\|z_+\|}$ is a unit normal of ∂K_{+div} (respectively, σ_{+div}). We need to show that $\langle \tilde{T}z_+, f_n(z)\rangle \neq 0$ for all $z \in \partial K_{+div}$ (respectively, for all $z \in \sigma_{+div}$). Any $z \in \partial K_{+div}$ can be expressed as $z = (x, u) = (-\frac{d}{c}u, u)$ for some $u > u_{\max}$ (see (3.8)). On any point $z \in \partial K_{+div}$, direct computation yields

$$\begin{aligned}
\langle \tilde{T}z_+, f_n(z)\rangle &= \langle (u_{\max}, \tfrac{d}{c}u_{\max}), (ax + bu_{\max}, cx + du)\rangle, \\
&= \langle (u_{\max}, \tfrac{d}{c}u_{\max}), (-\tfrac{ad}{c}u + bu_{\max}, 0)\rangle, \\
&= (-\tfrac{ad}{c}u + bu_{\max})u_{\max} = -\tfrac{1}{c}(adu - bcu_{\max})u_{\max},
\end{aligned}$$

for some $u > u_{\max}$, which shows $\langle \tilde{T}z_+, f_n(z)\rangle = 0$ if and only if $adu - bcu_{\max} = 0$.

For case (i), we have $ad \geq 0$, so that $adu \geq adu_{\max} > bcu_{\max}$, where the last inequality is due to (3.7). Then $adu - bcu_{\max} > 0$, and we have $\langle \tilde{T}z_+, f_n(z)\rangle \neq 0$ for all $z \in \partial K_{+div}$, as desired.

For case (ii), it can be verified that $ad < 0$ due in part to (3.6) ($d < -a < 0$ when $a > 0$). Then $\frac{bc}{ad} > 1$ due to (3.7) and $\sigma_{+div} \neq \emptyset$. On ∂K_{+div}, $\langle \tilde{T}z_+, f_n(z)\rangle = 0$ can hold if and only if $adu - bcu_{\max} = 0$. This is assured on any point $z = (x, u) \in \sigma_{+div} \subset \partial K_{+div}$ due to $u < \frac{bc}{ad}u_{\max}$. ■

Remark 3.25. For case (ii), the proof also shows that $\tilde{\sigma}_{+div} \setminus \{z_{eq+}\}$ is a transverse section to f_n. □

Proof of Claim 3.5.9

We will show that[14] if $z_0 \in \tilde{\sigma}_{+div}$, then $z_0 \notin R_n$. If $z_0 \in R_n$, we have $\phi_n(t, z_0) \to z_{eq0}$ as $t \to \infty$. Since $z_{eq0} \in K$, it is sufficient to show that if $z_0 \in \tilde{\sigma}_{+div}$, then $\phi_n(t, z_0) \notin K$ for all $t \geq 0$. It can be verified that $ad < 0$, due in part to (3.6) ($d < -a < 0$ when $a > 0$). Then (3.7) yields $\frac{bc}{ad} > 1$. Let $z_0 = (x_0, u_0) \in \tilde{\sigma}_{+div}$, so that $u_0 \geq \frac{bc}{ad} u_{\max} > u_{\max}$ and $cx_0 + du_0 = 0$ (see (3.14) and (3.8)). Since $u_0 > u_{\max}$, we have $\text{sat}(u) = u_{\max}$ for all u in a sufficiently small neighborhood of u_0. Consider the initial value problem (see (3.4))

$$\begin{aligned} \dot{x} &= ax + bu_{\max}, \qquad & x(0) &= x_0, \\ \dot{u} &= cx + du, \qquad & u(0) &= u_0, \end{aligned}$$

whose solution will coincide with $\phi_n(t, z_0)$ as long as it remains in $K_+ \cup \partial K_+$. Solving for $x(t)$ yields

$$x(t) = x_0 e^{at} + \tfrac{b}{a} u_{\max}(e^{at} - 1), \qquad \forall t \geq 0. \tag{3.26}$$

We will show that $u(t) \geq u_{\max}$ for all $t \geq 0$, so that $\phi_n(t, z_0) \notin K$ for all $t \geq 0$.

Consider case (i) ($a < 0$ and $d \in (0, -a)$). Define $v := u - u_0$ so that $\dot{v} = \dot{u} = cx + du = dv + (cx + du_0)$, and consider

$$\dot{v} = dv + (cx + du_0), \qquad v(0) = u(0) - u_0 = 0.$$

Clearly, if $v(t) = u(t) - u_0 \geq 0$ for all $t \geq 0$, then $u(t) \geq u_0 \geq \frac{bc}{ad} u_{\max} > u_{\max}$ for all $t \geq 0$, and the conclusion follows. Since $d > 0$, a sufficient condition for $v(t) \geq 0$ for all $t \geq 0$, is for the input of the preceding ODE to satisfy $cx(t) + du_0 \geq 0$ for all $t \geq 0$. Using $cx_0 + du_0 = 0$ and the solution of $x(t)$ (3.26), this condition follows from

$$\begin{aligned} cx(t) + du_0 = cx_0 e^{at} + \tfrac{bc}{a} u_{\max}(e^{at} - 1) + du_0 &= -du_0 e^{at} + \tfrac{bc}{a} u_{\max}(e^{at} - 1) + du_0, \\ = (du_0 - \tfrac{bc}{a} u_{\max})(1 - e^{at}) &= d(u_0 - \tfrac{bc}{ad} u_{\max})(1 - e^{at}) \geq 0, \end{aligned}$$

for all $t \geq 0$, where the final inequality is due to $a < 0$, $d > 0$ and $u_0 \geq \frac{bc}{ad} u_{\max}$.

Now consider case (ii) ($a > 0$). From (3.6), we have $d < -a < 0$. In turn, we have $\frac{bc}{a} u_{\max} \geq du_0$ due to $u_0 \geq \frac{bc}{ad} u_{\max}$. Because $a > 0$, we have $e^{at} - 1 \geq 0$ for all $t \geq 0$. The evolution of $cx(t)$ then satisfy (see (3.26))

$$cx(t) = cx_0 e^{at} + \tfrac{bc}{a} u_{\max}(e^{at} - 1) \geq cx_0 e^{at} + du_0(e^{at} - 1) = -du_0 = cx_0,$$

for all $t \geq 0$, due to $cx_0 + du_0 = 0$. Then $u(t)$ is governed by the differential inequality

$$\dot{u} = cx + du \geq cx_0 + du, \qquad u(0) = u_0.$$

In similar manner as the proof of Claim 3.5.4, define $\tilde{v} := -u$, so that

$$\dot{\tilde{v}} \leq d\tilde{v} - cx_0, \qquad \tilde{v}(0) = -u_0.$$

Applying the Comparison Lemma [37, Lemma 3.4, pp. 102 – 103] to the above differential inequality yields $\tilde{v}(t) \leq -u_0 e^{dt} - \frac{c}{d} x_0(e^{dt} - 1)$. Using $cx_0 + du_0 = 0$, we have

$$u(t) = -\tilde{v}(t) \geq u_0 e^{dt} + \tfrac{c}{d} x_0(e^{dt} - 1) = \tfrac{1}{d}(du_0 e^{dt} + cx_0(e^{dt} - 1)) = -\tfrac{c}{d} x_0 = u_0 > u_{\max},$$

[14] See Section 3.11.1 on page 102 for clarifications.

for all $t \geq 0$, as desired. ■

Proof of Claim 3.5.10

Let $z_0 = (x_0, u_0) \in K_{+out} \cap R_n$, so that $u_0 > u_{\max}$, $cx_0 + du_0 > 0$ (see (3.8)), and $\phi_n(t, z_0) \to z_{eq0}$ as $t \to \infty$. Since $z_{eq0} \in K$ and $z_0 \in K_{+out} \subset K_+$, Fact 3.1.2 shows that $\phi_n(t, z_0)$ must intersect ∂K_{+in} at some finite time. Let T_n be the first time instant that $\phi_n(t, z_0)$ intersects ∂K_{+in}. It is clear that T_n satisfies (3.15), and further, that $l(z_+, \phi_n(T_n, z_0)) \subset \partial K_{+in}$. This proves the first statement of the claim.

The solution of the nominal system, $\phi_n(t, z_0) = (x_n(t), u_n(t))$, is governed by (see (3.4))

$$\begin{aligned} \dot{x}_n &= ax_n + bu_{\max}, & x_n(0) &= x_0, \\ \dot{u}_n &= cx_n + du_n, & u_n(0) &= u_0, \end{aligned}$$

as long as $u_n(t) \geq u_{\max}$, i.e. for all $t \leq T_n$. The solution of the GPAW-compensated system, $\phi_g(t, z_0) = (x_g(t), u_g(t))$, is governed by (see (3.5))

$$\begin{aligned} \dot{x}_g &= ax_g + bu_{\max}, & x_g(0) &= x_0, \\ \dot{u}_g &= \begin{cases} 0, & \text{if } cx_g + du_g > 0, \\ cx_g + du_g, & \text{otherwise,} \end{cases} & u_g(0) &= u_0, \end{aligned}$$

as long as $u_g(t) \geq u_{\max}$. We need to show that there exists a $T_g \in (0, \infty)$ such that $T_g < T_n$ and $\phi_g(T_g, z_0) \in l(z_+, \phi_n(T_n, z_0))$.

Solving the initial value problem

$$\dot{x} = ax + bu_{\max}, \qquad x(0) = x_0,$$

yields

$$x(t) = \begin{cases} x_0 + bu_{\max}t, & \text{if } a = 0, \\ x_0 e^{at} + \frac{b}{a}u_{\max}(e^{at} - 1), & \text{otherwise,} \end{cases} \qquad \forall t \geq 0.$$

It can be seen that $x_n(t) = x(t)$ for all t such that $u_n(t) \geq u_{\max}$, and $x_g(t) = x(t)$ for all t such that $u_g(t) \geq u_{\max}$. Define

$$T_g := \inf\{\bar{t} \in (0, \infty) \mid u_g(\bar{t}) \leq u_{\max}\}, \qquad T := \min\{T_n, T_g\}. \tag{3.27}$$

Observe that T_g is the first time instant that $\phi_g(t, z_0)$ intersects ∂K_{+in}, or ∞ if $\phi_g(t, z_0)$ never intersects ∂K_{+in}. With T as defined, the preceding relations yield $x_n(t) = x_g(t) = x(t)$ for all $t \in [0, T]$. Hence $x_n(t)$ and $x_g(t)$ are well defined at least for all $t \in [0, T]$. Now, define

$$h_n(t, u_n(t)) := cx(t) + du_n(t), \qquad h_g(t, u_g(t)) := \begin{cases} 0, & \text{if } cx(t) + du_g(t) > 0, \\ cx(t) + du_g(t), & \text{otherwise.} \end{cases}$$

Observe that whenever $cx(t) + du_g(t) > 0$, then $h_n(t, u_g(t)) > h_g(t, u_g(t))$ holds. When $cx(t) + du_g(t) \leq 0$, we have $h_n(t, u_g(t)) = h_g(t, u_g(t))$. Hence $h_g(t, u_g(t)) \leq h_n(t, u_g(t))$ for all $u_g(t) \geq u_{\max}$, for all $t \in [0, T]$. Clearly, the solution of $u_g(t)$ is governed by the differential inequality

$$\dot{u}_g(t) = h_g(t, u_g(t)) \le h_n(t, u_g(t)), \qquad u_g(0) = u_0,$$

while the solution of $u_n(t)$ is governed by the ODE

$$\dot{u}_n(t) = h_n(t, u_n(t)), \qquad u_n(0) = u_0,$$

for all $t \in [0, T]$. By the Comparison Lemma [37, Lemma 3.4, pp. 102 – 103], we have $u_g(t) \le u_n(t)$ for all $t \in [0, T]$.

To obtain a *strict* inequality, observe that $cx(0)+du_n(0) = cx(0)+du_g(0) = cx_0+du_0 > 0$ holds with *strict* inequality (which implies $\dot{u}_n(0) > 0$ and $u_n(t)$ is strictly increasing at $t = 0$). Then there exists a sufficiently small $\delta_{\max} > 0$ such that for all $\delta \in (0, \delta_{\max})$, we have $cx(t) + du_n(t) > 0$ and $cx(t) + du_g(t) > 0$ for all $t \in [0, \delta]$. Since $u_n(t)$ is increasing at $t = 0$, for any such $\delta \in (0, \delta_{\max})$, there exists an $\epsilon = \epsilon(\delta) > 0$ such that $u_n(\delta) = u_0 + \epsilon$. Moreover, we have $u_g(\delta) = u_0$ due to $\dot{u}_g(t) = 0$ for all $t \in [0, \delta]$ (see the definition of h_g). In other words, defining $u_\delta := u_0 + \epsilon$, we have

$$\begin{aligned} \dot{u}_n(t) &= h_n(t, u_n(t)), \qquad u_n(\delta) = u_\delta, \\ \dot{u}_g(t) &\le h_n(t, u_g(t)), \qquad u_g(\delta) < u_\delta. \end{aligned}$$

Applying Lemma 3.11.1 in Section 3.11.2 (on page 102) to the preceding, we get the *strict* condition $u_g(t) < u_n(t)$ for all $t \in [\delta, T]$. Since $\delta > 0$ is only required to be smaller than $\delta_{\max}$ but otherwise arbitrary, we have $u_g(t) < u_n(t)$ for all $t \in (0, T]$.

Assume for the sake of contradiction that $\phi_g(t, z_0)$ never intersects ∂K_{+in}. Then $T_g = \infty$ and $T := \min\{T_n, T_g\} = T_n < \infty$ (see (3.27)). Since $\phi_n(T_n, z_0) \in \partial K_{+in}$, we have $u_n(T_n) = u_{\max}$. The condition $u_g(t) < u_n(t)$ for all $t \in (0, T_n]$ yields $u_g(T_n) < u_n(T_n) = u_{\max}$. This, coupled with $u_g(0) = u_0 > u_{\max}$ and continuity of $u_g(t)$ means that there exists a $\tilde{T} \in (0, T_n)$ such that $u_g(\tilde{T}) = u_{\max}$. This contradicts the assumption that $\phi_g(t, z_0)$ never intersects ∂K_{+in}, and also shows that $T_g = \tilde{T} < T_n < \infty$ and $\phi_g(T_g, z_0) \in \partial K_{+in}$. It remains to show that $\phi_g(T_g, z_0) \in l(z_+, \phi_n(T_n, z_0))$.

Since $z_0 \in K_{+out}$ and $\phi_n(T_n, z_0), \phi_g(T_g, z_0) \in \partial K_{+in}$, both $\phi_n(t, z_0)$ and $\phi_g(t, z_0)$ must intersect ∂K_{+div} at least once (see Fact 3.1.2 and the definition of $K_{+out}, K_{+in}, K_{+div}$ in (3.8)). Let $T_{ndiv} \in (0, T_n)$ and $T_{gdiv} \in (0, T_g)$ be the first time instants that $\phi_n(t, z_0)$ and $\phi_g(t, z_0)$ intersect ∂K_{+div} respectively, so that

$$\phi_n(t, z_0) \in K_{+out}, \forall t \in [0, T_{ndiv}), \qquad \phi_g(t, z_0) \in K_{+out}, \forall t \in [0, T_{gdiv}).$$

This implies $cx(t) + du_g(t) > 0$ for all $t \in [0, T_{gdiv})$, so that $\dot{u}_g(t) = 0$ and $u_g(t) = u_0$ for all $t \in [0, T_{gdiv}]$. It also implies that $\dot{u}_n(t) = cx(t) + du_n(t) > 0$ for all $t \in [0, T_{ndiv})$, so that $u_n(T_{ndiv}) > u_0$. Then we have $u_{\max} < u_0 = u_g(T_{gdiv}) < u_n(T_{ndiv})$, which implies $\phi_g(T_{gdiv}, z_0) \in l(z_+, \phi_n(T_{ndiv}, z_0)) \subset \partial K_{+div}$. Let $z_n := \phi_n(T_{ndiv}, z_0)$, $z_g := \phi_g(T_{gdiv}, z_0)$, and let $\tilde{D}(z_n)$ be the closed bounded region enclosed by the closed path

$$\tilde{\eta}(z_n) := l(z_+, z_n) \cup \tilde{\eta}_{\phi_n}(z_n) \cup l(z_+, \phi_n(T_n, z_0)) \cup \{z_+\},$$

where

$$\tilde{\eta}_{\phi_n}(z_n) := \{\bar{z} \in \mathbb{R}^2 \mid \bar{z} = \phi_n(t, z_0), \forall t \in [T_{ndiv}, T_n]\}.$$

It can be seen that starting from z_+, $\tilde{\eta}(z_n)$ traces the path along ∂K_{+div} towards the point where $\phi_n(t, z_0)$ first intersects it, proceeds along the solution $\phi_n(t, z_0)$ until it intersects ∂K_{+in}, then along ∂K_{+in} until it reaches its starting point z_+.

If the open-loop system is marginally stable ($a = 0$) or strictly stable with a stable controller ($a < 0$ and $d \leq 0$), Claim 3.5.8 shows that ∂K_{+div} is a transverse section to f_n. Since $\phi_n(t, z_0)$ traverses from K_{+out} through $z_n \in \partial K_{+div}$ to K_{+in}, all trajectories of Σ_n intersecting the transverse section ∂K_{+div} can only pass from K_{+out} to K_{+in}, i.e. they cannot pass from K_{+in} to K_{+out} through ∂K_{+div}. This implies that $\phi_n(t, z_0)$ can never return to K_{+out} within the interval $[T_{ndiv}, T_n]$, and $\tilde{D}(z_n)$ is contained in $K_{+in} \cup \partial K_{+div} \cup \partial K_{+in} \cup \{z_+\}$. Moreover, $l(z_+, z_n) \subset \partial K_{+div}$ is also a transverse section to f_n.

If the open-loop system is strictly stable with an unstable controller ($a < 0$ and $d \in (0, -a)$), or unstable ($a > 0$), the assumption $z_0 \in R_n$ and Claim 3.5.9 implies $\phi_n(t, z_0) \notin \tilde{\sigma}_{+div}$ for all $t \in [0, T_n]$, which in turn implies $z_n \in \sigma_{+div}$. Claim 3.5.8 shows that σ_{+div} is a transverse section to f_n, which by the same reasoning, implies that $\phi_n(t, z_0)$ can never return to K_{+out} within the interval $[T_{ndiv}, T_n]$, and $\tilde{D}(z_n)$ is contained in $K_{+in} \cup \sigma_{+div} \cup \partial K_{+in} \cup \{z_+\}$ ($\subset K_{+in} \cup \partial K_{+div} \cup \partial K_{+in} \cup \{z_+\}$). Moreover, $l(z_+, z_n) \subset \sigma_{+div}$ is also a transverse section to f_n.

By Claim 3.3.1, the solutions of system Σ_n are unique, so that no two different paths can intersect [135, pp. 38]. Hence, no solution starting in $\tilde{D}(z_n) \setminus \tilde{\eta}_{\phi_n}(z_n)$ can intersect $\tilde{\eta}_{\phi_n}(z_n)$, or exit $\tilde{D}(z_n)$ through the segment $\tilde{\eta}_{\phi_n}(z_n)$. This, together with the fact that $l(z_+, z_n)$ is a transverse section and $z_g \in l(z_+, z_n) \subset \tilde{D}(z_n) \setminus \tilde{\eta}_{\phi_n}(z_n)$, means that $\phi_n(t, z_g)$ can exit the region $\tilde{D}(z_n)$ only through the line segment $l(z_+, \phi_n(T_n, z_0)) \subset \partial K_{+in}$. By Fact 3.1.1, f_n and f_g coincide in $\tilde{D}(z_n) \subseteq K_{+in} \cup \partial K_{+div} \cup \partial K_{+in} \cup \{z_+\}$, so that $\phi_g(t, z_g) = \phi_n(t, z_g)$ at least until $\phi_n(t, z_g)$ exits $\tilde{D}(z_n)$, i.e. until $t = T_g - T_{gdiv}$, where $\phi_g(T_g - T_{gdiv}, z_g) = \phi_g(T_g - T_{gdiv}, \phi_g(T_{gdiv}, z_0)) = \phi_g(T_g, z_0) \in \partial K_{+in}$. It follows that $\phi_g(t, z_g)$ can exit $\tilde{D}(z_n)$ only through the line segment $l(z_+, \phi_n(T_n, z_0))$, i.e. $\phi_g(T_g, z_0) \in l(z_+, \phi_n(T_n, z_0))$, as desired. ■

Proof of Claim 3.5.11 (Invariance of $\bar{E}(z_0) \subset R_g$)

Let

$$\tilde{\sigma}_+ := \begin{cases} l(\phi_g(t_{int}, z_0), z_+) \cup \{z_+\}, & \text{if } t_{int} < \infty, \\ \sigma_+, & \text{otherwise.} \end{cases}$$

Observing from Fact 3.1.1 that f_n and f_g coincide on $\tilde{\sigma}_+ \subset K \cup \{z_+\}$, it can be verified as in the proof of Claim 3.5.1, that $\tilde{\sigma}_+$ is a transverse section to f_g, and f_g always points into $\bar{E}(z_0)$ on $\tilde{\sigma}_+$. It is clear that $l(z_0, z_+) \subset \partial K_{+in}$ is also a transverse section to f_g, and that f_g always points into $\bar{E}(z_0)$ on $l(z_0, z_+)$. Both of these results show that any solution of the GPAW-compensated system $\phi_g(t, z_0)$ originating in $\bar{E}(z_0)$ cannot exit $\bar{E}(z_0)$ through the line segments $\tilde{\sigma}_+$ or $l(z_0, z_+)$. Furthermore, the solution $\phi_g(t, z_0)$ is unique due to Proposition 3.3.3, which implies that no solution originating in $\bar{E}(z_0)$ can exit it through the boundary $\gamma_{0\phi_g}(z_0)$ (or $\gamma_{int\phi_g}(z_0)$ as appropriate) (see Remark 3.3). These show that the region $\bar{E}(z_0)$ enclosed by $\gamma(z_0)$ must be a positively invariant set for system Σ_g. Theorem 2.5.3 shows that $\phi_g(t, z_0) \in \bar{K}$ for all $t \geq 0$, which implies $\bar{E}(z_0) \subset \bar{K}$. This proves the first statement of the claim.

Since $z_0 \in R_g$, we have $\phi_g(t, z_0) \to z_{eq0}$ as $t \to \infty$. Since $\bar{E}(z_0)$ is a positively invariant set and $z_0 \in \bar{E}(z_0)$, we have $\phi_g(t, z_0) \in \bar{E}(z_0)$ for all $t \geq 0$. The conclusion $z_{eq0} \in \bar{E}(z_0)$ then follows from the fact that $\bar{E}(z_0)$ is *closed* and hence contains all its limit points.

It remains to show that $\bar{E}(z_0) \subset R_g$. Observe that

$$(l(\phi_g(t_{int}, z_0), z_+) \cup \{z_+\} \cup l(z_0, z_+)) \cap \partial K_{+out} = \emptyset, \qquad (\{z_{eq0}\} \cup \sigma_+ \cup l(z_0, z_+)) \cap \partial K_{+out} = \emptyset.$$

It follows from the definitions of $\gamma(z_0)$ (3.16) and $\bar{E}(z_0)$, that if $\bar{\gamma} := \bar{E}(z_0) \cap \partial K_{+out} \neq \emptyset$, then $\bar{\gamma}$ must lie in the line segments $\gamma_{0\phi_g}(z_0)$ (or $\gamma_{int\phi_g}(z_0)$), i.e. $\bar{\gamma} \subset \gamma_{0\phi_g}(z_0)$ (or $\bar{\gamma} \subset \gamma_{int\phi_g}(z_0)$). Hence any solution of Σ_g starting in $\bar{E}(z_0)$ that intersects ∂K_{+out} must intersect $\phi_g(t, z_0)$ at some time. Since $\lim_{t\to\infty} \phi_g(t, z_0) = z_{eq0}$, it follows from uniqueness of solutions (see Remark 3.3) that any solution starting from a point $\tilde{z} \in \bar{E}(z_0)$ that intersects ∂K_{+out} must intersect $\phi_g(t, z_0)$ and converge to z_{eq0}, i.e. $\tilde{z} \in R_g$. In similar manner, any solution starting from a point $\hat{z} \in \bar{E}(z_0)$ that intersects ∂K_{-out} must converge to z_{eq0}, i.e. $\hat{z} \in R_g$. These imply $(\bar{E}(z_0) \cap (\partial K_{+out} \cup \partial K_{-out})) \subset R_g$. It suffices to consider solutions that do not intersect $\partial K_{+out} \cup \partial K_{-out}$, i.e. solutions contained in $\tilde{E}(z_0) := \bar{E}(z_0) \setminus (\partial K_{+out} \cup \partial K_{-out})$.

It can be verified from Claim 3.4.3 that any equilibria of Σ_g apart from z_{eq0} contained in $\bar{E}(z_0)$ must lie in $\partial K_{+out} \cup \partial K_{-out}$. Then the only equilibrium point in $\tilde{E}(z_0)$ $(\subset \bar{E}(z_0))$ is z_{eq0}, which must be a stable node or focus. Observe that f_g is *continuously differentiable* in $\tilde{E}(z_0) \subset K \cup \partial K_{+in} \cup \partial K_{-in} \cup \{z_+, z_-\}$, so that Bendixson's Criterion [37, Lemma 2.2, pp. 67] applies in this region. As in the proof of Claim 3.5.2 (on page 104), Bendixson's Criterion [37, Lemma 2.2, pp. 67] and the absence of saddle points in $\tilde{E}(z_0)$ means that $\{z_{eq0}\}$ is the ω limit set of every solution contained in $\tilde{E}(z_0)$. Hence $\tilde{E}(z_0) \subset R_g$, and the conclusion follows. ■

Chapter 4

Geometric Properties and Region of Attraction Comparison Results

In this chapter, we present a geometric property of the GPAW-compensated controller and derive some region of attraction (ROA) comparison results similar in principle to Proposition 3.5.13. As shown in Section 2.5, the GPAW-compensated controller is defined by the online solution to a combinatorial optimization subproblem. In Section 4.1, we show that it can be equivalently defined by the online solution to a convex quadratic program, or a projection onto a convex polyhedral cone problem.[1] This is significant because it holds regardless of any nonlinearities in the plant or controller. Geometric properties of the projection operator are shown in Section 4.2, leading to a geometric bounding condition relating the vector fields of the nominal controller and GPAW-compensated controller. Section 4.3 derives a general ROA comparison result, that gives sufficient conditions to ensure the ROA of a system contains the ROA estimate of a related system. This is specialized in Section 4.4 to yield ROA comparison results between the nominal system and GPAW-compensated system. The chapter concludes by demonstration of these ROA comparison results on some simple systems in Section 4.5.

4.1 Quadratic Program Formulation of GPAW-Compensated Controllers

As shown in Section 2.5, the GPAW-compensated controller (2.27)

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g), \end{aligned} \tag{4.1}$$

is defined by the projection matrix (2.30)

$$R_{\mathcal{I}}(x_g) = \begin{cases} I - \Gamma N_{\mathcal{I}} (N_{\mathcal{I}}^{\mathrm{T}} \Gamma N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise}, \end{cases} \tag{4.2}$$

[1] When the controller has output of dimensions one or two, closed-form expressions for the GPAW-compensated controller are available in Appendices A and B.

and an online solution $\mathcal{I}^*$ to the combinatorial optimization subproblem (2.31)

$$\begin{aligned} \max_{\mathcal{I}\in\mathcal{J}} F(\mathcal{I}) &= f_c^{\mathrm{T}}(x_g,y,r)\Gamma^{-1}R_{\mathcal{I}}(x_g)f_c(x_g,y,r), \\ \text{subject to} \qquad & \operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|, \\ & N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}(x_g)R_{\mathcal{I}}(x_g)f_c(x_g,y,r) \leq 0. \end{aligned} \tag{4.3}$$

In (4.2) and (4.3), the matrix $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{q\times q}$ is the chosen GPAW parameter (see Remark 2.15), while matrix $N_{\mathcal{I}}(x_g)$ (2.29) and sets $\mathcal{I}_{\mathrm{sat}}$, $\mathcal{J}$ are defined by[2]

$$\begin{aligned} N_{\mathcal{I}}(x_g) &= \begin{cases} [\nabla h_{\sigma_{\mathcal{I}}(1)}(x_g), \nabla h_{\sigma_{\mathcal{I}}(2)}(x_g), \ldots, \nabla h_{\sigma_{\mathcal{I}}(|\mathcal{I}|)}(x_g)], & \text{if } \mathcal{I}\neq\emptyset, \\ 0, & \text{otherwise,} \end{cases} \\ \mathcal{I}_{\mathrm{sat}} &:= \mathcal{I}_{\mathrm{sat}}(x_g) = \{i \in \mathcal{I}_{2m} \mid h_i(x_g) \geq 0\}, \qquad \mathcal{J} := \{\mathcal{I} \subset \mathcal{I}_{\mathrm{sat}} \mid |\mathcal{I}| \leq q\}, \end{aligned} \tag{4.4}$$

where $h_i(x_g)$ for all $i \in \mathcal{I}_{2m}$ are the saturation constraint functions defined in (2.28) and $\sigma_{\mathcal{I}}$ is the bijection described in Remark 2.5. Proposition 2.5.1 shows that solutions to subproblem (4.3) always exist, and Remark 2.16 made the observation that they can always be found by an exhaustive search algorithm. This formulation has the advantage that when the dimension of the controller output m is small, closed form expressions can be derived from the combinatorial optimization subproblem (4.3) as shown in Appendices A and B. These allow a highly efficient realization as shown in Section B.1 of Appendix B.

In this section, we show that the same GPAW-compensated controller (4.1) can be defined instead by the online solution to a convex quadratic program, or a projection onto a convex polyhedral cone problem. This gives three different but equivalent ways to realize the GPAW-compensated controller, the suitability of which depends on the computational efficiency of available algorithms. See Appendix C for a summary of available methods to realize the GPAW-compensated controller.

First, recall the definitions of the *principal* and *complementary projection matrices* $P_{\mathcal{I}}(x_g)$, $S_{\mathcal{I}}(x_g)$ in (2.42), and their relations with $R_{\mathcal{I}}(x_g)$ in (2.43),

$$R_{\mathcal{I}}(x_g) = \Phi P_{\mathcal{I}}(x_g)\Phi^{-1} = I - \Phi S_{\mathcal{I}}(x_g)\Phi^{-1}, \tag{4.5}$$

where $\Phi \in \mathbb{R}^{q\times q}$ is a nonsingular matrix obtained from a decomposition of the GPAW parameter $\Gamma = \Phi\Phi^{\mathrm{T}}$ [124, Theorem 7.2.7, p. 406]. Recall also that both $P_{\mathcal{I}}(x_g)$ and $S_{\mathcal{I}}(x_g)$ are *idempotent* [126, p. 697] and satisfy (2.45). Using (4.5), (2.45), and (2.42), the objective function of subproblem (4.3) can be written as

$$\begin{aligned} F(\mathcal{I}) = f_c^{\mathrm{T}}\Gamma^{-1}R_{\mathcal{I}}f_c = f_c^{\mathrm{T}}\Phi^{-\mathrm{T}}\Phi^{-1}(I - \Phi S_{\mathcal{I}}\Phi^{-1})f_c &= \tilde{f}_c^{\mathrm{T}}\tilde{f}_c - \tilde{f}_c^{\mathrm{T}}S_{\mathcal{I}}\tilde{f}_c, \\ = \|\tilde{f}_c\|^2 - \tilde{f}_c^{\mathrm{T}}S_{\mathcal{I}}^{\mathrm{T}}S_{\mathcal{I}}\tilde{f}_c = \|\tilde{f}_c\|^2 - \|S_{\mathcal{I}}\tilde{f}_c\|^2 &= \|\tilde{f}_c\|^2 - \|(I-P_{\mathcal{I}})\tilde{f}_c\|^2, \end{aligned}$$

where $\tilde{f}_c := \Phi^{-1}f_c := \Phi^{-1}f_c(x_g,y,r)$ and all function arguments have been dropped. Now, observe that maximizing $F(\mathcal{I})$ is equivalent to minimizing $-F(\mathcal{I})$, and that adding a constant term to the objective function will not change the optimal solution(s). Since the first term $\|\tilde{f}_c\|^2$ does not vary with $\mathcal{I}$ (hence is a constant term), we can replace the maximization of $F(\mathcal{I})$ by the minimization of $\tilde{F}(\mathcal{I}) := -F(\mathcal{I}) + \|\tilde{f}_c\|^2 = \|(I-P_{\mathcal{I}})\tilde{f}_c\|^2$ without changing any optimal solution $\mathcal{I}^*$. Hence the GPAW-compensated controller (4.1) can be defined by

[2] Recall that $\mathcal{I}_i := \{1,2,\ldots,i\}$ for any positive integer i, and m is the dimension of the controller output.

an online solution $\mathcal{I}^*$ to the modified combinatorial *minimization* subproblem

$$\begin{aligned} \min_{\mathcal{I}\in\mathcal{J}} \tilde{F}(\mathcal{I}) &= \|\tilde{f}_c - P_{\mathcal{I}}\tilde{f}_c\|^2, \\ \text{subject to} \qquad & \operatorname{rank}(N_{\mathcal{I}}) = |\mathcal{I}|, \\ & \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} P_{\mathcal{I}}\tilde{f}_c \leq 0, \end{aligned} \tag{4.6}$$

where all function arguments have been dropped, and we have used $\tilde{N}_{\mathcal{I}}(x_g) := \Phi^{\mathrm{T}} N_{\mathcal{I}}(x_g)$, $\tilde{f}_c := \Phi^{-1} f_c$, and (4.5), to obtain $N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} R_{\mathcal{I}} f_c = \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \Phi^{-1}\Phi P_{\mathcal{I}} \Phi^{-1} f_c = \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} P_{\mathcal{I}} \tilde{f}_c$.

Now, consider the quadratic program

$$\begin{aligned} \min_{x\in\mathbb{R}^q} & \|\tilde{f}_c - x\|^2, \\ \text{subject to} \qquad & \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} x \leq 0, \end{aligned} \tag{4.7}$$

which is a convex optimization problem with a unique solution [126, p. 218]. We want to show that the unique optimal solution x^* to problem (4.7) is given by $x^* = P_{\mathcal{I}^*}\tilde{f}_c$ for any (not necessarily unique) optimal solution $\mathcal{I}^*$ to subproblem (4.6).

First, recall the definition of the (finitely generated) cone $\mathcal{K}$, generated by the columns of $\tilde{N}_{\mathcal{I}_{\text{sat}}}$ [126, Section 2.12.2.1, p. 146]

$$\mathcal{K} = \{\bar{x} \in \mathbb{R}^q \mid \bar{x} = \tilde{N}_{\mathcal{I}_{\text{sat}}} z, \forall z \geq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}\} \subset \mathbb{R}^q.$$

The *dual*[3] to $\mathcal{K}$, denoted by $\mathcal{K}^*$, is defined by [126, Section 2.13.1, p. 152]

$$\begin{aligned} \mathcal{K}^* &= \{\bar{y} \in \mathbb{R}^q \mid \langle \bar{y}, \bar{x}\rangle \geq 0, \forall \bar{x} \in \mathcal{K}\}, \\ &= \{\bar{y} \in \mathbb{R}^q \mid \langle \bar{y}, \tilde{N}_{\mathcal{I}_{\text{sat}}} z\rangle \geq 0, \forall z \geq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}\}, \\ &= \{\bar{y} \in \mathbb{R}^q \mid \langle \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{y}, z\rangle \geq 0, \forall z \geq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}\}. \end{aligned} \tag{4.8}$$

For any vector $z \geq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}$, the dot product $\langle \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{y}, z\rangle$ is the sum of the elements of the vector $\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{y} \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}$, each multiplied by some non-negative real number (the corresponding non-negative element of z). Clearly, $\langle \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{y}, z\rangle \geq 0$ can hold for *all* $z \geq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}$ if and only if the individual elements of $\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{y}$ are non-negative, i.e. $\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{y} \geq 0$. Hence an equivalent representation for the dual cone $\mathcal{K}^*$ is

$$\mathcal{K}^* = \{\bar{x} \in \mathbb{R}^q \mid \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{x} \geq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}\} \subset \mathbb{R}^q. \tag{4.9}$$

Since the (polyhedral) *polar cone*[3] $\mathcal{K}^\circ$ of $\mathcal{K}$ is the negative dual cone [126, Section 6.5.2.0.1, p. 512, footnote 2.54, p. 152], we have

$$\begin{aligned} \mathcal{K}^\circ &= \{\bar{y} \in \mathbb{R}^q \mid \langle \bar{y}, \bar{x}\rangle \leq 0, \forall \bar{x} \in \mathcal{K}\}, \\ &= \{\bar{x} \in \mathbb{R}^q \mid \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \bar{x} \leq 0 \in \mathbb{R}^{|\mathcal{I}_{\text{sat}}|}\} \subset \mathbb{R}^q. \end{aligned} \tag{4.10}$$

From (4.7), (4.10), and [126, Section E.9, p. 727], we see that the optimal solution x^* to problem (4.7) is the unique projection of $\tilde{f}_c$ onto the polar cone $\mathcal{K}^\circ$. The projection of $\tilde{f}_c$ onto $\mathcal{K}^\circ$, together with the cone $\mathcal{K}$ and its dual $\mathcal{K}^*$, are illustrated in Fig. 4-1.

Next, observe that the polar of $\mathcal{K}^\circ$, i.e. $\mathcal{K}^{\circ\circ}$, satisfy $\mathcal{K}^{\circ\circ} = \mathcal{K}$ for every finitely generated

[3] Some authors define the dual and polar cone in the opposite sense we have adopted, which is the convention used in [126].

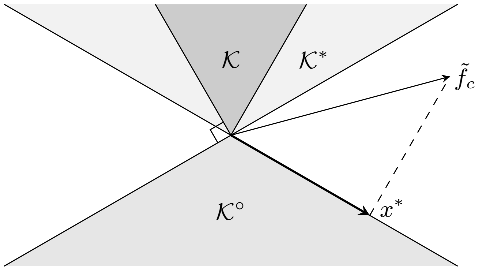


Figure 4-1: The projection x^* of $\tilde{f}_c$ onto the polyhedral cone $\mathcal{K}^\circ$, together with $\mathcal{K}$ and its dual $\mathcal{K}^*$.

cone $\mathcal{K}$ [169, Lemma 2.7.9, p. 54]. Moreover, a theorem due to Minkowski states that every polyhedral cone (e.g. $\mathcal{K}^\circ$) is finitely generated [169, Theorem 2.8.6, p. 55], and a theorem due to Weyl states that every finitely generated cone (e.g. $\mathcal{K}$) is polyhedral [169, Theorem 2.8.8, p. 56]. These allow results applicable to finitely generated cones to be applied to polyhedral cones, and vice versa. We will need the following result from [170].

Proposition 4.1.1 (Cone Projection as Subspace Projection [170, Proposition 2])**.** *Let x^* be the projection of a vector y of $\mathbb{R}^n$ into a convex polyhedral cone $\mathcal{K} = \mathcal{K}(S)$ that is generated by a set $S = \{s_1, \ldots, s_k\}$. Let R be the set of vectors s_i of S orthogonal to $y - x^*$. Then the vector x^* is equal to the projection of y into the subspace $L(R)$ generated by the vectors of R.*

Proposition 4.1.1 shows that projection onto a convex polyhedral cone is equivalent to projection onto a *subspace* spanned by the generating vectors of the face on which the projected point lies. Applied to the convex polyhedral polar cone $\mathcal{K}^\circ$ (permissible due to Minkowski's theorem [169, Theorem 2.8.6, p. 55]), Proposition 4.1.1 shows that the unique solution x^* to problem (4.7) is equal to the projection of $\tilde{f}_c$ onto the subspace spanned by a face of $\mathcal{K}^\circ$ that x^* resides in (including possibly the *face* $\mathcal{K}^\circ$). Observing that the set of candidate solutions $\mathcal{J}$ for subproblem (4.6) is *exhaustive* (see discussion after (2.8) on page 40), and hence must contain a subset $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$ such that $P_{\mathcal{I}}$ projects onto such a subspace, leads to the desired result.

Proposition 4.1.2 (Relation between Solutions of Combinatorial Optimization Subproblem and Convex Quadratic Program)**.** *The unique solution x^* to the convex quadratic program* (4.7) *satisfies $x^* = P_{\mathcal{I}^*}\tilde{f}_c$ for any solution $\mathcal{I}^*$ to the combinatorial optimization subproblems* (4.3) *or* (4.6)*.*

Proof. When $\text{rank}(\tilde{N}_{\mathcal{I}_{\text{sat}}}) = 0$ (which includes the case $\mathcal{I}_{\text{sat}} = \emptyset$ when no constraints are active), the matrix $\tilde{N}_{\mathcal{I}_{\text{sat}}}$ is the zero matrix/vector and the constraint $\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\text{T}} x \leq 0$ of problem (4.7) is automatically satisfied, i.e. $\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\text{T}} x = 0 \cdot x \leq 0$. It is clear that the unique optimal solution x^* to problem (4.7) is $x^* = \tilde{f}_c$. In this case, the constraint $\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\text{T}} P_{\mathcal{I}}\tilde{f}_c \leq 0$ of subproblem (4.6) is also automatically satisfied. Proposition 2.5.2 then shows $\emptyset$ to be an optimal solution to subproblem (4.6) (see also Remark 2.17), so that the objective function satisfies $\|\tilde{f}_c - P_{\mathcal{I}^*}\tilde{f}_c\|^2 = \tilde{F}(\mathcal{I}^*) = \tilde{F}(\emptyset) = 0$ (see (2.44)) for *any* optimal solution $\mathcal{I}^*$ to subproblem (4.6). This implies $P_{\mathcal{I}^*}\tilde{f}_c = \tilde{f}_c$ and $x^* = \tilde{f}_c = P_{\mathcal{I}^*}\tilde{f}_c$ as desired.

When $\text{rank}(\tilde{N}_{\mathcal{I}_{\text{sat}}}) \geq 1$, Proposition 4.1.1 shows that x^* is equal to the projection of $\tilde{f}_c$ onto the subspace L° spanned by a face of $\mathcal{K}^\circ$ that x^* resides in. As shown in [169, Section 2.13, pp. 69 – 70, in particular, equation (2.13.3)], the subspace spanned by any face of $\mathcal{K}^\circ$ is the orthogonal complement of a subspace spanned by some face of its polar $\mathcal{K}$

(due to $\mathcal{K}^{\circ\circ} = \mathcal{K}$ [169, Lemma 2.7.9, p. 54]), and vice versa. Hence the subspace L° is the orthogonal complement of a subspace L spanned by some face of $\mathcal{K}$. Observe that each face of $\mathcal{K}$ is generated by some subset of the columns of $\tilde{N}_{\mathcal{I}_{\text{sat}}}$, so that L is also spanned by some subset of columns of $\tilde{N}_{\mathcal{I}_{\text{sat}}}$. Since the definition of the set of candidate solutions $\mathcal{J}$ is exhaustive (see discussion after (2.8) on page 40), it must include a set of indices $\mathcal{I}_L \subset \mathcal{I}_{\text{sat}}$ such that the columns of $\tilde{N}_{\mathcal{I}_L}$ are linearly independent and spans L. The unique projection of $\tilde{f}_c$ onto the subspace orthogonal to the subspace spanned by the columns of $\tilde{N}_{\mathcal{I}_L}$, i.e. L°, is given by $P_{\mathcal{I}_L}\tilde{f}_c$ [100, Theorem 1]. Hence $x^* = P_{\mathcal{I}_L}\tilde{f}_c$ by Proposition 4.1.1 and $\tilde{F}(\mathcal{I}_L) = \|\tilde{f}_c - P_{\mathcal{I}_L}\tilde{f}_c\|^2 = \|\tilde{f}_c - x^*\|^2$. The minimization in problem (4.6) yields $\|\tilde{f}_c - P_{\mathcal{I}^*}\tilde{f}_c\|^2 = \tilde{F}(\mathcal{I}^*) \leq \tilde{F}(\mathcal{I}_L) = \|\tilde{f}_c - x^*\|^2$ for *any* optimal solution $\mathcal{I}^*$. Uniqueness of the solution x^* of problem (4.7) [126, pp. 218] ensures that $\|\tilde{f}_c - x^*\|^2 < \|\tilde{f}_c - y\|^2$ for all $y \neq x^*$. This, together with $\|\tilde{f}_c - P_{\mathcal{I}^*}\tilde{f}_c\|^2 \leq \|\tilde{f}_c - x^*\|^2$, shows that $x^* = P_{\mathcal{I}^*}\tilde{f}_c$, as desired. ■

Remark 4.1. Proposition 4.1.2 shows that even if subproblem (4.6) has no unique solutions, the projection of $\tilde{f}_c$ for *any* solution $\mathcal{I}^*$, namely $P_{\mathcal{I}^*}\tilde{f}_c$, *is unique*. Since Φ is nonsingular, it implies the uniqueness of $R_{\mathcal{I}^*}f_c = \Phi P_{\mathcal{I}^*}\tilde{f}_c$ as well. □

Proposition 4.1.2 shows that the GPAW-compensated controller (4.1) can be realized equivalently as

$$\begin{aligned} \dot{x}_g &= \Phi x^*(x_g, y, r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g), \end{aligned} \tag{4.11}$$

where at each fixed time (so that (x_g, y, r) are fixed), $x^*(x_g, y, r)$ is the unique optimal solution to the projection of $\Phi^{-1}f_c(x_g, y, r)$ onto the convex polyhedral cone $\mathcal{K}^\circ$ (4.10), or the quadratic program (4.7), rewritten as

$$\begin{aligned} &\min_{x \in \mathbb{R}^q} \|\Phi^{-1}f_c(x_g, y, r) - x\|^2, \\ &\text{subject to} \qquad N^{\mathrm{T}}_{\mathcal{I}_{\text{sat}}}(x_g)\Phi x \leq 0. \end{aligned} \tag{4.12}$$

This formulation may be useful for implementation due to availability of solvers and algorithms. However, most of the inherent structure of the GPAW-compensated controller is then concealed by this representation, which renders it ill suited for further analysis.

Numerous software packages are available[4] to solve the convex quadratic program (4.12), for which an example is the MATLAB®[5] function `quadprog` used in Section B.1 of Appendix B. For algorithms to project onto convex polyhedral cones, see [170–175] and the references therein. See also Appendix C for a summary of available methods to realize the GPAW-compensated controller.

4.1.1 Initializing the Quadratic Program

Convex optimization problems with a guaranteed *unique* solution like problem (4.12) have the advantage that their numerical solutions are insensitive to the initial guess [176, Section 1.4, pp. 9 – 11]. However, the computation time still depends to some extent on how far the initial guess is from the global optimum. When realizing the GPAW-compensated controller (4.11), the quadratic program (4.12) needs to be solved *at each point in time*, and

[4] See `http://www.numerical.rl.ac.uk/qp/qp.html` for a list of available quadratic program solvers.

[5] See `http://www.mathworks.com`.

an initial guess can usually be provided to the solver to aid its solution. Here, we provide some guidelines on initializing the quadratic program solver.

First, for any well-designed system and reference input, the system should be operating in the interior of the unsaturated region for the majority of time. Hence $\mathcal{I}_{\text{sat}} = \emptyset$ and the unique optimal solution to subproblem (4.12) would be $x^*(x_g, y, r) = \Phi^{-1} f_c(x_g, y, r)$. In this case, the GPAW-compensated controller (4.11) is fully defined and there is no need to solve subproblem (4.12).

Now, observe that apart from switchings on the saturation constraint boundaries causing the active constraint set $\mathcal{I}_{\text{sat}}$ to change, the matrix $N_{\mathcal{I}_{\text{sat}}}(x_g)$ in (4.12) would be of constant dimension and continuous in x_g. These switchings are instantaneous and insignificant[6] compared to the time when switchings do not occur. Hence for the most part, $N_{\mathcal{I}_{\text{sat}}}(x_g)$ is a constant dimensioned continuous matrix, which implies that the optimal solution $x^*(x_g, y, r)$ will vary continuously with (x_g, y, r). This motivates setting the initial guess to the last optimal solution. In other words, let t_1, t_2 $(t_1 < t_2)$, be two successive times when subproblem (4.12) needs to be solved, and let $x^*(x_g(t_1), y(t_1), r(t_1))$ be the optimal solution at time t_1. When solving subproblem (4.12) at time t_2, we set the initial guess to be $x_{ig}(t_2) = x^*(x_g(t_1), y(t_1), r(t_1))$, with the expectation that the true optimal solution will be close to $x_{ig}(t_2)$ except possibly at switching times.

In summary, we have

- when operating in the *interior* of the unsaturated region, set the optimal solution to subproblem (4.12) as $x^*(x_g, y, r) = \Phi^{-1} f_c(x_g, y, r)$;
- when operating on the saturation constraint boundaries or in the saturated region, set the initial guess x_{ig} to the last optimal solution.

4.2 A Geometric Bounding Condition

Here, we present a property of GPAW-compensated controllers when the unsaturated region is a *star domain*. First, we describe *star domains*, which are generalizations of *convex sets*.[7] For any two points $x_1, x_2 \in \mathbb{R}^n$, let $l(x_1, x_2)$ be the closed[8] line segment connecting them,

$$l(x_1, x_2) := \{\bar{x} \in \mathbb{R}^n \mid \bar{x} = \theta x_1 + (1-\theta) x_2, \forall \theta \in [0, 1]\}.$$

Definition 4.1 (Kernel of a Set [178, Definition 1.4, p. 5], [179])**.** Let $X \subset \mathbb{R}^n$ be a nonempty set. The *kernel* of X, denoted by $\ker(X)$, is defined by

$$\ker(X) := \{\bar{x} \in \mathbb{R}^n \mid l(\bar{x}, y) \subset X, \forall y \in X\} \subset X. \qquad \square$$

In other words, the kernel of X, $\ker(X)$, is the set of points for which every point within $\ker(X)$ can be connected to every other point in X by a straight line contained within X. It is clear that $\ker(X)$ must be a (possibly non-proper) subset of X.

Definition 4.2 (Star Domain [178, Definition 1.2, p. 4], [179])**.** A nonempty set $X \subset \mathbb{R}^n$ is a *star domain* (or *star-shaped*) if $\ker(X) \neq \emptyset$. $\square$

In other words, a nonempty set X is a star domain if there exists at least one point $x \in X$ such that for every $y \in X$, the line segment connecting x and y is contained within X.

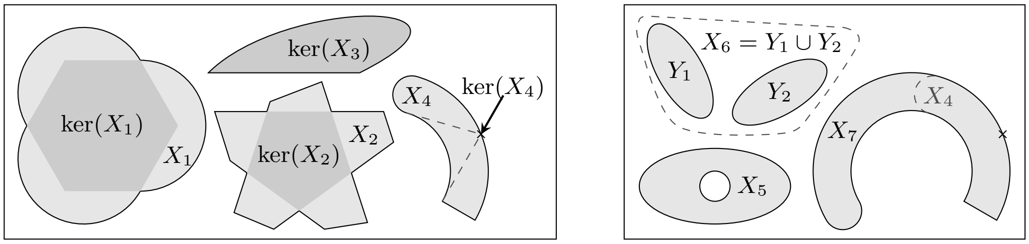
[6] In fact, the set of switching times would be a set of measure zero [154, p. 309].

[7] Recall that a set $X \in \mathbb{R}^n$ is *convex* if $\theta x_1 + (1-\theta) x_2 \in X$ whenever $x_1, x_2 \in X$ and $\theta \in [0, 1]$ [177, p. 10].

[8] Note that $l(\alpha, \beta)$ is defined as an *open* line segment in Section 3.5.

Remark 4.2. Clearly, any convex set X (including the case $X = \mathbb{R}^n$) is also a star domain with $\ker(X) = X$. For any non-convex star domain, $\ker(X)$ is a strict subset of X. □

In $\mathbb{R}^2$, Fig. 4-2(a) shows examples of star domains while Fig. 4-2(b) shows sets which are not star domains. It is clear from Definitions 4.2 and 4.1 that star domains must be simply connected,[9] as reflected in Fig. 4-2(a). In Fig. 4-2(a), X_1 and X_2 are non-convex star domains, X_3 is convex with $\ker(X_3) = X_3$, while X_4 is a non-convex star domain with the single point indicated as its kernel. In Fig. 4-2(b), sets X_5 and X_6 are not simply connected, and hence are not star domains. Notice that X_7 is a simply connected set obtained by "extending" the star domain X_4. However, X_7 has an empty kernel and hence is not a star domain. The corollary to the next result (Corollary 4.2.2) will be needed in the sequel.



(a) star domains, $\ker(X_i) \neq \emptyset$.

(b) not star domains, $\ker(X_i) = \emptyset$.

Figure 4-2: Examples and counterexamples of star domains in $\mathbb{R}^2$.

Lemma 4.2.1 (Cartesian Product of Star Domains)**.** *The sets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ are star domains if and only if $X \times Y$ is a star domain in $\mathbb{R}^{n+m}$ with kernel $\ker(X \times Y) = \ker(X) \times \ker(Y)$.*

Proof. ($\Rightarrow$) We first show that *if* $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ are star domains, *then* $X \times Y$ is a star domain in $\mathbb{R}^{n+m}$ with kernel $\ker(X \times Y) = \ker(X) \times \ker(Y)$. Assume X and Y are star domains, so that $\ker(X) \times \ker(Y) \neq \emptyset$. It is sufficient to show that

(i) $\ker(X \times Y) \supset \ker(X) \times \ker(Y)$ (which implies $\ker(X \times Y) \neq \emptyset$ and hence $X \times Y$ is a star domain); and

(ii) $\ker(X \times Y) \subset \ker(X) \times \ker(Y)$ (which, together with (i) means $\ker(X \times Y) = \ker(X) \times \ker(Y)$).

First, we show (i). Let $(x, y) \in \ker(X) \times \ker(Y)$ with $x \in \ker(X)$ and $y \in \ker(Y)$. We need to show that $(x, y) \in \ker(X \times Y)$. Definition 4.1 applied to the individual star domains X and Y yields

$$\begin{aligned} l(x, \tilde{x}) &\subset X, \qquad \forall \tilde{x} \in X, \\ l(y, \tilde{y}) &\subset Y, \qquad \forall \tilde{y} \in Y, \end{aligned} \tag{4.13}$$

which holds if and only if

$$l((x, y), (\tilde{x}, \tilde{y})) \subset X \times Y, \qquad \forall (\tilde{x}, \tilde{y}) \in X \times Y. \tag{4.14}$$

This means $(x, y) \in \ker(X \times Y)$ and concludes (i).

[9] Recall that a *simply connected* set X is such that every closed curve in X can be continuously contracted into a point without leaving X [125, Section 4.3-6, p. 90].

Next, we show (ii). Let $(x, y) \in \ker(X \times Y)$. We need to show that $(x, y) \in \ker(X) \times \ker(Y)$. Definition 4.1 applied to the star domain $X \times Y$ yields (4.14), which holds if and only if (4.13) holds. This implies $x \in \ker(X)$, $y \in \ker(Y)$, $(x, y) \in \ker(X) \times \ker(Y)$, and hence the conclusion of (ii) and ($\Rightarrow$).

($\Leftarrow$) To show the converse, assume $X \times Y \in \mathbb{R}^{n+m}$ is a star domain with kernel $\ker(X) \times \ker(Y)$. Since $X \times Y$ is a star domain, its kernel is non-empty. This implies $\ker(X) \neq \emptyset$ and $\ker(Y) \neq \emptyset$. Hence X and Y are star domains, as desired. ■

Corollary 4.2.2. *The set $X \subset \mathbb{R}^n$ is a star domain if and only if $X \times \mathbb{R}^m$ is a star domain in $\mathbb{R}^{n+m}$ with kernel $\ker(X \times \mathbb{R}^m) = \ker(X) \times \mathbb{R}^m$.*

When a star domain is defined by a set of constraint functions, the following gives a characterization of the gradient vectors of the constraint functions on the boundary of the star domain. It states that any vector pointing “out” from the star domain must lie in the cone generated by the gradient vectors of the active constraints.

Lemma 4.2.3 (Gradient Direction Bounds on Star Domain Boundary)**.** *Let X be defined by a set of m constraints*

$$X = \{\bar{x} \in \mathbb{R}^n \mid \tilde{h}_i(\bar{x}) \leq 0, \forall i \in \mathcal{I}_m\} \subset \mathbb{R}^n,$$

for some functions $\tilde{h}_i \colon \mathbb{R}^n \to \mathbb{R}$, $i \in \mathcal{I}_m$. For any boundary point $x \in \partial X$, define the set of indices of active constraints

$$\mathcal{I}_{\lim}(x) := \{i \in \mathcal{I}_m \mid \tilde{h}_i(x) = 0\}.$$

If X is a star domain, then for any $x_{ker} \in \ker(X)$ and any boundary point $x \in \partial X$, we have

$$\langle x - x_{ker}, \nabla \tilde{h}_i(x) \rangle \geq 0, \qquad \forall i \in \mathcal{I}_{\lim}(x).$$

Proof. From the definition of $\ker(X)$ (Definition 4.1), we have $y(\theta) := \theta x + (1-\theta)x_{ker} \in X$ for all $\theta \in [0,1]$. Hence $\tilde{h}_i(y(\theta)) \leq 0$ for all $\theta \in [0,1]$ and all $i \in \mathcal{I}_m$. View $\tilde{h}_i(y(\theta))$ as a function of θ with x, x_{ker} fixed. Since $x \in \partial X$, we have $\tilde{h}_i(y(1)) = \tilde{h}_i(x) = 0$ for all $i \in \mathcal{I}_{\lim}(x)$. Since $\tilde{h}_i(y(\theta)) \leq 0$ for all $\theta \in [0,1]$, $\tilde{h}_i(y(\theta))$ must be non-decreasing at $\theta = 1$. Hence by the chain rule,

$$\frac{d\tilde{h}_i}{d\theta}(y(\theta)) = \frac{\partial \tilde{h}_i}{\partial x}(y(\theta))\frac{dy}{d\theta} = \frac{\partial \tilde{h}_i}{\partial x}(y(\theta))(x - x_{ker}) \geq 0, \qquad \forall i \in \mathcal{I}_{\lim}(x),$$

at $\theta = 1$. This gives $\frac{d\tilde{h}_i}{dx}(y(1))(x - x_{ker}) = \frac{d\tilde{h}_i}{dx}(x)(x - x_{ker}) \geq 0$ for all $i \in \mathcal{I}_{\lim}(x)$, which can be written in dot product form with the gradient vector as stated. ■

For the next result, observe that any solution $\mathcal{I}^*$ of subproblem (4.3) or (4.6) satisfies $\text{rank}(N_{\mathcal{I}^*}(x_g)) = |\mathcal{I}^*|$. This ensures $R_{\mathcal{I}^*}(x_g, y, r)$, $P_{\mathcal{I}^*}(x_g, y, r)$, and $S_{\mathcal{I}^*}(x_g, y, r)$, are well-defined. Recall the definition of the unsaturated region (2.33)

$$K = \{\bar{x} \in \mathbb{R}^q \mid h_i(\bar{x}) \leq 0, \forall i \in \mathcal{I}_{2m}\},$$

where h_i are the saturation constraint functions (2.28)

$$\begin{aligned} h_i(x_g) &= g_{ci}(x_g) - u_{\max,i}, \\ h_{i+m}(x_g) &= -g_{ci}(x_g) + u_{\min,i}, \end{aligned} \qquad \forall i \in \mathcal{I}_m := \{1, 2, \ldots, m\}.$$

The following is the main result of this section, which gives a geometric bounding condition relating the vector fields of the GPAW-compensated controller and uncompensated controller when the unsaturated region is a star domain.[10]

Theorem 4.2.4 (Geometric Bounding Condition)**.** *If the unsaturated region $K \subset \mathbb{R}^q$ is a star domain, then for any $x \in K$ and any $x_{ker} \in \ker(K)$,*

$$\langle \Gamma^{-1}(x - x_{ker}), \tilde{f}_g(x,y,r)\rangle \leq \langle \Gamma^{-1}(x - x_{ker}), f_c(x,y,r)\rangle, \tag{4.15}$$

holds for all $(y,r) \in \mathbb{R}^{p+n_r}$, where $\tilde{f}_g(x,y,r) := R_{\mathcal{I}^*}(x,y,r)f_c(x,y,r)$ and $f_c(x,y,r)$ are the vector fields of the GPAW-compensated controller (4.1) *and uncompensated controller* (2.26) *respectively, and $\Gamma = \Gamma^{\mathrm{T}} > 0$ is the GPAW parameter.*

Proof. When x is in the interior of K, i.e. $h_i(x) < 0$ holds with *strict* inequality for all $i \in \mathcal{I}_{2m}$, $\mathcal{I}^* = \emptyset$ is the unique optimal solution to subproblems (4.3) and (4.6) (see Remark 2.17). By the definition of $R_{\mathcal{I}^*}(x,y,r)$ (see (4.2)), condition (4.15) holds with equality. It remains to show that (4.15) holds when x is on the boundary of K, i.e. $x \in \partial K$.

Condition (4.15) can be written as $\langle \Gamma^{-1}(x - x_{ker}), f_c(x,y,r) - \tilde{f}_g(x,y,r)\rangle \geq 0$, which, using $\Gamma = \Gamma^{\mathrm{T}} = \Phi\Phi^{\mathrm{T}}$ and (4.5), simplifies further to

$$\begin{aligned}(x - x_{ker})^{\mathrm{T}}\Gamma^{-1}(f_c - R_{\mathcal{I}^*}f_c) &= (x - x_{ker})^{\mathrm{T}}\Phi^{-\mathrm{T}}\Phi^{-1}(I - R_{\mathcal{I}^*})f_c,\\ &= (x - x_{ker})^{\mathrm{T}}\Phi^{-\mathrm{T}}\Phi^{-1}(I - \Phi P_{\mathcal{I}^*}\Phi^{-1})f_c,\\ &= (x - x_{ker})^{\mathrm{T}}\Phi^{-\mathrm{T}}(I - P_{\mathcal{I}^*})\Phi^{-1}f_c,\\ &= \tilde{x}^{\mathrm{T}}(I - P_{\mathcal{I}^*})\tilde{f}_c \geq 0, \end{aligned} \tag{4.16}$$

where $\tilde{x} := \Phi^{-1}(x - x_{ker})$, $\tilde{f}_c := \Phi^{-1}f_c(x,y,r)$, and all function arguments have been dropped.

Proposition 4.1.2 shows that $x^* = P_{\mathcal{I}^*}\tilde{f}_c$ where x^* is the unique optimal solution to problem (4.7). Since x^* is the projection of $\tilde{f}_c$ onto the convex polyhedral cone $\mathcal{K}^\circ$ (4.10), it satisfies $\tilde{f}_c - x^* \in \mathcal{K}^{\circ\circ}$ [180, Proposition 3.2.3, p. 50] where $\mathcal{K}^{\circ\circ} = \mathcal{K}$ [169, Lemma 2.7.9, p. 54]. Hence we have $\tilde{f}_c - x^* \in \mathcal{K}$. Using $x^* = P_{\mathcal{I}^*}\tilde{f}_c$ (by Proposition 4.1.2), we have

$$\tilde{f}_c - x^* = \tilde{f}_c - P_{\mathcal{I}^*}\tilde{f}_c = (I - P_{\mathcal{I}^*})\tilde{f}_c \in \mathcal{K}. \tag{4.17}$$

Lemma 4.2.3 applied to the (star) unsaturated region K yields $\langle x - x_{ker}, \nabla h_i(x)\rangle \geq 0$ for all $i \in \mathcal{I}_{\text{sat}}$, where $\mathcal{I}_{\text{sat}}$ (4.4) is the set of indices of active saturation constraints. Using $\tilde{N}_{\mathcal{I}_{\text{sat}}} := \Phi^{\mathrm{T}}N_{\mathcal{I}_{\text{sat}}}$ as defined for problem (4.7), this means (see the definition of $N_{\mathcal{I}}$ in terms of h_i in (4.4))

$$N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}}(x - x_{ker}) = N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}}\Phi\Phi^{-1}(x - x_{ker}) = \tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}}\tilde{x} \geq 0,$$

which implies that $\tilde{x}$ is in the dual of $\mathcal{K}$ (see (4.9)), i.e. $\tilde{x} \in \mathcal{K}^*$. From the definition of the dual cone $\mathcal{K}^*$ (4.8), any $\tilde{y} \in \mathcal{K}$ satisfies $\langle \tilde{y}, \tilde{x}\rangle \geq 0$. Since $(I - P_{\mathcal{I}^*})\tilde{f}_c \in \mathcal{K}$ by (4.17) and $\tilde{x} \in \mathcal{K}^*$, we have (4.16) (and hence (4.15)) as desired. ■

Remark 4.3. When x_{ker} is not in $\ker(K)$, it can be shown that condition (4.15) of Theorem 4.2.4 holds for all points $x \in K$ such that the line segment $l(x, x_{ker})$ is contained within

[10] Strictly speaking, these are time varying (or input-dependent) vector fields. In Theorem 4.2.4, we use $\tilde{f}_g$ to denote the vector field of the GPAW-compensated controller. In the next section, as in Chapter 3, f_g denotes the vector field of the GPAW-compensated *closed-loop* system.

K, i.e. $l(x, x_{ker}) \subset K$. For *any* fixed $x_{ker} \in K \setminus \ker(K)$, condition (4.15) holds for all x in any convex subset of K containing x_{ker}. □

Remark 4.4. Clearly, if the output function g_c in (4.1) is linear, i.e. $g_c(x_g) = C_c x_g$ for some constant matrix $C_c \in \mathbb{R}^{m \times q}$, the induced unsaturated region $K = \{\bar{x} \in \mathbb{R}^q \mid u_{\min} \leq C_c \bar{x} \leq u_{\max}\}$ is a convex polyhedral set [177, p. 170] and hence a star domain. □

The geometric condition (4.15) of Theorem 4.2.4 is illustrated in Fig. 4-3 with an example non-convex star domain for the unsaturated region. Theorem 4.2.4 states that the projection of $\tilde{f}_g$ onto the vector $\Gamma^{-1}(x - x_{ker})$ is bounded above by the projection of f_c onto the same vector. This is clear in Fig. 4-3 for the two cases illustrated when $\Gamma = I$.

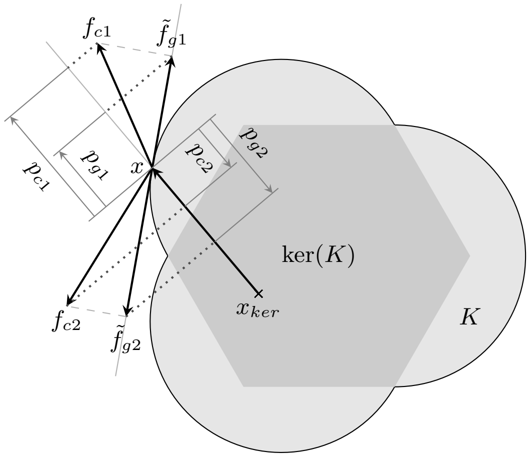


Figure 4-3: Illustration of the geometric bounding condition of Theorem 4.2.4 with an example non-convex star domain and $\Gamma = I$. The projection of $\tilde{f}_g$ onto $x - x_{ker}$ is always less positive ($0 < p_{g1} < p_{c1}$) or more negative ($p_{g2} < p_{c2} < 0$) than the projection of f_c onto the same vector.

4.3 A General Region of Attraction Comparison Result

Our next goal is to derive some stability results for the *regulatory* GPAW-compensated system *relative* to the uncompensated system (see Section 3.7) as suggested by the geometric bound of Theorem 4.2.4 and Fig. 4-3. Being regulatory systems imply that the closed-loop systems are *autonomous* and presents some simplifications. In particular, the region of attraction (ROA) is well defined for asymptotically stable equilibrium points. In this section, we present a general result that allows the size of the ROA of an autonomous system to be inferred from that of a related autonomous system. These results are similar in principle to those of [181, Section 3.4, pp. 134 – 140], where general stability properties (not restricted to ROA properties) for a system are inferred from those of a *comparison system*.

First, we recall some basic definitions. Consider the n-th order autonomous system

$$\dot{x} = f(x), \tag{4.18}$$

where for some domain $D \subset \mathbb{R}^n$, we assume the function $f \colon D \to \mathbb{R}^n$ satisfies some regularity conditions to ensure existence and uniqueness of solutions for all *forward* times, or for all times when appropriate. When f is continuous, several classical results based on Lipschitz continuity of f, e.g. [37, Section 3.1, pp. 88 –95], [135, pp. 12 – 25], are available to ensure existence and uniqueness of solutions for all (forward and reverse) times. In particular, global Lipschitz continuity of f ensures existence and uniqueness of solutions to system (4.18) for all times [37, Theorem 3.2, p. 93]. If f is *discontinuous*, regularity conditions are available to ensure existence and uniqueness of solutions for all *forward* times [120, §2.7, pp. 75 – 86, and §2.10, pp. 106 – 117].

Let $\phi(t, x_0)$ be the unique solution of system (4.18) starting from x_0 at time $t = 0$. A set M is said to be an *invariant set* for system (4.18) if the solution starting within M stays in M for *all* times, i.e. $x_0 \in M \Rightarrow (\phi(t, x_0) \in M, \forall t \in \mathbb{R})$ [37, p. 127]. A set M is a *positively invariant set* for system (4.18) if the solution starting within M stays in M for all *forward* times, i.e. $x_0 \in M \Rightarrow (\phi(t, x_0) \in M, \forall t \geq 0)$ [37, p. 127].

We assume that $x_{eq} \in D$ is an asymptotically stable equilibrium point for system (4.18). The ROA of the equilibrium x_{eq} for system (4.18) is defined by [37, p. 314]

$$R_A(x_{eq}) := \{\bar{x} \in \mathbb{R}^n \mid \phi(t, \bar{x}) \to x_{eq} \text{ as } t \to \infty\}.$$

In other words, the ROA is the set of all points such that the solution starting from any point within it converges to the equilibrium. Clearly, $R_A(x_{eq})$ must contain the equilibrium x_{eq}. As shown in [37, Lemma 8.1, p. 314], the ROA is an open connected invariant set. This implies that there exists a sufficiently small neighborhood of x_{eq} contained in the ROA $R_A(x_{eq})$. A set $\Omega \subset R_A(x_{eq})$ is said to be an *estimate* of $R_A(x_{eq})$ if every solution starting in Ω approaches x_{eq} as $t \to \infty$ [37, p. 316], i.e. $x_0 \in \Omega \Rightarrow \lim_{t\to\infty} \phi(t, x_0) = x_{eq}$. Clearly, a set is an estimate of $R_A(x_{eq})$ if and only if it is a subset of $R_A(x_{eq})$ (including the *subset* $R_A(x_{eq})$). Numerous methods are available to estimate ROAs [155–166]. Our purpose is *not* to estimate ROAs, but, *given* an ROA estimate for a system, to find conditions for which the same ROA estimate is valid for a related system. This is similar in objective to Proposition 3.5.13, but significantly weaker.[11] We will need the following definition of class $\mathcal{K}$ functions, for which some useful properties are presented in [37, Section 4.4, pp. 144 – 147]. Note that class $\mathcal{K}$ functions are *not* related to the polyhedral cone $\mathcal{K}$ in Section 4.1.

Definition 4.3 (Class $\mathcal{K}$ Functions [37, Definition 4.2, p. 144])**.** A continuous function $\alpha\colon [0, a) \to [0, \infty)$ is said to belong to class $\mathcal{K}$ if it is strictly increasing and $\alpha(0) = 0$. □

An ROA estimate $\Omega \subset R_A(x_{eq})$ is said to be associated with a Lyapunov function $V\colon \Omega \to \mathbb{R}$ for system (4.18) if

$$\begin{aligned} V(x_{eq}) = 0, \qquad V(\bar{x}) > 0, \qquad & \forall \bar{x} \in \Omega \setminus \{x_{eq}\}, \\ \dot{V}(\bar{x}) = \frac{\partial V(\bar{x})}{\partial x} f(\bar{x}) \leq -\alpha(\|\bar{x} - x_{eq}\|), \qquad & \forall \bar{x} \in \Omega, \end{aligned} \tag{4.19}$$

for some class $\mathcal{K}$ function $\alpha\colon [0, a) \to [0, \infty)$, where $a := \sup_{\bar{x}\in\Omega} \|\bar{x} - x_{eq}\|$. In this case, we indicate the association by denoting such an estimate by Ω_V. Note that numerous ROA estimation methods yield estimates that are associated with Lyapunov functions.

The next result gives sufficient conditions for an ROA estimate of a system associated with some Lyapunov function to be a valid ROA estimate for a related system, i.e. the estimate is contained within the ROA of this related system.

[11] Proposition 3.5.13 states that the ROA of the GPAW-compensated system contains the ROA of the uncompensated system, which is *independent* of any Lyapunov functions. The results in this section are associated with some particular (non-unique) Lyapunov functions. See also Remark 3.14.

Lemma 4.3.1. *Consider two n-th order autonomous systems*

$$\dot{x} = f_1(x), \tag{4.20}$$
$$\dot{x} = f_2(x). \tag{4.21}$$

For some $D \subset \mathbb{R}^n$, assume $f_1 \colon D \to \mathbb{R}^n$ and $f_2 \colon D \to \mathbb{R}^n$ are such that solutions $\phi_1(t, x_0)$ and $\phi_2(t, x_0)$ to systems (4.20) *and* (4.21) *respectively exist and are unique for all $x_0 \in D$ and all $t \geq 0$. Assume $x_{eq} \in D$ is an asymptotically stable equilibrium point for both systems, and let $R_{A1}(x_{eq})$, $R_{A2}(x_{eq})$ be the ROAs of systems* (4.20) *and* (4.21) *respectively. Let $\Omega_V \subset R_{A1}(x_{eq})$ be an estimate of $R_{A1}(x_{eq})$ associated with a Lyapunov function $V \colon \Omega_V \to \mathbb{R}$ satisfying* (4.19) *(with f_1, Ω_V in place of f, Ω). If Ω_2 is a subset of Ω_V (possibly $\Omega_2 = \Omega_V$) and also a positively invariant set for system* (4.21)*, and in addition,*

$$\frac{\partial V(\bar{x})}{\partial x} f_2(\bar{x}) \leq \frac{\partial V(\bar{x})}{\partial x} f_1(\bar{x}), \qquad \forall \bar{x} \in \Omega_2, \tag{4.22}$$

then Ω_2 is an estimate of $R_{A2}(x_{eq})$, i.e. $\Omega_2 \subset R_{A2}(x_{eq})$. Moreover, x_{eq} is in the closure of Ω_2, i.e. $x_{eq} \in \bar{\Omega}_2 \subset \bar{\Omega}_V$.

Proof. The proof has some similarities with the proof of [104, Lemma 1]. We need to show that $\lim_{t\to\infty} \phi_2(t, x_0) = x_{eq}$ for all $x_0 \in \Omega_2$. Since x_{eq} is an asymptotically stable equilibrium for system (4.21), $R_{A2}(x_{eq})$ is non-empty and necessarily contains x_{eq}, i.e. $x_{eq} \in R_{A2}(x_{eq}) \neq \emptyset$. Since $R_{A2}(x_{eq})$ is open [37, Lemma 8.1, p. 314], it contains some sufficiently small neighborhood of x_{eq}. Hence there exists a sufficiently small $\epsilon > 0$ such that $B_\epsilon := \{\bar{x} \in \mathbb{R}^n \mid V(\bar{x}) \leq \epsilon\} \subset R_{A2}(x_{eq})$. Fix any such $\epsilon > 0$.

From $\Omega_2 \subset \Omega_V$, (4.22), and (4.19), we have

$$\frac{\partial V(\bar{x})}{\partial x} f_2(\bar{x}) \leq \frac{\partial V(\bar{x})}{\partial x} f_1(\bar{x}) \leq -\alpha(\|\bar{x} - x_{eq}\|), \qquad \forall \bar{x} \in \Omega_2.$$

Let ∂B_ϵ be the boundary of B_ϵ, i.e. $\partial B_\epsilon = \{\bar{x} \in \mathbb{R}^n \mid V(\bar{x}) = \epsilon\}$, $\beta := \min_{\bar{x} \in \partial B_\epsilon} \|\bar{x} - x_{eq}\|$, and $\delta := \alpha(\beta)$. Clearly, we have $\beta > 0$ and $\delta > 0$ due to $\epsilon > 0$ and (4.19). Then for all $\bar{x} \in \Omega_2 \setminus B_\epsilon$, we have $\|\bar{x} - x_{eq}\| \geq \beta$, $\alpha(\|\bar{x} - x_{eq}\|) \geq \alpha(\beta) = \delta$, and

$$\frac{\partial V(\bar{x})}{\partial x} f_2(\bar{x}) \leq -\alpha(\|\bar{x} - x_{eq}\|) \leq -\delta, \qquad \forall \bar{x} \in \Omega_2 \setminus B_\epsilon. \tag{4.23}$$

Let $x_0 \in \Omega_2$. Since Ω_2 is a positively invariant set for system (4.21), we have $\phi_2(t, x_0) \in \Omega_2$ for all $t \geq 0$. The Lyapunov function V evaluated along $\phi_2(t, x_0)$ is given by

$$V(\phi_2(t, x_0)) = V(x_0) + \int_0^t \frac{\partial V(\phi_2(\tau, x_0))}{\partial x} f_2(\phi_2(\tau, x_0))\, d\tau, \qquad \forall t \geq 0.$$

Clearly, $\phi_2(t, x_0)$ can remain in $\Omega_2 \setminus B_\epsilon$ only for a finite amount of time. Otherwise, $\phi_2(t, x_0) \in \Omega_2 \setminus B_\epsilon$ for all $t \geq 0$ and (4.23) imply

$$V(\phi_2(t, x_0)) \leq V(x_0) - \delta \int_0^t d\tau = V(x_0) - \delta t, \qquad \forall t \geq 0,$$

which shows that $V(\phi_2(t, x_0)) < 0$ for sufficiently large t, a contradiction to V being positive definite. Hence $\phi_2(t, x_0)$ must enter B_ϵ at some finite time, and must approach x_{eq} since

$B_\epsilon \subset R_{A2}(x_{eq})$, i.e. $\lim_{t\to\infty} \phi_2(t, x_0) = x_{eq}$.

Finally, since Ω_2 is a positively invariant set for system (4.21), $\phi_2(t, x_0) \in \Omega_2$ for all $t \geq 0$. Since $\lim_{t\to\infty} \phi_2(t, x_0) = x_{eq}$, the equilibrium x_{eq} must lie in the closure of Ω_2, and hence also the closure of Ω_V, i.e. $x_{eq} \in \bar{\Omega}_2 \subset \bar{\Omega}_V$. ■

Remark 4.5. Note that the ROA estimate Ω_V in Lemma 4.3.1 need not be compact.[12] Moreover, the conclusion $\Omega_2 \subset R_{A2}(x_{eq})$ shows that the (possibly unknown) ROA $R_{A2}(x_{eq})$ is at least as large as Ω_2. □

Lemma 4.3.1 will be used in the next section to derive some ROA comparison results specific to GPAW-compensated systems.

4.4 Region of Attraction Comparison for GPAW-Compensated Systems

For any well-posed anti-windup problem, the input-constrained nominal system must be stable in some sense at least locally, e.g. locally asymptotically stable with respect to some equilibrium. When it is indeed locally asymptotically stable with respect to some equilibrium, the fundamental question is whether an anti-windup scheme can maintain/enlarge the ROA while achieving performance enhancements in the presence of control saturation (see also Section 3.5).

In this section, we use Lemma 4.3.1 to derive some results on the ROA of the *regulatory* GPAW-compensated system relative to the uncompensated system (see Section 3.7). First, we describe the closed-loop systems, for which we recall the open-loop plant (1.1) and uncompensated controller (2.26),

$$\begin{aligned} \dot{x} &= f(x, \mathrm{sat}(u)), \\ y &= g(x, \mathrm{sat}(u)), \end{aligned} \qquad \text{and} \qquad \begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \\ u_c &= g_c(x_c), \end{aligned}$$

respectively. Making appropriate substitutions, the nominal uncompensated system with $u = u_c$ can be written as

$$\Sigma_n \colon \begin{cases} \dot{x} = f(x, \mathrm{sat}(g_c(x_c))), \\ \dot{x}_c = f_c(x_c, g(x, \mathrm{sat}(g_c(x_c))), r), \end{cases} \qquad \text{or} \qquad \Sigma_n \colon \dot{z}_n = f_n(z_n), \tag{4.24}$$

while the GPAW-compensated system comprising the plant and GPAW-compensated controller (4.1) (with $u = u_g$) is described by

$$\Sigma_g \colon \begin{cases} \dot{x} = f(x, \mathrm{sat}(g_c(x_g))), \\ \dot{x}_g = R_{\mathcal{I}^*} f_c(x_g, g(x, \mathrm{sat}(g_c(x_g))), r), \end{cases} \qquad \text{or} \qquad \Sigma_g \colon \dot{z}_g = f_g(z_g), \tag{4.25}$$

where $R_{\mathcal{I}^*} f_c(x_g, y, r) := R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r)$. In (4.24), $f_n \colon \mathbb{R}^{n+q} \to \mathbb{R}^{n+q}$ is the vector field of the nominal system with state $z_n := (x, x_c)$, while in (4.25), $f_g \colon \mathbb{R}^{n+q} \to \mathbb{R}^{n+q}$ is the vector field of the GPAW-compensated system with state $z_g := (x, x_g)$.

[12] Recall the Heine-Borel theorem [154, Theorem 2.41, p. 40] which states that in finite dimensional Euclidean spaces like $\mathbb{R}^n$, a set is compact if and only if it is closed and bounded.

Theorem 2.5.3 (controller state-output consistency) shows that the unsaturated region[13]

$$\mathbb{R}^n \times K, \qquad K := \{\bar{x} \in \mathbb{R}^q \mid \mathrm{sat}(g_c(\bar{x})) = g_c(\bar{x})\}, \tag{4.26}$$

is a positively invariant set [37, p. 127] for the GPAW-compensated system Σ_g (see also Remark 2.18). When the controller state is initialized such that $x_g(0) \in K$ (see Remark 2.19), Theorem 2.5.3 shows that the GPAW-compensated system can be written as

$$\Sigma_{gu}\colon \begin{cases} \dot{x} = f(x, g_c(x_g)), \\ \dot{x}_g = R_{\mathcal{I}^*} f_c(x_g, g(x, g_c(x_g)), r), \end{cases} \qquad \text{or} \qquad \Sigma_{gu}\colon \dot{z}_g = f_{gu}(z_g), \tag{4.27}$$

where the saturation function $\mathrm{sat}(\cdot)$ has been eliminated.[14] For this system, comparison would be made against the *unconstrained* system

$$\Sigma_u\colon \begin{cases} \dot{x} = f(x, g_c(x_c)), \\ \dot{x}_c = f_c(x_c, g(x, g_c(x_c)), r), \end{cases} \qquad \text{or} \qquad \Sigma_u\colon \dot{z}_u = f_u(z_u), \tag{4.28}$$

with state $z_u := (x, x_c)$.

We assume that solutions of the nominal system Σ_n and unconstrained system Σ_u exist and are unique for all initial conditions and all (forward and reverse) times. This is a mild assumption as existence and uniqueness of solutions to the nominal and unconstrained systems are usually guaranteed in the design of the nominal controller even when not explicitly sought. For the GPAW-compensated systems Σ_g and Σ_{gu}, we assume that their solutions exist and are unique for all *forward* times when started within the unsaturated region[15] $\mathbb{R}^n \times K$. We restrict its existence to *forward* times since in general, the vector field of the GPAW-compensated system will be *discontinuous* at least on the saturation constraint boundary $\mathbb{R}^n \times \partial K$ (see Section 3.2). For the simple planar LTI system considered in Chapter 3, Proposition 3.3.3 yields the desired result. We leave the proof for existence and uniqueness of solutions to general GPAW-compensated systems Σ_g and Σ_{gu} as future work (see Section 7.1.9). For brevity, this assumption on existence and uniqueness of solutions will not be repeated in the statements of the results in this section.

The preceding systems (4.24), (4.25), (4.27), and (4.28), are regulatory (in contrast to tracking systems) when the controller reference r is constant.[16] We are interested in asymptotic stability, for which it is meaningful only if the constant r induces some *isolated* equilibria.[17] Which of the (possibly multiple) isolated equilibria is to be taken as reference should be apparent for the particular application. We assume that r induces an isolated

[13] We will use the term "unsaturated region" to refer to both $K \subset \mathbb{R}^q$ and $\mathbb{R}^n \times K \subset \mathbb{R}^{n+q}$, which should cause no confusion.

[14] This is discussed in Section 2.5 where (2.36) (equivalent to Σ_{gu} (4.27)) is obtained.

[15] The restriction to initial conditions within the unsaturated region $\mathbb{R}^n \times K$ is for simplicity, since this can usually be enforced (see Remark 2.19).

[16] When r is a constant, it can be taken as part of the description of the function f_c in (4.24), (4.25), (4.27), and (4.28). Then we can write f_n, f_g, f_{gu}, and f_u, in (4.24), (4.25), (4.27), and (4.28) respectively, not as a function of r. These result in autonomous system descriptions.

[17] Observe that not all systems have isolated equilibria. For example, the scalar system $\dot{x} = x^2 + 1$ has no real equilibrium point, while equilibria of the system $\dot{x} = 0$ is the entire line $\mathbb{R}$, and hence not isolated. If any equilibrium point x_{eq} is not isolated, it cannot be asymptotically stable, since the solution can stay identically on an arbitrarily close neighboring equilibrium point and never reach x_{eq}. We stress the need for isolated equilibrium points because the GPAW-compensated system can have a continuum of equilibria even when the nominal system has only isolated equilibria, as seen in Section 3.4.

equilibrium point $z_{eq} := (\tilde{x}, \tilde{x}_c) \in \mathbb{R}^{n+q}$ for the nominal system (4.24) satisfying

$$f(\tilde{x}, g_c(\tilde{x}_c)) = 0, \qquad f_c(\tilde{x}_c, g(\tilde{x}, g_c(\tilde{x}_c)), r) = 0, \qquad u_{\min} < g_c(\tilde{x}_c) < u_{\max}, \tag{4.29}$$

where $u_{\max}$ and $u_{\min}$ are the vectors of saturation limits in (1.2). Notice that the last condition of (4.29) implies that the equilibrium z_{eq} lies in the *interior* of the unsaturated region $\mathbb{R}^n \times K$. It is clear that any isolated equilibrium within the *interior* of the unsaturated region for system (4.24) must also be an isolated equilibrium for systems (4.25), (4.27), and (4.28), since their vector fields coincide in $\mathbb{R}^n \times (K \setminus \partial K)$, i.e.

$$f_n(\bar{z}) = f_g(\bar{z}) = f_{gu}(\bar{z}) = f_u(\bar{z}), \qquad \forall \bar{z} \in \mathbb{R}^n \times (K \setminus \partial K).$$

For this r, we assume that z_{eq} is a locally asymptotically stable equilibrium for the nominal system (4.24), which is *a standard assumption in the anti-windup setting.* The preceding relation and (4.29) implies that it is also a locally asymptotically stable equilibrium for systems (4.25), (4.27), and (4.28). Let the ROAs of z_{eq} for systems (4.24), (4.25), (4.27), and (4.28) be $R_n(z_{eq})$, $R_g(z_{eq})$, $R_{gu}(z_{eq})$, and $R_u(z_{eq})$, respectively. We are interested in establishing ROA containment results like Proposition 3.5.13, where the size of $R_g(z_{eq})$ and $R_{gu}(z_{eq})$ is to be inferred from that of $R_n(z_{eq})$ and $R_u(z_{eq})$ respectively. However, recognizing that the controller state can usually be initialized arbitrarily (see Remark 2.19), and hence it is possible to ensure $x_g(0) \in K$, we are interested only in the size of $R_g(z_{eq})$ and $R_{gu}(z_{eq})$ within the unsaturated region, i.e. $R_g(z_{eq}) \cap (\mathbb{R}^n \times K)$ and $R_{gu}(z_{eq}) \cap (\mathbb{R}^n \times K)$. From the definitions of f_g (4.25) and f_{gu} (4.27), we have

$$f_g(\bar{z}) = f_{gu}(\bar{z}), \qquad \forall \bar{z} \in \mathbb{R}^n \times K,$$

while Theorem 2.5.3 shows that $\mathbb{R}^n \times K$ is a positively invariant set for systems Σ_g and Σ_{gu}. This implies

$$R_g(z_{eq}) \cap (\mathbb{R}^n \times K) = R_{gu}(z_{eq}) \cap (\mathbb{R}^n \times K).$$

Then, with the initial controller state $x_g(0) \in K$ and the initial plant state $x(0)$, the solutions of the GPAW-compensated systems Σ_g and Σ_{gu} will converge to z_{eq} if $(x(0), x_g(0)) \in R_g(z_{eq}) \cap (\mathbb{R}^n \times K)$.

What follows are three ROA comparison results that provide sufficient conditions to assert that some ROA estimate Ω_V of the nominal system Σ_n or unconstrained system Σ_u is also an ROA estimate of the GPAW-compensated system Σ_g or Σ_{gu} respectively. For clarity, we separate the two cases:

(i) the ROA estimate Ω_V for the nominal system Σ_n or unconstrained system Σ_u does not contain the unsaturated region fully, i.e. $(\mathbb{R}^n \times K) \setminus \Omega_V \neq \emptyset$;

(ii) the ROA estimate Ω_V for the nominal system Σ_n or unconstrained system Σ_u contains the unsaturated region, i.e. $(\mathbb{R}^n \times K) \subset \Omega_V$. Clearly, this requires $R_n(z_{eq})$ or $R_u(z_{eq})$ to contain the unsaturated region.

Case (ii) is really a specialization of case (i), and leads to conditions that are simpler and easier to verify, but may be more restrictive. Note that in Theorem 4.4.1 below (corresponding to case (i)), an additional restriction (in contrast to Lemma 4.3.1) has been placed on the description of the ROA estimate Ω_V, namely that it is a sublevel set of the associated Lyapunov function,[18] i.e. $\Omega_V = \{\bar{z} \in \mathbb{R}^{n+q} \mid V(\bar{z}) \leq c\}$ for some $c > 0$. Numerous ROA

[18] Observe that Ω_V having this form implies that it contains the equilibrium z_{eq} in its interior.

estimation methods yield estimates of this form, e.g. [155, 164–166], so that this restriction is well justified. Theorems 4.4.2 and 4.4.3 correspond to case (ii).

Theorem 4.4.1 (Region of Attraction Bounds for GPAW-Compensated System)**.** *Consider the nominal system Σ_n (4.24) and GPAW-compensated system Σ_g (4.25). Assume that $z_{eq} = (\tilde{x}, \tilde{x}_c)$ is an asymptotically stable equilibrium for systems Σ_n and Σ_g that satisfies (4.29), and let $R_n(z_{eq})$, $R_g(z_{eq})$ be the respective ROAs for Σ_n and Σ_g. Let $\Omega_V = \{\bar{z} \in \mathbb{R}^{n+q} \mid V(\bar{z}) \leq c\} \subset R_n(z_{eq})$ for some $c > 0$ be an estimate of $R_n(z_{eq})$ associated with a Lyapunov function $V \colon \mathbb{R}^{n+q} \to \mathbb{R}$, $z = (x, x_c) \mapsto V(x, x_c) = V(z)$, satisfying*

$$\begin{aligned} V(z_{eq}) = 0, \qquad V(\bar{z}) > 0, \qquad\qquad & \forall \bar{z} \in \Omega_V \setminus \{z_{eq}\}, \\ \dot{V}(\bar{z}) = \frac{\partial V(\bar{z})}{\partial z} f_n(\bar{z}) \leq -\alpha(\|\bar{z} - z_{eq}\|), \qquad & \forall \bar{z} \in \Omega_V, \end{aligned} \tag{4.30}$$

for some class $\mathcal{K}$ function α, where f_n is the vector field of the nominal system Σ_n (4.24). If there exists a $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{q \times q}$ such that[19]

$$\frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c} (I - R_{\mathcal{I}^*}) f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r) \geq 0, \qquad \forall (\bar{x}, \bar{x}_c) \in \Omega_V \cap (\mathbb{R}^n \times \partial K), \tag{4.31}$$

then GPAW compensation with parameter Γ yields system Σ_g (4.25) whose ROA contains $\Omega_{VK} := \Omega_V \cap (\mathbb{R}^n \times K)$, i.e. $\Omega_{VK} \subset R_g(z_{eq})$.

Remark 4.6. Observe that (4.31) is independent of the saturation function sat(·), and it needs to hold only on the *boundary* $\mathbb{R}^n \times \partial K$ of the unsaturated region $\mathbb{R}^n \times K$. □

Proof. We will be applying Lemma 4.3.1 with (4.20) and (4.21) representing Σ_n (4.24) and Σ_g (4.25) respectively. First, we establish the analogue of (4.22) with respect to the subset $\Omega_{VK} := (\Omega_V \cap (\mathbb{R}^n \times K)) \subset \Omega_V$. Define

$$f(\bar{x}, \bar{x}_c) := f(\bar{x}, \mathrm{sat}(g_c(\bar{x}_c))), \qquad f_c(\bar{x}, \bar{x}_c) := f_c(\bar{x}_c, g(\bar{x}, \mathrm{sat}(g_c(\bar{x}_c))), r),$$

and observe that $\mathrm{sat}(g_c(\bar{x}_c)) = g_c(\bar{x}_c)$ for all $\bar{x}_c \in K$. Then using (4.24), (4.25), and (4.31), we have

$$\begin{aligned} \frac{\partial V(\bar{z})}{\partial z} f_g(\bar{z}) &= \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x} f(\bar{x}, \bar{x}_c) + \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c} R_{\mathcal{I}^*} f_c(\bar{x}, \bar{x}_c), \\ &\leq \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x} f(\bar{x}, \bar{x}_c) + \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c} f_c(\bar{x}, \bar{x}_c) = \frac{\partial V(\bar{z})}{\partial z} f_n(\bar{z}), \end{aligned}$$

for all $\bar{z} \in \Omega_V \cap (\mathbb{R}^n \times \partial K)$. For all $\bar{z}$ in the interior of the unsaturated region, i.e. $\bar{z} \in \mathbb{R}^n \times (K \setminus \partial K)$, no saturation constraints are active, so that $R_{\mathcal{I}^*} = I$ (see (4.2) and Remark 2.17) and hence $f_g(\bar{z}) = f_n(\bar{z})$. Together, these yield

$$\frac{\partial V(\bar{z})}{\partial z} f_g(\bar{z}) \leq \frac{\partial V(\bar{z})}{\partial z} f_n(\bar{z}), \qquad \forall \bar{z} \in \Omega_{VK}, \tag{4.32}$$

which is analogous to (4.22).

Next, we show that Ω_{VK} is a positively invariant set for Σ_g. Since $\Omega_V = \{\bar{z} \in \mathbb{R}^{n+q} \mid V(\bar{z}) \leq c\}$, we can express $\Omega_{VK} = \Omega_V \cap (\mathbb{R}^n \times K)$ as $\Omega_{VK} = \{\bar{z} \in \mathbb{R}^n \times K \mid V(\bar{z}) \leq c\}$.

[19] Recall that $R_{\mathcal{I}^*}$ is defined with parameter Γ (see (4.2)).

Theorem 2.5.3 shows $\mathbb{R}^n \times K$ to be a positively invariant set for Σ_g, so that for any $z_0 \in \Omega_{VK}$, the solution $\phi_g(t, z_0)$ of Σ_g satisfy $\phi_g(t, z_0) \in \mathbb{R}^n \times K$ for all $t \geq 0$. Conditions (4.32) and (4.30) show that $\frac{\partial V(\bar{z})}{\partial z} f_g(\bar{z}) \leq 0$ for all $\bar{z} \in \Omega_{VK}$. Together, these show Ω_{VK} to be a positively invariant set for Σ_g.

By making the following identifications with quantities in Lemma 4.3.1,

$$\begin{aligned} &f_1 \sim f_n, \qquad R_{A1}(x_{eq}) \sim R_n(z_{eq}), \qquad x_{eq} \sim z_{eq}, \\ &f_2 \sim f_g, \qquad R_{A2}(x_{eq}) \sim R_g(z_{eq}), \qquad \Omega_2 \sim \Omega_{VK}, \end{aligned}$$

it can be verified that all its hypotheses are satisfied. Application of Lemma 4.3.1 then yields $\Omega_{VK} \subset R_g(z_{eq})$, as desired. ■

Remark 4.7. From [37, Lemma 4.3, p. 145], we see that a class $\mathcal{K}$ function α exists to satisfy the second condition of (4.30) when $\dot{V}$ is continuous and negative definite with respect to the equilibrium z_{eq} in Ω_V, i.e. when $\dot{V}$ is continuous and

$$\dot{V}(z_{eq}) = 0, \qquad \dot{V}(\bar{z}) < 0, \qquad \forall \bar{z} \in \Omega_V \setminus \{z_{eq}\}.$$

Clearly, $\dot{V}$ is continuous when V is continuously differentiable and f_n is continuous. □

Remark 4.8. The main condition in Theorem 4.4.1 is (4.31), which as mentioned, is independent of the saturation function. It can be verified that the proof is valid with minor modifications[20] when Σ_n, Σ_g, f_n, f_g, $R_n(z_{eq})$, $R_g(z_{eq})$, are replaced by Σ_u, Σ_{gu}, f_u, f_{gu}, $R_u(z_{eq})$, $R_{gu}(z_{eq})$, respectively. For brevity, we will not re-state the result for this case. □

The observation of Remark 4.8 means that the ROA of the GPAW-compensated system can be lower bounded in size by two ways: comparing against an ROA estimate of the nominal (saturated) system, or an ROA estimate of the *unconstrained* system. When comparing against an ROA estimate of the nominal system, condition (4.31) is a sufficient condition for GPAW compensation to yield ROA improvements over the nominal system with respect to the estimate. Usually, the ROA of the unconstrained system will be at least as large as the ROA of the constrained nominal system. There may exist a Γ satisfying (4.31), yet no Γ such that the analogue of (4.31) can hold with respect to a larger ROA estimate for the unconstrained system. ROA comparison against the nominal system is a relative result, while comparisons against the ROA of the unconstrained system is in some sense an "absolute" result, prevalent in current anti-windup literature. As shown in Section 3.7, such "absolute" results may give some confidence in the application of the anti-windup scheme, but may not reveal any advantages gained by its adoption.

Remark 4.9. Theorem 4.4.1 gives sufficient conditions for GPAW compensation to yield ROA improvements, without any indication as to the existence of Γ satisfying (4.31), nor its choice. Moreover, the definition of $R_{\mathcal{I}^*}$ requires the solution of an optimization problem in general (see Section 4.1), on every point in $\Omega_V \cap (\mathbb{R}^n \times \partial K)$. In general, it would not be easy to find a Γ satisfying (4.31). However, it will be shown in Section 4.5.2 that Theorem 4.4.1 can be applied at least to a simple nonlinear system. In view of these, we recommend to start with Γ being the identity matrix when attempting to verify (4.31). We also note that (4.31) may be more readily verified if the controller admits closed-form expressions (e.g. see Appendices A and B). Using the notions of polar and dual cones as in

[20] The only modifications needed are that the condition $\text{sat}(g_c(\bar{x}_c)) = g_c(\bar{x}_c)$ for all $\bar{x}_c \in K$ is *not needed*, and define $f(\bar{x}, \bar{x}_c) := f(\bar{x}, g_c(\bar{x}_c))$, $f_c(\bar{x}, \bar{x}_c) := f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r)$ without the saturation function.

Section 4.1 is also likely to be more tractable. Indeed, using (4.5), condition (4.31) can be written as (compare with (4.16))

$$\frac{\partial V(\bar{x},\bar{x}_c)}{\partial x_c}\Phi(I-P_{\mathcal{I}^*})\Phi^{-1}f_c(\bar{x},\bar{x}_c) \geq 0, \qquad \forall(\bar{x},\bar{x}_c) \in \Omega_V \cap (\mathbb{R}^n \times \partial K),$$

where $\Gamma = \Phi\Phi^{\mathrm{T}}$ [124, Theorem 7.2.7, p. 406] and $f_c(\bar{x},\bar{x}_c) := f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r)$. As seen from the proof of Theorem 4.2.4, the preceding condition holds if the vector $\left(\frac{\partial V(\bar{x},\bar{x}_c)}{\partial x_c}\Phi\right)^{\mathrm{T}}$ lies in the dual to the polyhedral cone generated by the (transformed) gradients of the active saturation constraints, restricted to points on $\Omega_V \cap (\mathbb{R}^n \times \partial K)$. □

The next result specializes Theorem 4.4.1 to the case where the ROA $R_n(z_{eq})$ (or ROA estimate Ω_V) of the nominal system contains the unsaturated region $\mathbb{R}^n \times K$. This includes the case when global asymptotic stability of the equilibrium for the nominal system can be assured.[21] Since only the unsaturated region is of concern, we call the conclusion "global asymptotic stability" (as in [103, 104]) to mean that the ROA contains the unsaturated region $\mathbb{R}^n \times K$. One minor difference with Theorem 4.4.1 is that the ROA estimate is not required to be a sublevel set of the associated Lyapunov function. The proof is largely similar to that for Theorem 4.4.1.

Theorem 4.4.2 (Global Asymptotic Stability for GPAW-Compensated System)**.** *Consider the nominal system Σ_n (4.24) and GPAW-compensated system Σ_g (4.25). Assume that $z_{eq} = (\tilde{x}, \tilde{x}_c)$ is an asymptotically stable equilibrium for systems Σ_n and Σ_g that satisfies (4.29), and let $R_n(z_{eq})$, $R_g(z_{eq})$ be the respective ROAs for Σ_n and Σ_g. Assume that $V\colon \mathbb{R}^{n+q} \to \mathbb{R}$, $z = (x, x_c) \mapsto V(x, x_c) = V(z)$, is a Lyapunov function that satisfies*

$$\begin{aligned} V(z_{eq}) = 0, \qquad V(\bar{z}) > 0, \qquad\qquad & \forall \bar{z} \in (\mathbb{R}^n \times K) \setminus \{z_{eq}\}, \\ \dot{V}(\bar{z}) = \frac{\partial V(\bar{z})}{\partial z} f_n(\bar{z}) \leq -\alpha(\|\bar{z} - z_{eq}\|), \qquad & \forall \bar{z} \in \mathbb{R}^n \times K, \end{aligned} \tag{4.33}$$

for some class $\mathcal{K}$ function α, and that $R_n(z_{eq})$ contains the unsaturated region, i.e. $(\mathbb{R}^n \times K) \subset R_n(z_{eq})$. If there exists a $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{q\times q}$ such that

$$\frac{\partial V(\bar{x},\bar{x}_c)}{\partial x_c}(I - R_{\mathcal{I}^*}) f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r) \geq 0, \qquad \forall(\bar{x},\bar{x}_c) \in \mathbb{R}^n \times \partial K, \tag{4.34}$$

then GPAW compensation with parameter Γ yields system Σ_g (4.25) whose ROA contains the unsaturated region, i.e. $(\mathbb{R}^n \times K) \subset R_g(z_{eq})$.

Proof. We will be applying Lemma 4.3.1 with (4.20) and (4.21) representing Σ_n (4.24) and Σ_g (4.25) respectively. As in the proof of Theorem 4.4.1, the analogue of (4.22) with respect to the unsaturated region $\mathbb{R}^n \times K$ can be readily verified. Define $f(\bar{x},\bar{x}_c) := f(\bar{x}, \mathrm{sat}(g_c(\bar{x}_c)))$, $f_c(\bar{x},\bar{x}_c) := f_c(\bar{x}_c, g(\bar{x}, \mathrm{sat}(g_c(\bar{x}_c))), r)$, and observe that $\mathrm{sat}(g_c(\bar{x}_c)) = g_c(\bar{x}_c)$ for all $\bar{x}_c \in K$. Then using (4.24), (4.25), and (4.34), we have

$$\begin{aligned} \frac{\partial V(\bar{z})}{\partial z} f_g(\bar{z}) &= \frac{\partial V(\bar{x},\bar{x}_c)}{\partial x} f(\bar{x},\bar{x}_c) + \frac{\partial V(\bar{x},\bar{x}_c)}{\partial x_c} R_{\mathcal{I}^*} f_c(\bar{x},\bar{x}_c), \\ &\leq \frac{\partial V(\bar{x},\bar{x}_c)}{\partial x} f(\bar{x},\bar{x}_c) + \frac{\partial V(\bar{x},\bar{x}_c)}{\partial x_c} f_c(\bar{x},\bar{x}_c) = \frac{\partial V(\bar{z})}{\partial z} f_n(\bar{z}), \end{aligned}$$

[21] Recall that global asymptotic stability can be achieved for the closed-loop system only when the (unsaturated) open-loop plant is not unstable [34].

for all $\bar{z} \in \mathbb{R}^n \times \partial K$. For all $\bar{z}$ in the interior of the unsaturated region, i.e. $\bar{z} \in \mathbb{R}^n \times (K \backslash \partial K)$, no saturation constraints are active, so that $R_{\mathcal{I}^*} = I$ (see (4.2) and Remark 2.17) and hence $f_g(\bar{z}) = f_n(\bar{z})$. Together, these yield

$$\frac{\partial V(\bar{z})}{\partial z} f_g(\bar{z}) \le \frac{\partial V(\bar{z})}{\partial z} f_n(\bar{z}), \qquad \forall \bar{z} \in \mathbb{R}^n \times K,$$

which is analogous to (4.22).

Theorem 2.5.3 shows the unsaturated region $\mathbb{R}^n \times K$ to be a positively invariant set for Σ_g. By making the following identifications with quantities in Lemma 4.3.1,

$$\begin{aligned} &f_1 \sim f_n, \qquad && R_{A1}(x_{eq}) \sim R_n(z_{eq}), \qquad && x_{eq} \sim z_{eq}, \\ &f_2 \sim f_g, \qquad && R_{A2}(x_{eq}) \sim R_g(z_{eq}), \qquad && \Omega_2 \sim \mathbb{R}^n \times K, \end{aligned}$$

it can be verified that all its hypotheses are satisfied. Application of Lemma 4.3.1 then yields $(\mathbb{R}^n \times K) \subset R_g(z_{eq})$, as desired. ■

Remarks 4.7, 4.8, and 4.9 apply to Theorem 4.4.2 with (4.34) in place of (4.31). However, we note that in this case, comparison against ROAs (or ROA estimates) of the nominal or unconstrained system result in the same conclusion irrespective of any difference in size of the ROAs (or ROA estimates).

The main drawback of Theorems 4.4.1 and 4.4.2 is that conditions (4.31) and (4.34) are hard to verify in general (see Remark 4.9). For the next result, we impose further restrictions on the nominal system and use the geometric bounding condition of Theorem 4.2.4 to yield a result with a *prescribed* choice of the GPAW parameter Γ. In particular, we assume that the unsaturated region $\mathbb{R}^n \times K$ (or equivalently, K, due to Corollary 4.2.2) is a star domain (see Definition 4.2), the equilibrium point $z_{eq} = (\tilde{x}, \tilde{x}_c)$ lies in its kernel (see Definition 4.1), and the Lyapunov function is "decoupled" in the sense that it takes the form $V(x, x_c) = V_x(x) + (x_c - \tilde{x}_c)^{\mathrm{T}} P_c (x_c - \tilde{x}_c)$ for some symmetric positive definite matrix $P_c = P_c^{\mathrm{T}} > 0 \in \mathbb{R}^{q \times q}$. While not proven, it is conjectured that only systems where the open-loop plant is not unstable admits Lyapunov functions of this structure.

Theorem 4.4.3 (Global Asymptotic Stability for GPAW-Compensated System with Star Unsaturated Region)**.** *Consider the nominal system Σ_n* (4.24) *and GPAW-compensated system Σ_g* (4.25)*. Assume that $z_{eq} = (\tilde{x}, \tilde{x}_c)$ is an asymptotically stable equilibrium for systems Σ_n and Σ_g that satisfies* (4.29)*, and let $R_n(z_{eq})$, $R_g(z_{eq})$ be the respective ROAs for Σ_n and Σ_g. Assume the unsaturated region $\mathbb{R}^n \times K$* (4.26) *is a star domain and z_{eq} lies in its kernel, i.e. $z_{eq} \in \mathbb{R}^n \times \ker(K)$ or $\tilde{x}_c \in \ker(K)$. Assume that $V \colon \mathbb{R}^{n+q} \to \mathbb{R}$, $z = (x, x_c) \mapsto V(x, x_c) = V(z)$, is a Lyapunov function of the form*

$$V(x, x_c) = V_x(x) + (x_c - \tilde{x}_c)^{\mathrm{T}} P_c (x_c - \tilde{x}_c), \qquad P_c = P_c^{\mathrm{T}} > 0, \tag{4.35}$$

that satisfies (4.33)*, and that $R_n(z_{eq})$ contains the unsaturated region, i.e. $(\mathbb{R}^n \times K) \subset R_n(z_{eq})$. Then GPAW compensation with parameter $\Gamma = P_c^{-1}$ yields system Σ_g* (4.25) *whose ROA contains the unsaturated region, i.e. $(\mathbb{R}^n \times K) \subset R_g(z_{eq})$.*

Proof. We will be applying Theorem 4.4.2 to yield the desired conclusion. Observe that except for condition (4.34), all hypotheses of Theorem 4.4.2 are satisfied. It is sufficient to show that with $\Gamma = P_c^{-1}$, condition (4.34) holds with the additional assumptions, namely (i) K is a star domain; (ii) $\tilde{x}_c \in \ker(K)$; and (iii) V has the form (4.35).

Since K is a star domain and $\tilde{x}_c \in \ker(K)$, Theorem 4.2.4 yields the geometric bounding condition

$$\langle P_c(\bar{x}_c - \tilde{x}_c), R_{\mathcal{I}^*} f_c(\bar{x}, \bar{x}_c)\rangle \leq \langle P_c(\bar{x}_c - \tilde{x}_c), f_c(\bar{x}, \bar{x}_c)\rangle, \qquad \forall (\bar{x}, \bar{x}_c) \in \mathbb{R}^n \times K,$$

where $f_c(\bar{x}, \bar{x}_c) := f_c(\bar{x}_c, g(\bar{x}, \mathrm{sat}(g_c(\bar{x}_c))), r)$, and we have used $\Gamma = P_c^{-1}$. The preceding can be written as

$$\begin{aligned}\langle P_c(\bar{x}_c - \tilde{x}_c), (I - R_{\mathcal{I}^*}) f_c(\bar{x}, \bar{x}_c)\rangle &= (\bar{x}_c - \tilde{x}_c)^{\mathrm{T}} P_c (I - R_{\mathcal{I}^*}) f_c(\bar{x}, \bar{x}_c), \\ &= \frac{1}{2} \cdot \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c} (I - R_{\mathcal{I}^*}) f_c(\bar{x}, \bar{x}_c) \geq 0,\end{aligned}$$

for all $(\bar{x}, \bar{x}_c) \in \mathbb{R}^n \times K$. The last condition implies (4.34) with $\Gamma = P_c^{-1}$. Application of Theorem 4.4.2 yields the conclusion. ■

Remark 4.10. As in Remark 4.8, Theorem 4.4.3 is valid when Σ_n, Σ_g, $R_n(z_{eq})$, $R_g(z_{eq})$, and f_n, are replaced by Σ_u, Σ_{gu}, $R_u(z_{eq})$, $R_{gu}(z_{eq})$, and f_u, respectively. □

We note that the star domain assumption on the unsaturated region holds in particular when K is convex, which is the case when g_c in (4.1) is linear in its argument (see also Remark 4.4). If K is convex, then condition (4.29) implies $\tilde{x} \in \ker(K)$. The most restrictive assumption in Theorem 4.4.3 is the requirement for the Lyapunov function to have the form of (4.35). As will be shown in the next section, this is satisfied for some simple systems with open-loop stable plants.

4.5 Applications of Region of Attraction Comparison Results

In this section, we demonstrate how to apply Theorems 4.4.1, 4.4.2, and 4.4.3. Since the planar LTI system studied in Chapter 3 is simple and well understood, we use this system to demonstrate the applications of Theorems 4.4.2 and 4.4.3 in Section 4.5.1. In Section 4.5.2, we apply Theorem 4.4.1 to a simple planar nonlinear system. A further simplification is due to the controllers being first order, hence is independent of any GPAW parameters (see Remark B.1 in Appendix B).

4.5.1 Input Constrained Planar LTI Systems

Recall the input constrained planar LTI nominal system (3.4)

$$\Sigma_n\colon \begin{cases} \dot{x} = ax + b\,\mathrm{sat}(u), \\ \dot{u} = cx + du, \end{cases} \qquad \text{or} \qquad \Sigma_n\colon \dot{z} = f_n(z), \tag{4.36}$$

and GPAW-compensated system (3.5)

$$\Sigma_g\colon \begin{cases} \dot{x} = ax + b\,\mathrm{sat}(u), \\ \dot{u} = \begin{cases} 0, & \text{if } u \geq u_{\max} \wedge cx + du > 0, \\ 0, & \text{if } u \leq u_{\min} \wedge cx + du < 0, \\ cx + du, & \text{otherwise}, \end{cases} \end{cases} \qquad \text{or} \qquad \Sigma_g\colon \dot{z} = f_g(z). \tag{4.37}$$

Corresponding to these systems are the unconstrained system

$$\Sigma_u\colon \begin{cases} \dot{x} = ax + bu, \\ \dot{u} = cx + du, \end{cases} \qquad \text{or} \qquad \Sigma_u\colon \dot{z} = f_u(z) = Az, \qquad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \tag{4.38}$$

and alternate form of the GPAW-compensated system[22]

$$\Sigma_{gu}\colon \begin{cases} \dot{x} = ax + bu, \\ \dot{u} = \begin{cases} 0, & \text{if } u \geq u_{\max} \wedge cx + du > 0, \\ 0, & \text{if } u \leq u_{\min} \wedge cx + du < 0, \\ cx + du, & \text{otherwise}, \end{cases} \end{cases} \qquad \text{or} \qquad \Sigma_{gu}\colon \dot{z} = f_{gu}(z), \tag{4.39}$$

where the saturation function has been eliminated. The objective is to regulate the system state about the origin $z_{eq} := (0, 0)$, which can be verified to satisfy (4.29). Note that given any initial condition (x_0, u_0), initializing the system state with $(x(0), u(0)) = (x_0, \mathrm{sat}(u_0))$ ensures that the system solutions start within the unsaturated region $\mathbb{R} \times [u_{\min}, u_{\max}]$.

For definiteness, let the plant and controller parameters be

$$(a, b, c, d, u_{\max}, u_{\min}) = (-1, 1, -1, -1, 1, -1),$$

which can be verified to satisfy Assumption 3.1. Satisfaction of Assumption 3.1 implies that $A = \left[\begin{smallmatrix} -1 & 1 \\ -1 & -1 \end{smallmatrix}\right]$ is Hurwitz, the origins of Σ_n, Σ_g, Σ_{gu} are locally exponentially stable, and the origin of Σ_u is globally exponentially stable. Observe that the unsaturated open-loop plant is strictly stable since $a = -1 < 0$. It can be verified that the Lyapunov equation

$$PA + A^{\mathrm{T}}P = -Q, \qquad P = \begin{bmatrix} p_1 & 0 \\ 0 & p_2 \end{bmatrix} = I = P^{\mathrm{T}} > 0, \qquad Q = 2I = Q^{\mathrm{T}} \geq 0,$$

holds. Defining $V(\bar{z}) := \bar{z}^{\mathrm{T}} P \bar{z}$, we have

$$\begin{gathered} V(z_{eq}) = 0, \qquad V(\bar{z}) > 0, \qquad \forall \bar{z} \neq z_{eq}, \\ \dot{V}_u(\bar{z}) := \frac{\partial V(\bar{z})}{\partial z} f_u(\bar{z}) = \bar{z}^{\mathrm{T}}(PA + A^{\mathrm{T}}P)\bar{z} = -\bar{z}^{\mathrm{T}} Q \bar{z} = -2\|\bar{z}\|^2, \qquad \forall \bar{z} \in \mathbb{R}^2, \end{gathered} \tag{4.40}$$

so that $V(\bar{z})$ is a Lyapunov function for system Σ_u. Clearly, V satisfies (4.33) with the class $\mathcal{K}$ function $\alpha(r) = 2r^2$. Since the vector fields f_n (4.36), f_g (4.37), f_u (4.38), f_{gu} (4.39), coincide in the interior of the unsaturated region, i.e.

$$f_n(\bar{z}) = f_g(\bar{z}) = f_{gu}(\bar{z}) = f_u(\bar{z}), \qquad \forall \bar{z} \in \mathbb{R} \times (u_{\min}, u_{\max}),$$

we see that V is also a Lyapunov function for systems Σ_n, Σ_g, and Σ_{gu}.

Application of Theorem 4.4.2

To apply Theorem 4.4.2, we need to verify (4.33), that the ROA $R_u(z_{eq})$ contains the unsaturated region $\mathbb{R} \times [u_{\min}, u_{\max}]$, and condition (4.34) holds. Clearly, (4.33) holds with f_u in place of f_n by virtue of (4.40). Since A is Hurwitz, it follows from [37, Theorem 4.5,

[22] The alternate form of the GPAW-compensated system (4.39) is valid when the controller state is initialized to satisfy $u_{\min} \leq u(0) \leq u_{\max}$.

p. 134] that the origin of Σ_u is globally asymptotically stable, so that the ROA contains the unsaturated region, i.e. $\mathbb{R} \times [u_{\min}, u_{\max}] \subset R_u(z_{eq})$.

It remains to verify (4.34). Using the definitions of f_{gu} (4.39) and f_u (4.38), for all $(\bar{x}, \bar{x}_c) \in \mathbb{R}^2$, we have

$$\begin{aligned} \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c}(I - R_{\mathcal{I}^*})f_c(\bar{x}, \bar{x}_c) &= 2p_2\bar{x}_c(I - R_{\mathcal{I}^*})f_c(\bar{x}, \bar{x}_c), \\ &= \begin{cases} 2p_2\bar{x}_c(c\bar{x} + d\bar{x}_c), & \text{if } \bar{x}_c \geq u_{\max} \wedge c\bar{x} + d\bar{x}_c > 0, \\ 2p_2\bar{x}_c(c\bar{x} + d\bar{x}_c), & \text{if } \bar{x}_c \leq u_{\min} \wedge c\bar{x} + d\bar{x}_c < 0, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{4.41}$$

where $f_c(\bar{x}, \bar{x}_c) := f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r) = c\bar{x} + d\bar{x}_c$. To verify (4.34), we need to show that $\frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c}(I - R_{\mathcal{I}^*})f_c(\bar{x}, \bar{x}_c) \geq 0$ for all $(\bar{x}, \bar{x}_c) \in \mathbb{R} \times \partial K = \mathbb{R} \times \{u_{\min}, u_{\max}\}$. For any $(\bar{x}, \bar{x}_c) \in \mathbb{R} \times \{u_{\min}, u_{\max}\}$, we have from (4.41) that

$$\frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c}(I - R_{\mathcal{I}^*})f_c(\bar{x}, \bar{x}_c) = \begin{cases} 2p_2u_{\max}(c\bar{x} + du_{\max}), & \text{if } \bar{x}_c = u_{\max} \wedge c\bar{x} + du_{\max} > 0, \\ 2p_2u_{\min}(c\bar{x} + du_{\min}), & \text{if } \bar{x}_c = u_{\min} \wedge c\bar{x} + du_{\min} < 0, \\ 0, & \text{otherwise.} \end{cases}$$

Now, consider the first condition and function value in the preceding. If the first condition ($\bar{x}_c = u_{\max} \wedge c\bar{x} + du_{\max} > 0$) is satisfied, then we have $c\bar{x} + du_{\max} > 0$ and the first function value satisfies $2p_2u_{\max}(c\bar{x} + du_{\max}) > 0$ since $p_2u_{\max} > 0$. In similar manner, we can conclude that if the second condition is satisfied, then the second function value satisfies $2p_2u_{\min}(c\bar{x} + du_{\min}) > 0$ because $p_2u_{\min} < 0$. Collecting these observations establishes $\frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c}(I - R_{\mathcal{I}^*})f_c(\bar{x}, \bar{x}_c) \geq 0$ for all $(\bar{x}, \bar{x}_c) \in \mathbb{R} \times \partial K$, which shows that (4.34) holds.

All hypotheses of Theorem 4.4.2 are satisfied, and we conclude from its application that the ROA of system Σ_{gu} contains the unsaturated region $\mathbb{R} \times [u_{\min}, u_{\max}]$. Since the vector fields of systems Σ_g (4.37) and Σ_{gu} (4.39) coincide in the unsaturated region, i.e.

$$f_g(\bar{z}) = f_{gu}(\bar{z}), \qquad \forall \bar{z} \in \mathbb{R} \times [u_{\min}, u_{\max}],$$

and Theorem 2.5.3 shows the unsaturated region to be positively invariant, we have that their ROAs coincide in this region, i.e. $R_g(z_{eq}) \cap (\mathbb{R} \times [u_{\min}, u_{\max}]) = R_{gu}(z_{eq}) \cap (\mathbb{R} \times [u_{\min}, u_{\max}])$. This means that the ROA of the GPAW-compensated system Σ_g (4.37) also contains the unsaturated region.

Application of Theorem 4.4.3

Next, we show how to apply Theorem 4.4.3. While comparison against the ROA of the unconstrained system Σ_u (as in the preceding) may be simpler, we will compare against the ROA of the nominal system Σ_n for illustration. To apply Theorem 4.4.3, we need to verify that the unsaturated region $\mathbb{R} \times [u_{\min}, u_{\max}]$ is a star domain, z_{eq} lies in its kernel, V has the form of (4.35) and satisfies (4.33), and that the ROA $R_n(z_{eq})$ contains the unsaturated region.

Since the interval $[u_{\min}, u_{\max}]$ is convex, it is a star domain. By Corollary 4.2.2, the unsaturated region $\mathbb{R} \times [u_{\min}, u_{\max}]$ must also be a star domain. It is clear that $z_{eq} = (0, 0)$ lies in $\mathbb{R} \times [u_{\min}, u_{\max}]$, which is the kernel of itself, i.e. $\ker(\mathbb{R} \times [u_{\min}, u_{\max}]) = \mathbb{R} \times [u_{\min}, u_{\max}]$. Moreover, it can be seen that $V(x, x_c) = p_1x^2 + p_2x_c^2$ is of the form (4.35) with $P_c = p_2 > 0$

and $V_x(x) = p_1 x^2$. To show that V satisfies (4.33), observe that the vector fields f_n (4.36) and f_u (4.38) coincide in the unsaturated region, i.e.

$$f_n(\bar{z}) = f_u(\bar{z}), \qquad \forall \bar{z} \in \mathbb{R} \times [u_{\min}, u_{\max}].$$

Then (4.33) follows from (4.40), restricted to points in $\mathbb{R} \times [u_{\min}, u_{\max}]$. It remains to show that $R_n(z_{eq})$ contains the unsaturated region. Since both the open-loop plant and the nominal controller are strictly stable, Claim 3.7.1 shows that the origin of system Σ_n is globally asymptotically stable, so that $R_n(z_{eq})$ must contain the unsaturated region.

All hypotheses of Theorem 4.4.3 are satisfied, and we conclude from its application that the ROA of system Σ_g contains the unsaturated region $\mathbb{R} \times [u_{\min}, u_{\max}]$.

4.5.2 A Nonlinear Example for Application of Theorem 4.4.1

The following system is adapted from [37, Example 8.9, pp. 318 – 320]

$$\Sigma_n\colon \begin{cases} \dot{x} = -\operatorname{sat}(u), \\ \dot{u} = x + (x^2 - 1)u, \end{cases} \qquad \operatorname{sat}(u) = \max\{\min\{u, u_{\max}\}, u_{\min}\},$$

which corresponds to the constrained nominal system (4.24). The objective is to regulate the system state about the origin $z_{eq} = (0, 0)$. We take $u_{\max} = 1$ and $u_{\min} = -1$. Using the closed-form expressions (A.5) (in Appendix A) for the GPAW-compensated controller, the GPAW-compensated system can be written as[23]

$$\Sigma_g\colon \begin{cases} \dot{x} = -\operatorname{sat}(u), \\ \dot{u} = \begin{cases} 0, & \text{if } u \geq u_{\max} \wedge x + (x^2 - 1)u > 0, \\ 0, & \text{if } u \leq u_{\min} \wedge x + (x^2 - 1)u < 0, \\ x + (x^2 - 1)u, & \text{otherwise.} \end{cases} \end{cases}$$

Since an ROA estimate for the associated unconstrained system (4.28)

$$\Sigma_u\colon \begin{cases} \dot{x} = -u, \\ \dot{u} = x + (x^2 - 1)u, \end{cases} \qquad \text{or} \qquad \Sigma_u\colon \dot{z} = f_u(z), \tag{4.42}$$

is readily available in [37, Example 8.9, pp. 318 – 320], we will use it to demonstrate the application of Theorem 4.4.1 (see Remark 4.8). The conclusion yields a containment result for the ROA of the alternate form of the GPAW-compensated system

$$\Sigma_{gu}\colon \begin{cases} \dot{x} = -u, \\ \dot{u} = \begin{cases} 0, & \text{if } u \geq u_{\max} \wedge x + (x^2 - 1)u > 0, \\ 0, & \text{if } u \leq u_{\min} \wedge x + (x^2 - 1)u < 0, \\ x + (x^2 - 1)u, & \text{otherwise.} \end{cases} \end{cases} \tag{4.43}$$

Remark 4.11. In Chapter 6, we will be comparing the GPAW scheme against three state-of-the-art anti-windup schemes proposed in [24, 65, 128]. We note that none of these methods can be applied to this simple system because the nominal controller (described by

[23] We have used $g_c(u) = u$, so that $\nabla g_c(u) = 1$ in (A.5).

$\dot{u} = x + (x^2 - 1)u$) is not a feedback linearizing controller, the unconstrained system is not globally asymptotically stable, and the nominal controller is nonlinear. $\square$

As shown in [37, Example 8.9, pp. 318 – 320], a Lyapunov function for system Σ_u is

$$V(\bar{z}) = \bar{z}^{\mathrm{T}} P \bar{z}, \qquad P = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 1 \end{bmatrix},$$

and an ROA estimate associated with V is the sublevel set $\Omega_V = \{\bar{z} \in \mathbb{R}^2 \mid V(\bar{z}) \leq c\}$ with $c = 2.25$. This choice of $c = 2.25$ ensures $\dot{V}(\bar{z}) = \frac{\partial V(\bar{z})}{\partial z} f_u(\bar{z})$ is negative definite in Ω_V. Since V is continuously differentiable and f_u is continuous, (4.30) holds (see Remark 4.7). The ROA estimate Ω_V is illustrated in Fig. 4-4 together with $R_u(z_{eq})$ and $R_{gu}(z_{eq})$, the true ROAs of the unconstrained system Σ_u and GPAW-compensated system Σ_{gu} respectively.[24]

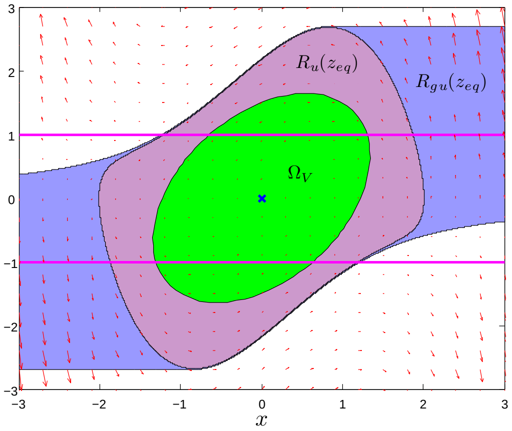


Figure 4-4: ROA estimate Ω_V of a planar nonlinear system together with $R_u(z_{eq})$ and $R_{gu}(z_{eq})$, the true ROAs of the unconstrained system Σ_u and GPAW-compensated system Σ_{gu} respectively.

While not proven, Fig. 4-4 suggests that GPAW compensation may enlarge the ROA of the *unconstrained* system.

To apply Theorem 4.4.1, we need to verify (4.31). Note that because the controller is first order, it is independent of any GPAW parameter Γ (see Remark B.1 in Appendix B). From (4.42) and (4.43), we see that $f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r)$ in (4.31) translates to $\bar{x} + (\bar{x}^2 - 1)\bar{x}_c$. Defining $\gamma(\bar{x}, \bar{x}_c) := \frac{\partial V(\bar{x}, \bar{x}_c)}{\partial x_c}(I - R_{\mathcal{I}^*}) f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r)$, we have that for all $(\bar{x}, \bar{x}_c) \in \mathbb{R}^2$,

$$\begin{aligned} \gamma(\bar{x}, \bar{x}_c) &= (2\bar{x}_c - \bar{x})(I - R_{\mathcal{I}^*}) f_c(\bar{x}_c, g(\bar{x}, g_c(\bar{x}_c)), r), \\ &= \begin{cases} (2\bar{x}_c - \bar{x})(\bar{x} + (\bar{x}^2 - 1)\bar{x}_c), & \text{if } \bar{x}_c \geq u_{\max} \wedge \bar{x} + (\bar{x}^2 - 1)\bar{x}_c > 0, \\ (2\bar{x}_c - \bar{x})(\bar{x} + (\bar{x}^2 - 1)\bar{x}_c), & \text{if } \bar{x}_c \leq u_{\min} \wedge \bar{x} + (\bar{x}^2 - 1)\bar{x}_c < 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

[24] The ROAs $R_u(z_{eq})$ and $R_{gu}(z_{eq})$ are found numerically by a trial and error process together with backward-in-time simulations.

For any $(\bar{x}, \bar{x}_c) \in \mathbb{R} \times \partial K = \mathbb{R} \times \{u_{\min}, u_{\max}\}$, we have from the preceding,

$$\gamma(\bar{x}, \bar{x}_c) = \begin{cases} (2u_{\max} - \bar{x})(\bar{x} + (\bar{x}^2 - 1)u_{\max}), & \text{if } \bar{x}_c = u_{\max} \wedge \bar{x} + (\bar{x}^2 - 1)u_{\max} > 0, \\ (2u_{\min} - \bar{x})(\bar{x} + (\bar{x}^2 - 1)u_{\min}), & \text{if } \bar{x}_c = u_{\min} \wedge \bar{x} + (\bar{x}^2 - 1)u_{\min} < 0, \\ 0, & \text{otherwise.} \end{cases}$$

By inspection of the preceding, we see that $\gamma(\bar{x}, \bar{x}_c) \geq 0$ for all $(\bar{x}, \bar{x}_c) = (\bar{x}, u_{\max})$ when $\bar{x} \leq 2u_{\max} = 2$, and all $(\bar{x}, \bar{x}_c) = (\bar{x}, u_{\min})$ when $\bar{x} \geq 2u_{\min} = -2$. In other words, we have

$$\gamma(\bar{x}, \bar{x}_c) \geq 0, \qquad \forall (\bar{x}, \bar{x}_c) \in ((-\infty, 2] \times \{u_{\max}\}) \cup ([-2, \infty) \times \{u_{\min}\}) \subset \mathbb{R} \times \partial K. \quad (4.44)$$

Condition (4.31) requires $\gamma(\bar{x}, \bar{x}_c) \geq 0$ for all $(\bar{x}, \bar{x}_c) \in \Omega_V \cap (\mathbb{R} \times \partial K)$. Using the definition of Ω_V ($\Omega_V = \{\bar{z} \in \mathbb{R}^2 \mid V(\bar{z}) \leq c\}$ for $c = 2.25$), it can be verified that

$$\Omega_V \cap (\mathbb{R} \times \partial K) = (\beta_+ \times \{u_{\max}\}) \cup (\beta_- \times \{u_{\min}\}),$$

where $\beta_+ := \left[\frac{1}{3} - \sqrt{\frac{17}{18}}, \frac{1}{3} + \sqrt{\frac{17}{18}}\right] = [-0.638, 1.305]$ and $\beta_- := \left[-\frac{1}{3} - \sqrt{\frac{17}{18}}, -\frac{1}{3} + \sqrt{\frac{17}{18}}\right] = [-1.305, 0.638]$. Since $\beta_+ \subset (-\infty, 2]$ and $\beta_- \subset [-2, \infty)$, we see from (4.44) that $\gamma(\bar{x}, \bar{x}_c) \geq 0$ for all $(\bar{x}, \bar{x}_c) \in (\beta_+ \times \{u_{\max}\}) \cup (\beta_- \times \{u_{\min}\}) = \Omega_V \cap (\mathbb{R} \times \partial K)$, which shows that (4.31) holds.

All hypotheses of Theorem 4.4.1 are satisfied, and we conclude from its application that the ROA of system Σ_{gu} contains $\Omega_V \cap (\mathbb{R} \times K)$. Two sets of solutions, one starting from $z_0 = (-1.2, -0.7) \in (\Omega_V \cap (\mathbb{R} \times K)) \subset R_u(z_{eq}) \subset R_{gu}(z_{eq})$ and another starting from $z_0 = (2.8, 0) \in R_{gu}(z_{eq}) \setminus R_u(z_{eq})$ are shown in Fig. 4-5.

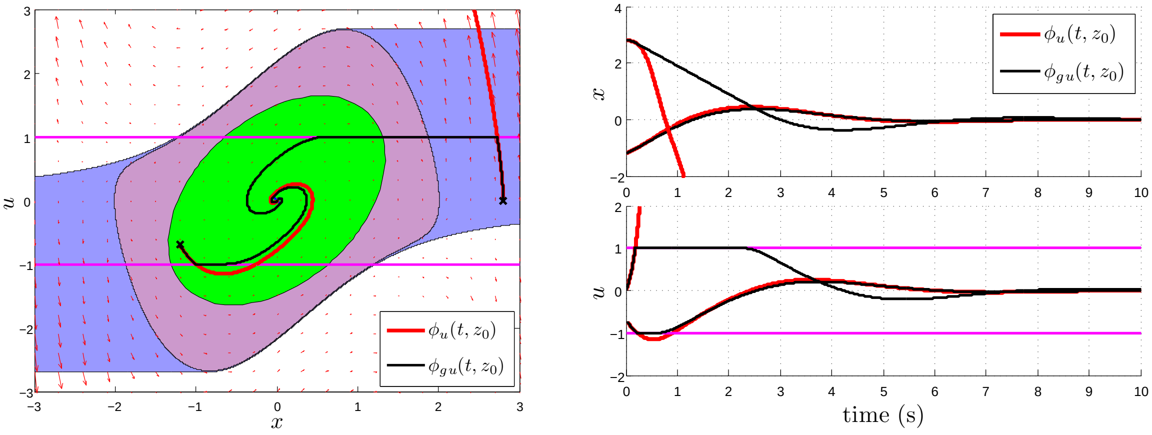


Figure 4-5: Solutions of unconstrained and GPAW-compensated systems.

4.6 Chapter Summary

We showed that the GPAW-compensated controller, apart from being defined by the online solution to a combinatorial optimization subproblem (see Section 2.5), can also be defined by the online solution to a convex quadratic program or a projection onto a convex polyhedral cone problem. Geometric properties of the projection operator were presented, leading to a geometric bounding condition relating the vector fields of the nominal controller and

GPAW-compensated controller. The main results of this chapter are the ROA comparison results Theorems 4.4.1, 4.4.2, and 4.4.3. Of note is Theorem 4.4.3, which yields an admissible GPAW parameter when it is applicable. These ROA comparison results were demonstrated on some simple (linear and nonlinear) planar systems. We note that attempts to apply Theorems 4.4.1, 4.4.2, and 4.4.3 to some simple systems comprising open-loop unstable plants have been unsuccessful. Whether or not Theorems 4.4.1, 4.4.2, and 4.4.3 are applicable to systems with open-loop unstable plants requires further studies.

Chapter 5

Input Constrained MIMO LTI Systems

In this chapter, we restrict consideration to regulatory GPAW-compensated systems comprising input-constrained MIMO LTI plants driven by MIMO LTI controllers. In Section 5.1, we present the system descriptions which are simply specializations of the nonlinear case. A stability result (Theorem 5.2.1) specialized from Theorem 4.4.3 of Chapter 4 is presented in Section 5.2. Theorem 5.2.1 gives sufficient conditions to assert global asymptotic stability of the origin for the GPAW-compensated system. It is applicable only to systems with open-loop stable plants, and is verified by solving a system of linear matrix inequalities, for which efficient solvers are readily available. While there are some attractive features, numerical experience suggest Theorem 5.2.1 to be a conservative result.

In Section 5.3, we study the familiar similarity transformation well known for LTI systems, as applied to GPAW-compensated controllers. We show that the transformed GPAW-compensated controller derived from some nominal controller is equivalent to the GPAW-compensated controller derived from the transformed nominal controller, with the GPAW parameters related through the associated transformation matrix. Despite the GPAW-compensated controller being defined by the online solution to an optimization problem, similarity transformations can be easily performed.

In Section 5.4, we describe *linear systems with partial state constraints*, which has been studied in [103–106]. We present a way to transform the nominal controller into a canonical form more convenient for GPAW compensation in Section 5.5. This canonical form is then used in Section 5.6 to show that under a non-unique choice of the GPAW parameter, the GPAW-compensated system can be transformed into a *linear system with partial state constraints*. This allows results in existing literature (e.g. [103–106]) to be applied to this class of GPAW-compensated systems, and vice versa.

5.1 System Descriptions

Here, we describe the *regulatory* nominal system, unconstrained system, and GPAW-compensated system when the unconstrained plant and nominal controller are both LTI. The input-constrained LTI plant corresponding to (1.1) is described by

$$\begin{aligned} \dot{x} &= Ax + B\,\mathrm{sat}(u), \\ y &= Cx + D\,\mathrm{sat}(u), \end{aligned} \tag{5.1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control input, $y \in \mathbb{R}^p$ is the measurement, $\mathrm{sat}\colon \mathbb{R}^m \to \mathbb{R}^m$ is the saturation function (1.2), and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$ are constant real matrices. We restrict consideration to regulatory systems, for which we assume $r \equiv 0$ in (2.26). Then the nominal LTI controller corresponding to (2.26) is described by[1]

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_c y, \\ u_c &= C_c x_c, \end{aligned} \tag{5.2}$$

where $x_c \in \mathbb{R}^q$ is the state, $y \in \mathbb{R}^p$ is the measurement, $u_c \in \mathbb{R}^m$ is the controller output, and $A_c \in \mathbb{R}^{q\times q}$, $B_c \in \mathbb{R}^{q\times p}$, $C_c \in \mathbb{R}^{m\times q}$ are constant real matrices. The nominal closed-loop system comprising (5.1) and (5.2) with $u = u_c$ can be written as

$$\Sigma_n\colon \begin{cases} \dot{x} = Ax + B\,\mathrm{sat}(C_c x_c), \\ \dot{x}_c = B_c C x + A_c x_c + B_c D\,\mathrm{sat}(C_c x_c), \end{cases} \qquad \text{or} \qquad \Sigma_n\colon \dot{z} = f_n(z), \tag{5.3}$$

where $z := (x, x_c)$ is the system state. Clearly, the unconstrained closed-loop system is described by

$$\Sigma_u\colon \begin{cases} \dot{x} = Ax + BC_c x_c, \\ \dot{x}_c = B_c C x + (A_c + B_c D C_c) x_c, \end{cases} \qquad \text{or} \qquad \Sigma_u\colon \dot{z} = f_u(z) = A_u z, \tag{5.4}$$

with $A_u = \left[\begin{smallmatrix} A & BC_c \\ B_c C & A_c + B_c D C_c \end{smallmatrix}\right] \in \mathbb{R}^{(n+q)\times(n+q)}$.

Using the construction in Section 2.5, the GPAW-compensated controller corresponding to (2.27) and derived from the nominal controller (5.2) can be shown to be

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g, y)(A_c x_g + B_c y), \qquad x_g(0) = x_c(0), \\ u_g &= C_c x_g, \end{aligned} \tag{5.5}$$

where $x_g \in \mathbb{R}^q$ is the state, $y \in \mathbb{R}^p$ is the measurement, $u_g \in \mathbb{R}^m$ is the controller output, (A_c, B_c, C_c) are defined by the nominal controller (5.2), and the projection operator $R_{\mathcal{I}^*}(x_g, y)$ is to be defined next.[2] Let the matrix $C_c \in \mathbb{R}^{m\times q}$ in (5.2) and (5.5) be decomposed into its rows as $C_c = [c_1, c_2, \ldots, c_m]^{\mathrm{T}}$, where $c_i \in \mathbb{R}^q$ for all[3] $i \in \mathcal{I}_m$. Following the construction in Section 2.5, define the $2m$ saturation constraint functions h_i corresponding to (2.28) by

$$h_i(x_g) = c_i^{\mathrm{T}} x_g - u_{\max,i}, \qquad h_{i+m}(x_g) = -c_i^{\mathrm{T}} x_g + u_{\min,i}, \qquad \forall i \in \mathcal{I}_m,$$

whose *constant* gradients are

$$\nabla h_i = -\nabla h_{i+m} = c_i, \qquad \forall i \in \mathcal{I}_m.$$

For any index set $\mathcal{I} \subset \mathcal{I}_{2m}$, define the $q \times \max\{|\mathcal{I}|, 1\}$ matrix corresponding to (2.29) by

$$N_{\mathcal{I}} = \begin{cases} [\nabla h_{\sigma_{\mathcal{I}}(1)}, \nabla h_{\sigma_{\mathcal{I}}(2)}, \ldots, \nabla h_{\sigma_{\mathcal{I}}(|\mathcal{I}|)}], & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases}$$

[1] See Section 2.6 when the nominal LTI controller is of more general structure.

[2] The construction of $R_{\mathcal{I}^*}(x_g, y)$ is presented in Section 2.5 and specialized for LTI systems here.

[3] Recall that $\mathcal{I}_i := \{1, 2, \ldots, i\}$ for any positive integer i.

where $\sigma_{\mathcal{I}}\colon \{1,2,\dots,|\mathcal{I}|\} \to \mathcal{I}$ is a chosen bijection described in Remark 2.5. For any $\mathcal{I} \subset \mathcal{I}_{2m}$ such that $\operatorname{rank}(N_{\mathcal{I}}) = |\mathcal{I}|$, define the projection matrix corresponding to (2.30) by

$$R_{\mathcal{I}} = \begin{cases} I - \Gamma N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}}\Gamma N_{\mathcal{I}})^{-1}N_{\mathcal{I}}^{\mathrm{T}}, & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise,} \end{cases} \tag{5.6}$$

where $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{q\times q}$ is the GPAW parameter. Define the index set of active saturation constraints and candidate solution set

$$\mathcal{I}_{\text{sat}} = \{i \in \mathcal{I}_{2m} \mid h_i(x_g) \geq 0\}, \qquad \mathcal{J} = \{\mathcal{I} \subset \mathcal{I}_{\text{sat}} \mid |\mathcal{I}| \leq q\}.$$

Then at any fixed time (so that $(x_g(t), y(t))$ are fixed), the projection operator $R_{\mathcal{I}^*}(x_g, y)$ in (5.5) is defined by (5.6) and a solution $\mathcal{I}^*$ to the combinatorial optimization subproblem corresponding to (2.31),

$$\begin{aligned} \max_{\mathcal{I}\in\mathcal{J}} F(\mathcal{I}) &= (A_c x_g + B_c y)^{\mathrm{T}}\Gamma^{-1}R_{\mathcal{I}}(A_c x_g + B_c y), \\ \text{subject to} \qquad & \operatorname{rank}(N_{\mathcal{I}}) = |\mathcal{I}|, \\ & N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} R_{\mathcal{I}}(A_c x_g + B_c y) \leq 0. \end{aligned} \tag{5.7}$$

Proposition 2.5.1 ensures that solutions $\mathcal{I}^*$ to subproblem (5.7) always exist. See also Theorem 2.5.3 for the controller state-output consistency property, and Appendix C for a procedure to apply GPAW compensation.

The GPAW-compensated closed-loop system comprising (5.1) and (5.5) with $u = u_g$ can be written as

$$\Sigma_g\colon \begin{cases} \dot{x} = Ax + B\,\mathrm{sat}(C_c x_g), \\ \dot{x}_g = R_{\mathcal{I}^*}(B_c C x + A_c x_g + B_c D\,\mathrm{sat}(C_c x_g)), \end{cases} \qquad \text{or} \qquad \Sigma_g\colon \dot{z}_g = f_g(z_g), \tag{5.8}$$

where $z_g := (x, x_g)$ and $R_{\mathcal{I}^*}(A_c x_g + B_c y) := R_{\mathcal{I}^*}(x_g, y)(A_c x_g + B_c y)$. When the state of the GPAW-compensated controller is initialized such that $\mathrm{sat}(x_g(0)) = x_g(0)$, Theorem 2.5.3 (controller state-output consistency) yields $\mathrm{sat}(x_g(t)) = x_g(t)$ for all $t \geq 0$. Then the GPAW-compensated system can be simplified to

$$\Sigma_{gu}\colon \begin{cases} \dot{x} = Ax + BC_c x_g, \\ \dot{x}_g = R_{\mathcal{I}^*}(B_c C x + (A_c + B_c D C_c)x_g), \end{cases} \qquad \text{or} \qquad \Sigma_{gu}\colon \dot{z}_g = f_{gu}(z_g), \tag{5.9}$$

for all $t \geq 0$, with the $\mathrm{sat}(\cdot)$ function eliminated.[4] As observed in Remark 2.19, the controller state can usually be initialized arbitrarily.

We assume the control objective is to regulate the system state about the origin $z_{eq} = (0,0) \in \mathbb{R}^{n+q}$, which is clearly an equilibrium for systems Σ_n, Σ_u, Σ_g, and Σ_{gu}. To ensure that z_{eq} lies in the interior of the unsaturated region, assume also that the saturation limits in (1.2) satisfy $u_{\min,i} < 0 < u_{\max,i}$ for all $i \in \mathcal{I}_m$.

[4] This is discussed in Section 2.5 where (2.36) (analogous to Σ_{gu} (5.9)) is obtained.

5.2 A Stability Result for Systems with Open-loop Stable Plants

Three region of attraction (ROA) comparison results were presented in Section 4.4. Of note is Theorem 4.4.3, which yields a prescribed GPAW parameter $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{q\times q}$ such that GPAW compensation with parameter Γ ensures "global asymptotic stability" for the equilibrium of the GPAW-compensated system Σ_g, in the sense that the ROA contains the unsaturated region as in [103, 104]. In this section, we use Theorem 4.4.3 to derive a stability result for the GPAW-compensated system, for which the GPAW parameter is found by solving a linear matrix inequality (LMI) problem [36, Section 2.2.1, p. 9]. As will be explained in Remark 5.1 below, this result applies only to systems with open-loop stable plants. A significant aspect is that the result depends only on properties of the *unconstrained* system Σ_u.

Theorem 5.2.1 (Global Asymptotic Stability for LTI GPAW-Compensated System)**.** *Consider the unconstrained system Σ_u (5.4). If there exist symmetric positive definite matrices $P_1 = P_1^{\mathrm{T}} > 0 \in \mathbb{R}^{n\times n}$ and $P_2 = P_2^{\mathrm{T}} > 0 \in \mathbb{R}^{q\times q}$ such that*

$$\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \begin{bmatrix} A & BC_c \\ B_cC & A_c + B_cDC_c \end{bmatrix} + \begin{bmatrix} A & BC_c \\ B_cC & A_c + B_cDC_c \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} < 0, \tag{5.10}$$

then GPAW compensation with parameter $\Gamma = P_2^{-1}$ yields systems Σ_g (5.8) and Σ_{gu} (5.9) whose ROAs contain the unsaturated region $\mathbb{R}^n \times K$, where $K = \{\bar{x} \in \mathbb{R}^q \mid \mathrm{sat}(C_c\bar{x}) = C_c\bar{x}\}$.

Proof. We will be applying Theorem 4.4.3 by comparing against the ROA of the unconstrained system Σ_u (see Remark 4.10). Define $P := \left[\begin{smallmatrix} P_1 & 0 \\ 0 & P_2 \end{smallmatrix}\right]$ and let $V(z) := z^{\mathrm{T}}Pz$ be a Lyapunov function candidate for system Σ_u. Since P_1 and P_2 are symmetric positive definite, P is also symmetric positive definite, so that V is a positive definite function. It is also clear that V is radially unbounded. Since $A_u = \left[\begin{smallmatrix} A & BC_c \\ B_cC & A_c+B_cDC_c \end{smallmatrix}\right]$ and (5.10) holds, we have

$$\frac{\partial V(\bar{z})}{\partial z} f_u(\bar{z}) = \bar{z}^{\mathrm{T}}(PA_u + A_u^{\mathrm{T}}P)\bar{z} < 0, \qquad \forall \bar{z} \neq 0. \tag{5.11}$$

By [37, Theorem 4.2, p. 124], the origin z_{eq} is a globally asymptotically stable equilibrium for the unconstrained system Σ_u, so that the ROA of system Σ_u contains the unsaturated region $\mathbb{R}^n \times K$. Since the vector fields f_u (5.4) and f_{gu} (5.9) coincide in the interior of the unsaturated region $\mathbb{R}^n \times (K \setminus \partial K)$, it follows that z_{eq} is also a locally asymptotically stable equilibrium for system Σ_{gu}.[5]

It can be verified that z_{eq} satisfies (4.29), i.e. it is an equilibrium for the unconstrained system Σ_u that lies within the interior of the unsaturated region. Moreover, since the output equation of the nominal controller (5.2) is linear in the controller state, the unsaturated region K is convex (see Remark 4.4), so that K and $\mathbb{R}^n \times K$ are star domains with kernels $\ker(K)$ and $\mathbb{R}^n \times \ker(K)$ respectively (see Remark 4.2 and Corollary 4.2.2). It can be verified that $z_{eq} \in \mathbb{R}^n \times \ker(K)$, and that $V(z) = V(x, x_c)$ has the form of (4.35) with $V_x(x) = x^{\mathrm{T}}P_1x$ and $P_c = P_2$. Moreover, since V is continuously differentiable and f_u is continuous, it follows that $\dot{V}(z) = \frac{\partial V(z)}{\partial z} f_u(z)$ is continuous, and is also negative definite

[5] The vector fields f_n and f_g also coincide with f_u in the interior of the unsaturated region, and z_{eq} is also a locally asymptotically stable equilibrium for systems Σ_n and Σ_g. However, these are not needed in the proof.

due to (5.11). By [37, Lemma 4.3, p. 145], there exists a class $\mathcal{K}$ function α such that V and f_u satisfy (4.33).

All hypotheses of Theorem 4.4.3 are satisfied, and its application shows that GPAW compensation with parameter $\Gamma = P_c^{-1} = P_2^{-1}$ yields system Σ_{gu} whose ROA contains the unsaturated region $\mathbb{R}^n \times K$. From the definitions of f_g (5.8) and f_{gu} (5.9), it can be seen that f_g and f_{gu} coincide in the unsaturated region, which is a positively invariant set of systems Σ_g and Σ_{gu} due to Theorem 2.5.3. These imply that the ROAs of systems Σ_g and Σ_{gu} within the unsaturated region coincide, and yields the second conclusion, namely that the ROA of system Σ_g (with parameter $\Gamma = P_2^{-1}$) also contains the unsaturated region. ■

Remark 5.1. When the matrix operations on the left-hand-side of (5.10) are carried out, it becomes

$$\begin{bmatrix} P_1 A + A^{\mathrm{T}} P_1 & P_1 B C_c + C^{\mathrm{T}} B_c^{\mathrm{T}} P_2 \\ (P_1 B C_c + C^{\mathrm{T}} B_c^{\mathrm{T}} P_2)^{\mathrm{T}} & P_2 (A_c + B_c D C_c) + (A_c + B_c D C_c)^{\mathrm{T}} P_2 \end{bmatrix} < 0. \tag{5.12}$$

As implied by [124, Theorem 7.7.6, p. 472], *necessary* conditions for (5.12) (and (5.10)) to hold are that the diagonal blocks, i.e. $P_1 A + A^{\mathrm{T}} P_1$ and $P_2(A_c + B_c D C_c) + (A_c + B_c D C_c)^{\mathrm{T}} P_2$, must be negative definite. Hence it is *necessary* for A and $A_c + B_c D C_c$ to be Hurwitz. This observation actually follows from [182, Proposition 3.5]. This means that Theorem 5.2.1 can only be applied to systems with stable open-loop plants. Moreover, if $D = 0$, $A_c + B_c D C_c = A_c$ being Hurwitz means the nominal controller must also be stable. □

The following example adapted from [182, Example 3.6] shows that A and $A_c + B_c D C_c$ being Hurwitz is *not sufficient* to ensure the existence of $P_1 = P_1^{\mathrm{T}} > 0$ and $P_2 = P_2^{\mathrm{T}} > 0$ satisfying (5.10).

Example 5.2.1. Let

$$A_u = \begin{bmatrix} A & B C_c \\ B_c C & A_c + B_c D C_c \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 2 & -1 \end{bmatrix}.$$

It is clear that $A = -1$ and $A_c + B_c D C_c = -1$ are both Hurwitz. However, it can be verified that A_u is not Hurwitz, with eigenvalues of -3 and $+1$. Hence no $P_1 = P_1^{\mathrm{T}} > 0$ and $P_2 = P_2^{\mathrm{T}} > 0$ exist to satisfy (5.10). △

Necessary and sufficient conditions to ensure existence of $P_1 = P_1^{\mathrm{T}} > 0$ and $P_2 = P_2^{\mathrm{T}} > 0$ that satisfy (5.10) are available in [182, Theorem 3.10] for a more general case. We note that condition (5.12) (and hence (5.10)) is an LMI, which admits efficient numerical solutions. This will be demonstrated in the next section on a simple example.

When the LMI (5.10) is feasible, Theorem 5.2.1 yields a GPAW parameter defined by $\Gamma = P_2^{-1}$. As discussed in Remark 2.25, it is desirable for Γ to have a small condition number [124, p. 336]. In view of this, we formulate a generalized eigenvalue problem [36, Section 2.2.3, pp. 10 – 11] to minimize the condition number of the resultant GPAW parameter, applicable whenever (5.10) is feasible. Since $\Gamma = P_2^{-1}$, the definition of its condition number $\kappa(\Gamma)$ [124, p. 336] yields

$$\kappa(\Gamma) = \|\Gamma^{-1}\| \|\Gamma\| = \|P_2\| \|P_2^{-1}\| = \kappa(P_2).$$

Hence minimizing $\kappa(\Gamma)$ is equivalent to minimizing $\kappa(P_2)$. From [36, Section 3.2, p. 38], it can be seen that the solution $P_2 = P_2^{\mathrm{T}} > 0$ to the generalized eigenvalue problem

$$\begin{gathered}\min_{P_1, P_2, \mu, \gamma} \gamma, \\ \text{subject to} \qquad P_1 > 0, \qquad \mu > 0, \qquad \mu I < P_2 < \gamma \mu I, \\ \begin{bmatrix} P_1 A + A^{\mathrm{T}} P_1 & P_1 B C_c + C^{\mathrm{T}} B_c^{\mathrm{T}} P_2 \\ (P_1 B C_c + C^{\mathrm{T}} B_c^{\mathrm{T}} P_2)^{\mathrm{T}} & P_2 (A_c + B_c D C_c) + (A_c + B_c D C_c)^{\mathrm{T}} P_2 \end{bmatrix} < 0, \end{gathered} \tag{5.13}$$

is of minimal condition number, with P_1, P_2 satisfying (5.10). Application of Theorem 5.2.1 remains unchanged, as will be shown in the next section.

5.2.1 Numerical Example

Here, we demonstrate an application of Theorem 5.2.1 on a simple system comprising a saturated second-order SISO LTI plant driven by a nominal second-order SISO LTI controller, where the objective is to regulate the system state about the origin.

The unconstrained stable plant is represented by the transfer function $G(s) = \frac{1}{s^2+s+1}$, which induces a saturated plant with state-space representation (5.1) where

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad D = 0.$$

Assume a nominal controller with transfer function $K(s) = -\frac{22.8s+11}{s^2+8.6s+25}$ has been designed to improve the transient response of the system, that is to be interconnected with the plant by *positive feedback*. The nominal controller has a state-space representation (5.2) where

$$A_c = \begin{bmatrix} -8.6 & -6.25 \\ 4 & 0 \end{bmatrix}, \qquad B_c = \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \qquad C_c = \begin{bmatrix} -5.7 & -0.6875 \end{bmatrix}.$$

The matrices (A, B, C, D, A_c, B_c, C_c) define the unconstrained system Σ_u (5.4) completely, which is all the data required to apply Theorem 5.2.1. Using the LMI solver (or function) `feasp` of the MATLAB® Robust Control Toolbox [183], symmetric positive definite matrices $P_{1f} := P_1$ and $P_{2f} := P_2$ that satisfy the LMI (5.12) (and hence (5.10)) were found to be

$$P_{1f} = \begin{bmatrix} 1.6090 & 0.31710 \\ 0.31710 & 0.66628 \end{bmatrix}, \qquad P_{2f} = \begin{bmatrix} 1.3255 & 0.13730 \\ 0.13730 & 1.7704 \end{bmatrix}.$$

Thus Theorem 5.2.1 shows that GPAW compensation with parameter

$$\Gamma_f = P_{2f}^{-1} = \begin{bmatrix} 0.76056 & -0.058984 \\ -0.058984 & 0.56941 \end{bmatrix},$$

yields the GPAW-compensated systems Σ_g (5.8) and Σ_{gu} (5.9) whose ROAs contain the unsaturated region.

Using the solver (or function) `gevp` of the MATLAB® Robust Control Toolbox [183], the solution to the generalized eigenvalue problem (5.13) was found to be

$$\mu = 1.5046 \times 10^{-26}, \qquad \gamma = 1.2012,$$

$$P_{1o} = \begin{bmatrix} 1.9132 & 0.39208 \\ 0.39208 & 0.78664 \end{bmatrix} \times 10^{-26}, \qquad P_{2o} = \begin{bmatrix} 1.5573 & 0.11354 \\ 0.11354 & 1.7544 \end{bmatrix} \times 10^{-26}.$$

Since we can always scale the GPAW parameter by a constant positive scalar (see Remark 2.15), we can use

$$\Gamma_o = 2.3948 \times 10^{-25} P_{2o}^{-1} = \begin{bmatrix} 15.451 & -1 \\ -1 & 13.715 \end{bmatrix},$$

as the normalized GPAW parameter. It can be verified that the condition numbers of Γ_f and Γ_o are $\kappa(\Gamma_f) = 1.4064 > \kappa(\Gamma_o) = 1.1997$, which shows a marginal improvement (decrease) when solving the generalized eigenvalue problem (5.13) to obtain P_1 and P_2.

Remark 5.2. When the nominal controller (with transfer function $K(s) = -\frac{22.8s+11}{s^2+8.6s+25}$) is represented (equivalently) by matrices

$$A_c = \begin{bmatrix} 0 & 1 \\ -25 & -8.6 \end{bmatrix}, \qquad B_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad C_c = \begin{bmatrix} -11 & -22.8 \end{bmatrix},$$

in (5.2), the numerical solutions change significantly. In particular, the condition numbers become $\kappa(\Gamma_f) = 22.324$ and $\kappa(\Gamma_o) = 17.441$, representing a drastic deterioration. This shows that the numerical solutions are sensitive to coordinate transformations. Moreover, attempts to use nominal controllers with higher bandwidth, e.g. with $K(s) = -\frac{31s+17.4}{s^2+9.6s+30.6}$, have failed in the sense that (5.10) and (5.13) becomes numerically infeasible. This suggests that Theorem 5.2.1 is a conservative result. □

The GPAW-compensated controller (5.5) can be implemented in a few ways summarized in Appendix C. Here, we use the closed-form expressions (A.7) in Appendix A for a more efficient solution. The closed-form expressions (A.7) for the GPAW-compensated controller (5.5) with parameter Γ are[6]

$$\dot{x}_g = \begin{cases} \left(I - \frac{\Gamma C_c^{\mathrm{T}} C_c}{C_c \Gamma C_c^{\mathrm{T}}}\right)(A_c x_g + B_c y), & \text{if } (u_g \geq u_{\max}) \wedge (C_c(A_c x_g + B_c y) > 0), \\ \left(I - \frac{\Gamma C_c^{\mathrm{T}} C_c}{C_c \Gamma C_c^{\mathrm{T}}}\right)(A_c x_g + B_c y), & \text{if } (u_g \leq u_{\min}) \wedge (C_c(A_c x_g + B_c y) < 0), \\ A_c x_g + B_c y, & \text{otherwise,} \end{cases}$$

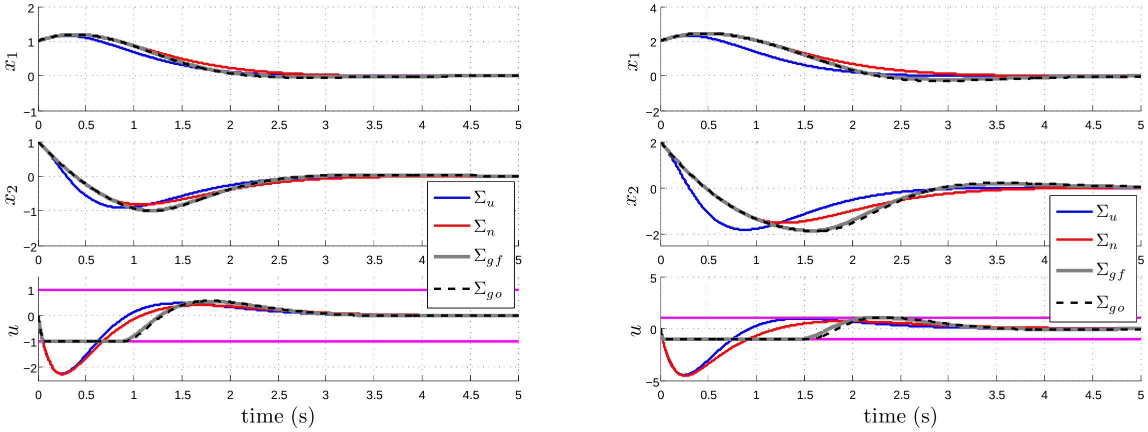
$$u_g = C_c x_g, \qquad x_g(0) = x_c(0).$$

For $\Gamma = \Gamma_f$ and $\Gamma = \Gamma_o$, we have the matrices

$$I - \frac{\Gamma_f C_c^{\mathrm{T}} C_c}{C_c \Gamma_f C_c^{\mathrm{T}}} = \begin{bmatrix} 0.0015497 & -0.12043 \\ -0.012848 & 0.99845 \end{bmatrix}, \qquad I - \frac{\Gamma_o C_c^{\mathrm{T}} C_c}{C_c \Gamma_o C_c^{\mathrm{T}}} = \begin{bmatrix} 0.0051210 & -0.12000 \\ -0.042458 & 0.99488 \end{bmatrix}.$$

Two sets of solutions for the unconstrained system Σ_u (5.4), nominal system Σ_n (5.3), and GPAW-compensated system Σ_g (5.8) are shown in Fig. 5-1. The GPAW-compensated systems with parameters Γ_f and Γ_o are denoted by Σ_{gf} and Σ_{go} respectively. In Fig. 5-1(a), the plant initial condition is $x(0) = (1,1)$, while in Fig. 5-1(b), the plant initial condition is $x(0) = (2,2)$. In both cases, the controller states are set to zero and the plant state is decomposed as $x = [x_1, x_2]^{\mathrm{T}}$. While Theorem 5.2.1 ensures global asymptotic stability for the origin of Σ_{gf} and Σ_{go} in the sense of [103], the time responses in Fig. 5-1 suggests

[6] Note that c_c in (A.7) is given by $c_c = C_c^{\mathrm{T}}$. Moreover, $\wedge$ denotes the logical AND operator.



(a) $x(0) = (1, 1)$

(b) $x(0) = (2, 2)$

Figure 5-1: Comparison of time responses of unconstrained system Σ_u, nominal system Σ_n, and GPAW-compensated systems Σ_{gf} ($\Gamma = \Gamma_f$) and Σ_{go} ($\Gamma = \Gamma_o$), all with zero initial conditions for the associated controllers.

there is little performance improvement when adopting the GPAW scheme. This can be attributed to the conservativeness of Theorem 5.2.1.

5.3 GPAW-Compensated Controller Transformations

Similarity transformations are fundamental operations for LTI systems. In this section, we establish the relation between the *transformed GPAW-compensated controller* and the GPAW-compensated controller derived from the *transformed nominal controller*.

For some nonsingular matrix $T \in \mathbb{R}^{q \times q}$, a similarity transformation is defined by $\hat{x}_g = Tx_g$, where x_g is the state of the GPAW-compensated controller (5.5) derived from (5.2). The transformed GPAW-compensated controller with state $\hat{x}_g$ is described by

$$\begin{aligned} \dot{\hat{x}}_g &= T\dot{x}_g = TR_{\mathcal{I}^*}(T^{-1}\hat{x}_g, y)(A_cT^{-1}\hat{x}_g + B_cy), \qquad \hat{x}_g(0) = Tx_c(0), \\ u_g &= C_cT^{-1}\hat{x}_g, \end{aligned} \tag{5.14}$$

where $x_g = T^{-1}\hat{x}_g$. Define $\hat{R}_{\mathcal{I}} := TR_{\mathcal{I}}T^{-1}$ where $R_{\mathcal{I}}$ is given by (5.6), written explicitly as

$$\hat{R}_{\mathcal{I}} = \begin{cases} I - T\Gamma N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}}\Gamma N_{\mathcal{I}})^{-1}N_{\mathcal{I}}^{\mathrm{T}}T^{-1}, & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise.} \end{cases} \tag{5.15}$$

Then the *transformed* GPAW-compensated controller (5.14) becomes

$$\begin{aligned} \dot{\hat{x}}_g &= \hat{R}_{\mathcal{I}^*}(\hat{x}_g, y)(\tilde{A}_c\hat{x}_g + \tilde{B}_cy), \qquad \hat{x}_g(0) = Tx_c(0), \\ u_g &= \tilde{C}_c\hat{x}_g, \end{aligned} \tag{5.16}$$

where $\mathcal{I}^*$ is a solution to subproblem (5.7) with $x_g = T^{-1}\hat{x}_g$, and

$$\tilde{A}_c := TA_cT^{-1}, \qquad \tilde{B}_c := TB_c, \qquad \tilde{C}_c := C_cT^{-1}. \tag{5.17}$$

Now, consider the transformation of the nominal controller (5.2) defined by the same matrix T and $\tilde{x}_c = Tx_c$, where x_c is the state of the nominal controller. The *transformed* nominal controller with state $\tilde{x}_c$ is given by

$$\begin{aligned} \dot{\tilde{x}}_c &= \tilde{A}_c\tilde{x}_c + \tilde{B}_c y, \qquad \tilde{x}_c(0) = Tx_c(0), \\ u_c &= \tilde{C}_c\tilde{x}_c, \end{aligned} \tag{5.18}$$

where $\tilde{A}_c$, $\tilde{B}_c$, $\tilde{C}_c$ are defined in (5.17). In accordance with the construction in Section 2.5, applying GPAW compensation to the *transformed* nominal controller (5.18) with parameter $\tilde{\Gamma} = \tilde{\Gamma}^{\mathrm{T}} > 0 \in \mathbb{R}^{q\times q}$ and state $\tilde{x}_g$ yields

$$\begin{aligned} \dot{\tilde{x}}_g &= \tilde{R}_{\mathcal{I}^*}(\tilde{x}_g, y)(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y), \qquad \tilde{x}_g(0) = Tx_c(0), \\ u_g &= \tilde{C}_c\tilde{x}_g, \end{aligned} \tag{5.19}$$

where $\tilde{R}_{\mathcal{I}^*}(\tilde{x}_g, y)$ is to be defined. Observe that (5.19) differs from (5.16) only in the projection operator $\tilde{R}_{\mathcal{I}^*}(\tilde{x}_g, y)$ and state definition $\tilde{x}_g$. We will show that under a particular choice of $\tilde{\Gamma}$, the two projection operators $\hat{R}_{\mathcal{I}^*}(\bar{x}_g, y)$ and $\tilde{R}_{\mathcal{I}^*}(\bar{x}_g, y)$ are in fact equal, which implies the equivalence of the GPAW-compensated controllers (5.16) and (5.19).

Observe that the i-th row of $\tilde{C}_c = C_cT^{-1}$ is given by $\tilde{c}_i^{\mathrm{T}} = c_i^{\mathrm{T}}T^{-1}$, where c_i is the i-th row of C_c. Analogous to the saturation constraint functions h_i, their gradients ∇h_i, and the matrices $N_{\mathcal{I}}$, $R_{\mathcal{I}}$ in Section 5.1, we have

$$\begin{gathered} \tilde{h}_i(\tilde{x}_g) = c_i^{\mathrm{T}}T^{-1}\tilde{x}_g - u_{\max,i}, \qquad \tilde{h}_{i+m}(\tilde{x}_g) = -c_i^{\mathrm{T}}T^{-1}\tilde{x}_g + u_{\min,i}, \\ \nabla\tilde{h}_i = -\nabla\tilde{h}_{i+m} = T^{-\mathrm{T}}c_i, \end{gathered} \qquad \forall i \in \mathcal{I}_m, \tag{5.20}$$

$$\tilde{N}_{\mathcal{I}} = \begin{cases} [\nabla\tilde{h}_{\sigma_{\mathcal{I}}(1)}, \nabla\tilde{h}_{\sigma_{\mathcal{I}}(2)}, \ldots, \nabla\tilde{h}_{\sigma_{\mathcal{I}}(|\mathcal{I}|)}], & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \tag{5.21}$$

$$\tilde{R}_{\mathcal{I}} = \begin{cases} I - \tilde{\Gamma}\tilde{N}_{\mathcal{I}}(\tilde{N}_{\mathcal{I}}^{\mathrm{T}}\tilde{\Gamma}\tilde{N}_{\mathcal{I}})^{-1}\tilde{N}_{\mathcal{I}}^{\mathrm{T}}, & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise.} \end{cases} \tag{5.22}$$

With the active saturation constraint set and candidate solution set

$$\mathcal{I}_{\mathrm{sat}} = \{i \in \mathcal{I}_{2m} \mid \tilde{h}_i(\tilde{x}_g) \geq 0\}, \qquad \mathcal{J} = \{\mathcal{I} \subset \mathcal{I}_{\mathrm{sat}} \mid |\mathcal{I}| \leq q\},$$

the projection operator $\tilde{R}_{\mathcal{I}^*}(\tilde{x}_g, y)$ in (5.19) is defined by (5.22) and a solution $\mathcal{I}^*$ to the combinatorial optimization subproblem analogous to (5.7)

$$\begin{aligned} \max_{\mathcal{I}\in\mathcal{J}} \tilde{F}(\mathcal{I}) &= (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}}\tilde{\Gamma}^{-1}\tilde{R}_{\mathcal{I}}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y), \\ \text{subject to} \qquad & \operatorname{rank}(\tilde{N}_{\mathcal{I}}) = |\mathcal{I}|, \\ & \tilde{N}_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}\tilde{R}_{\mathcal{I}}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) \leq 0. \end{aligned} \tag{5.23}$$

In relation to ∇h_i and $N_{\mathcal{I}}$ defined in Section 5.1, it can be seen that $\nabla\tilde{h}_i = T^{-\mathrm{T}}\nabla h_i$ and $\tilde{N}_{\mathcal{I}} = T^{-\mathrm{T}}N_{\mathcal{I}}$. Moreover, from (5.22) and (5.15), direct computation shows that $\tilde{R}_{\mathcal{I}} = \hat{R}_{\mathcal{I}}$ $(= TR_{\mathcal{I}}T^{-1})$ if and only if $\tilde{\Gamma} = T\Gamma T^{\mathrm{T}}$. Therefore, when $\tilde{\Gamma} = T\Gamma T^{\mathrm{T}}$, the *transformed* GPAW-compensated controller (5.16) is equivalent to the GPAW-compensated controller (5.19) *derived* from the transformed nominal controller (5.18), provided that subproblem (5.23) is equivalent to subproblem (5.7) when $\tilde{x}_g = Tx_g$. Using (5.17), $\tilde{x}_g = Tx_g$, $\tilde{\Gamma} = T\Gamma T^{\mathrm{T}}$,

$\tilde{R}_{\mathcal{I}} = TR_{\mathcal{I}}T^{-1}$, and $\tilde{N}_{\mathcal{I}} = T^{-\mathrm{T}}N_{\mathcal{I}}$, the objective function of subproblem (5.23) is

$$\begin{aligned}\tilde{F}(\mathcal{I}) &= (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}}\tilde{\Gamma}^{-1}\tilde{R}_{\mathcal{I}}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y),\\ &= (TA_cT^{-1}Tx_g + TB_c y)^{\mathrm{T}}(T\Gamma T^{\mathrm{T}})^{-1}TR_{\mathcal{I}}T^{-1}(TA_cT^{-1}Tx_g + TB_c y),\\ &= (A_cx_g + B_cy)^{\mathrm{T}}\Gamma^{-1}R_{\mathcal{I}}(A_cx_g + B_cy) = F(\mathcal{I}),\end{aligned}$$

and the constraint functions are

$$\operatorname{rank}(\tilde{N}_{\mathcal{I}}) = \operatorname{rank}(T^{-\mathrm{T}}N_{\mathcal{I}}) = \operatorname{rank}(N_{\mathcal{I}}),$$
$$\tilde{N}_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}\tilde{R}_{\mathcal{I}}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) = N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}T^{-1}TR_{\mathcal{I}}T^{-1}(TA_cT^{-1}Tx_g + TB_cy) = N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}R_{\mathcal{I}}(A_cx_g + B_cy),$$

which correspond to the objective and constraint functions of subproblem (5.7). Hence subproblem (5.23) is equivalent to subproblem (5.7), which implies the equivalence of the GPAW-compensated controllers (5.16) and (5.19). We summarize this result in the following.

Proposition 5.3.1 (Similarity Transformation of GPAW-Compensated Controllers)**.** *Let* $T \in \mathbb{R}^{q\times q}$ *be* nonsingular*, and consider the nominal controller* (5.2)*, the GPAW-compensated controller* (5.5)*, and the transformed nominal controller* (5.18) *under the similarity transformation defined by* $\tilde{x}_c = Tx_c$*. Under the similarity transformation defined by* $\hat{x}_g = Tx_g$*, the* transformed *GPAW-compensated controller* (5.16) *with parameter* Γ *is equivalent to the GPAW-compensated controller* (5.19) *with parameter* $T\Gamma T^{\mathrm{T}}$ derived *from the transformed nominal controller* (5.18)*.*

With controllers (5.2), (5.18), (5.5), (5.16), and (5.19) denoted by Σ_{cn}, $\tilde{\Sigma}_{cn}$, Σ_{cg}, $\hat{\Sigma}_{cg}$, and $\tilde{\Sigma}_{cg}$ respectively, this result is illustrated in Fig. 5-2. Proposition 5.3.1 shows that given some GPAW-compensated LTI controller (5.5) with parameter Γ, we can always transform it into an equivalent GPAW-compensated controller with parameter $\tilde{\Gamma} = I$, the identity matrix. To see this, observe that $\Gamma \in \mathbb{R}^{q\times q}$ is symmetric positive definite, so that it can always be decomposed as $\Gamma = \Phi\Phi^{\mathrm{T}}$ for some nonsingular $\Phi \in \mathbb{R}^{q\times q}$ [124, Theorem 7.2.7, p. 406]. Using $T = \Phi^{-1}$ as the transformation matrix yields $\tilde{\Gamma} = T\Gamma T^{\mathrm{T}} = \Phi^{-1}\Phi\Phi^{\mathrm{T}}\Phi^{-\mathrm{T}} = I$. Proposition 5.3.1 shows that the transformed GPAW-compensated controller (5.16) (with $T = \Phi^{-1}$) is equivalent to the GPAW-compensated controller (5.19) (with parameter $\tilde{\Gamma} = T\Gamma T^{\mathrm{T}} = I$) derived from the transformed nominal controller (5.18) (with $T = \Phi^{-1}$). Such a transformation may be convenient for realization purposes or further analysis.

$$\begin{tikzcd}
\boxed{\Sigma_{cn}} \arrow[r, Rightarrow, "{\text{GPAW}, \Gamma}"] \arrow[d, leftrightarrow, "T"'] & \boxed{\Sigma_{cg}} \arrow[d, leftrightarrow, "T"] \\
\boxed{\tilde{\Sigma}_{cn}} \arrow[r, Rightarrow, "{\text{GPAW}, T\Gamma T^{\mathrm{T}}}"] & \boxed{\tilde{\Sigma}_{cg} \equiv \hat{\Sigma}_{cg}}
\end{tikzcd}$$

Figure 5-2: Illustration of transformation equivalence in Proposition 5.3.1 for some *nonsingular* matrix $T \in \mathbb{R}^{q\times q}$ and symmetric positive definite $\Gamma \in \mathbb{R}^{q\times q}$.

5.4 Linear Systems with Partial State Constraints

Our next goal is to establish a link between GPAW-compensated LTI systems and *linear systems with partial state constraints*, which has been studied in [103–106]. Such a link, like the relation with *projected dynamical systems* [107–110] (see Section 3.2), is strategic in nature, and allows cross utilization of ideas and methods. Here, we describe linear systems with partial state constraints.

Define the m-dimensional *hypercube* $\mathbb{D}^m \subset \mathbb{R}^m$ by

$$\mathbb{D}^m := \{(\bar{\phi}_1, \bar{\phi}_2, \ldots, \bar{\phi}_m) \in \mathbb{R}^m \mid -1 \leq \bar{\phi}_i \leq 1, \forall i \in \mathcal{I}_m\}. \tag{5.24}$$

Linear systems with partial state constraints are described by ODEs of the form [103]

$$\begin{aligned} \dot{\theta} &= A_g\theta + B_g\phi, & \theta(0) &= \theta_0, \\ \dot{\phi} &= h(C_g\theta + D_g\phi), & \phi(0) &= \phi_0, \end{aligned} \tag{5.25}$$

where $\theta, \theta_0 \in \mathbb{R}^{n_\theta}$, $\phi, \phi_0 \in \mathbb{D}^m$ are the (partial) states and initial states, and A_g, B_g, C_g, D_g are real matrices of appropriate dimensions. Decomposing the vectors ϕ and $C_g\theta + D_g\phi$ into $\phi = [\phi_1, \phi_2, \ldots, \phi_m]^{\mathrm{T}}$ and $C_g\theta + D_g\phi = \psi = [\psi_1, \psi_2, \ldots, \psi_m]^{\mathrm{T}}$ respectively, the function $h(C_g\theta + D_g\phi) = h(\psi)$ is defined by

$$\begin{aligned} h(C_g\theta + D_g\phi) &= [\tilde{h}(\psi_1, \phi_1), \tilde{h}(\psi_2, \phi_2), \ldots, \tilde{h}(\psi_m, \phi_m)]^{\mathrm{T}}, \\ \tilde{h}(\tilde{\psi}, \tilde{\phi}) &= \begin{cases} 0, & \text{if } |\tilde{\phi}| = 1, \tilde{\psi}\tilde{\phi} \geq 0, \\ \tilde{\psi}, & \text{otherwise.} \end{cases} \end{aligned} \tag{5.26}$$

Existence and uniqueness of solutions to system (5.25) have been established in [103], together with some sufficient conditions for the ROA of the origin to contain $\mathbb{R}^{n_\theta} \times \mathbb{D}^m$. This ROA containment is called “global asymptotic stability” in [103, 104], which we have adopted. See [104–106] for further results on linear systems with partial state constraints.

5.5 A Canonical Form for GPAW Compensation

Here, we transform the nominal controller (5.2) to an equivalent realization that is more convenient for GPAW compensation. This realization has the form

$$\begin{aligned} \dot{x}_{cc} &= A_{cc}x_{cc} + B_{cc}y, & x_{cc}(0) &= x_{cc0}, \\ u_c &= C_{cc}x_{cc} = \begin{bmatrix} 0 & I_m \end{bmatrix} x_{cc}, \end{aligned} \tag{5.27}$$

where I_j (for some positive integer j) is the $j \times j$ identity matrix, $x_{cc}, x_{cc0} \in \mathbb{R}^{\tilde{q}}$ ($\tilde{q} \geq q$) are the (possibly augmented) state and initial state, and the controller output u_c is *part of the controller state* as implied by the special structure of $C_{cc} = \begin{bmatrix} 0 & I_m \end{bmatrix}$ (or $C_{cc} = I_m$ when $\tilde{q} = m$). Since $C_c \in \mathbb{R}^{m \times q}$ (see (5.2)), it is clear that $\mathrm{rank}(C_c) \leq \min\{m, q\} \leq m$ [124, p. 13]. Moreover, for any meaningful control design, $C_c \neq 0$ must hold, so that $\mathrm{rank}(C_c) > 0$. When $\mathrm{rank}(C_c) = m$, the transformation will yield $\tilde{q} = q$, i.e. the controller order is retained. When $\mathrm{rank}(C_c) < m$ the transformation will yield $\tilde{q} > q$, i.e. the transformed controller (5.27) is of higher order and must be a non-minimal realization of (5.2). As will be shown in Section 5.6, application of GPAW compensation on controllers of the form (5.27) with the particular form of C_{cc} yields a significant amount of simplification.

5.5.1 Full Row Rank C_c

Consider when $\operatorname{rank}(C_c) = m$, which implies $q \geq m$. Then all m rows of C_c are *linearly independent*. Defining $\tilde{q} := q$, we can choose a full rank matrix $\tilde{T} \in \mathbb{R}^{(\tilde{q}-m)\times\tilde{q}}$ such that $T := \left[\begin{smallmatrix}\tilde{T}\\ C_c\end{smallmatrix}\right] \in \mathbb{R}^{\tilde{q}\times\tilde{q}}$ is nonsingular (or define $T := C_c$ if $q = m$). By defining $x_{cc} := Tx_c$, it can be verified that similarity transformation of (5.2) yields (5.27) with

$$A_{cc} = TA_cT^{-1}, \qquad B_{cc} = TB_c, \qquad C_{cc} = C_cT^{-1}, \qquad x_{cc0} = Tx_c(0).$$

Since $TT^{-1} = \left[\begin{smallmatrix}\tilde{T}\\ C_c\end{smallmatrix}\right]T^{-1} = \left[\begin{smallmatrix}\tilde{T}T^{-1}\\ C_cT^{-1}\end{smallmatrix}\right] = I_{\tilde{q}} = \left[\begin{smallmatrix}I_{\tilde{q}-m} & 0\\ 0 & I_m\end{smallmatrix}\right]$, we see that C_cT^{-1} comprises the lower m rows of the identity matrix $I_{\tilde{q}}$, so that we have $C_{cc} = C_cT^{-1} = \begin{bmatrix}0 & I_m\end{bmatrix}$ as desired.

5.5.2 Row Rank Deficient C_c

Now, consider when $\alpha := \operatorname{rank}(C_c) < m$. Clearly, we have $0 < \alpha < m$, and there are α *linearly independent* rows in C_c. We can choose a suitable *permutation matrix* $P \in \mathbb{R}^{m\times m}$, $P^{-1} = P^{\mathrm{T}}$ [124, pp. 25 – 26] such that the topmost α rows of PC_c are *linearly independent*. Then the matrices P and PC_c can be partitioned as $PC_c = \left[\begin{smallmatrix}P_1\\ P_2\end{smallmatrix}\right]C_c = \left[\begin{smallmatrix}P_1C_c\\ P_2C_c\end{smallmatrix}\right] = \left[\begin{smallmatrix}C_{c1}\\ C_{c2}\end{smallmatrix}\right]$ where the rows of $C_{c1} := P_1C_c \in \mathbb{R}^{\alpha\times q}$ are *linearly independent*, and $C_{c2} := P_2C_c \in \mathbb{R}^{(m-\alpha)\times q}$ has all its rows *linearly dependent* on the rows of C_{c1}. For convenience, define $\tilde{u}_{c1} := C_{c1}x_c$, $\tilde{u}_{c2} := C_{c2}x_c$ and $\tilde{u}_c := (\tilde{u}_{c1}, \tilde{u}_{c2})$, so that $\tilde{u}_c = PC_cx_c = Pu_c$ and $u_c = P^{\mathrm{T}}\tilde{u}_c$ (due to $P^{-1} = P^{\mathrm{T}}$).

Since $C_{c1} \in \mathbb{R}^{\alpha\times q}$ has full row rank (which implies $q \geq \alpha$), we can choose a full rank matrix $\tilde{T} \in \mathbb{R}^{(q-\alpha)\times q}$ such that $\hat{T} := \left[\begin{smallmatrix}\tilde{T}\\ C_{c1}\end{smallmatrix}\right] \in \mathbb{R}^{q\times q}$ is nonsingular (or define $\hat{T} := C_{c1}$ if $q = \alpha$). Defining $\hat{x}_c := \hat{T}x_c$, the transformed state equation of (5.2) and intermediate output $\tilde{u}_{c1}$ are governed by

$$\begin{aligned}\dot{\hat{x}}_c &= \hat{T}A_c\hat{T}^{-1}\hat{x}_c + \hat{T}B_cy, \qquad\qquad \hat{x}_c(0) = \hat{T}x_c(0),\\ \tilde{u}_{c1} &= C_{c1}\hat{T}^{-1}\hat{x}_c = \begin{bmatrix}0 & I_\alpha\end{bmatrix}\hat{x}_c,\end{aligned} \tag{5.28}$$

where $C_{c1}\hat{T}^{-1} = \begin{bmatrix}0 & I_\alpha\end{bmatrix}$ follows from $\hat{T}\hat{T}^{-1} = \left[\begin{smallmatrix}\tilde{T}\\ C_{c1}\end{smallmatrix}\right]\hat{T}^{-1} = I_q$. Taking the time derivative of $\tilde{u}_{c2}$ and expressing in terms of $\hat{x}_c$ yields

$$\dot{\tilde{u}}_{c2} = C_{c2}A_c\hat{T}^{-1}\hat{x}_c + C_{c2}B_cy, \qquad \tilde{u}_{c2}(0) = C_{c2}x_c(0). \tag{5.29}$$

Based on (5.28) and (5.29), we can define an intermediate controller with augmented state $\bar{x}_c := (\hat{x}_c, \tilde{u}_{c2}) \in \mathbb{R}^{\tilde{q}}$, $\tilde{q} = q + m - \alpha$, and output u_c, described by

$$\begin{aligned}\dot{\bar{x}}_c &= \tilde{A}_c\bar{x}_c + \tilde{B}_cy, \qquad \bar{x}_c(0) = \bar{x}_{c0},\\ u_c &= P^{\mathrm{T}}\tilde{u}_c = P^{\mathrm{T}}\begin{bmatrix}0 & I_\alpha & 0\\ 0 & 0 & I_{m-\alpha}\end{bmatrix}\bar{x}_c = P^{\mathrm{T}}\begin{bmatrix}0 & I_m\end{bmatrix}\bar{x}_c,\end{aligned} \tag{5.30}$$

where

$$\tilde{A}_c = \begin{bmatrix}\hat{T}A_c\hat{T}^{-1} & 0\\ C_{c2}A_c\hat{T}^{-1} & 0\end{bmatrix}, \qquad \tilde{B}_c = \begin{bmatrix}\hat{T}\\ C_{c2}\end{bmatrix}B_c, \qquad \bar{x}_{c0} = \begin{bmatrix}\hat{T}\\ C_{c2}\end{bmatrix}x_c(0).$$

Observe that $\tilde{u}_c = (\tilde{u}_{c1}, \tilde{u}_{c2})$ forms the lower m elements of $\bar{x}_c$. It can be verified that the intermediate controller (5.30) can be obtained by similarity transformation of the $\tilde{q}$-th order *augmented nominal controller* with state $x_{ca} := (x_c, \tilde{u}_{c2})$,

$$\begin{aligned} \dot{x}_{ca} &= \begin{bmatrix} A_c & 0 \\ C_{c2}A_c & 0 \end{bmatrix} x_{ca} + \begin{bmatrix} I_q \\ C_{c2} \end{bmatrix} B_c y, \qquad x_{ca}(0) = \begin{bmatrix} I_q \\ C_{c2} \end{bmatrix} x_c(0), \\ u_c &= P^{\mathrm{T}} \begin{bmatrix} C_{c1} & 0 \\ 0 & I_{m-\alpha} \end{bmatrix} x_{ca}, \end{aligned} \tag{5.31}$$

by defining $\bar{x}_c := T_{ca}x_{ca}$ with $T_{ca} = \left[\begin{smallmatrix} \hat{T} & 0 \\ 0 & I_{m-\alpha} \end{smallmatrix}\right]$.

Remark 5.3. Note that an equivalent realization of (5.31) is with the output equation $u_c = \begin{bmatrix} C_c & 0 \end{bmatrix} x_{ca}$. However, the form as presented in (5.31) is needed for effective use of the augmented state $\tilde{u}_{c2}$ in GPAW compensation, which manipulates the controller state to enforce the saturation constraints. The realization in (5.31) renders each output element independent of all other output elements, so that each can be saturated independently of all others, allowing full exploitation of available control authority. See Section 5.5.3 for further discussions. □

The output equation of (5.30) can be written as

$$u_c = P^{\mathrm{T}} \begin{bmatrix} 0 & I_m \end{bmatrix} \bar{x}_c = \begin{bmatrix} 0 & P^{\mathrm{T}} \end{bmatrix} \bar{x}_c = \begin{bmatrix} 0 & I_m \end{bmatrix} \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{bmatrix} \bar{x}_c = \begin{bmatrix} 0 & I_m \end{bmatrix} T\bar{x}_c,$$

where $T := \left[\begin{smallmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{smallmatrix}\right] \in \mathbb{R}^{(q+m-\alpha)\times(q+m-\alpha)}$ is nonsingular. The special structure of (5.27) can then be obtained by a final similarity transformation defined by $x_{cc} := T\bar{x}_c = TT_{ca}x_{ca}$, which immediately yields $C_{cc} = \begin{bmatrix} 0 & I_m \end{bmatrix}$ in (5.27). Partition the inverse matrix $\hat{T}^{-1}$ as $\hat{T}^{-1} = \begin{bmatrix} \hat{T}_{i1} & \hat{T}_{i2} \end{bmatrix}$, and observe that $\left[\begin{smallmatrix} \hat{T} \\ C_{c2} \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{T} \\ PC_c \end{smallmatrix}\right]$. The remaining vectors and matrices in (5.27) are (see (5.30))

$$\begin{aligned} A_{cc} = T\tilde{A}_cT^{-1} &= T \begin{bmatrix} \hat{T}A_c\hat{T}^{-1} & 0 \\ C_{c2}A_c\hat{T}^{-1} & 0 \end{bmatrix} \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P \end{bmatrix}, \\ &= T \begin{bmatrix} \hat{T}A_c\hat{T}_{i1} & \hat{T}A_c\hat{T}_{i2} & 0 \\ C_{c2}A_c\hat{T}_{i1} & C_{c2}A_c\hat{T}_{i2} & 0 \end{bmatrix} \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P_1 \\ 0 & P_2 \end{bmatrix} = T \begin{bmatrix} \hat{T}A_c\hat{T}_{i1} & \hat{T}A_c\hat{T}_{i2}P_1, \\ C_{c2}A_c\hat{T}_{i1} & C_{c2}A_c\hat{T}_{i2}P_1 \end{bmatrix}, \\ &= T \begin{bmatrix} \hat{T} \\ C_{c2} \end{bmatrix} A_c \begin{bmatrix} \hat{T}_{i1} & \hat{T}_{i2}P_1 \end{bmatrix} = \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \tilde{T} \\ PC_c \end{bmatrix} A_c \begin{bmatrix} \hat{T}_{i1} & \hat{T}_{i2}P_1 \end{bmatrix}, \\ &= \begin{bmatrix} \tilde{T} \\ C_c \end{bmatrix} A_c \begin{bmatrix} \hat{T}_{i1} & \hat{T}_{i2}P_1 \end{bmatrix} = \begin{bmatrix} \tilde{T} \\ C_c \end{bmatrix} A_c\hat{T}^{-1} \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P_1 \end{bmatrix}, \\ B_{cc} = T\tilde{B}_c &= \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \hat{T} \\ C_{c2} \end{bmatrix} B_c = \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \tilde{T} \\ PC_c \end{bmatrix} B_c = \begin{bmatrix} \tilde{T} \\ C_c \end{bmatrix} B_c, \\ x_{cc0} = T\bar{x}_{c0} &= \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \hat{T} \\ C_{c2} \end{bmatrix} x_c(0) = \begin{bmatrix} I_{q-\alpha} & 0 \\ 0 & P^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \tilde{T} \\ PC_c \end{bmatrix} x_c(0) = \begin{bmatrix} \tilde{T} \\ C_c \end{bmatrix} x_c(0). \end{aligned}$$

In the preceding expressions, recall that $\tilde{T} \in \mathbb{R}^{(q-\alpha)\times q}$ is a full rank matrix chosen such that $\hat{T}$ is nonsingular, and P_1 is the top partition of the permutation matrix P.

5.5.3 Comments on Nominal Controller Transformations

Instead of the preceding techniques, the nominal controller (5.2) can always be transformed into an equivalent non-minimal $(q+m)$-th order controller with state $x_{cc} := (x_c, u)$, described by

$$\begin{aligned}\dot{x}_{cc} &= \begin{bmatrix} A_c & 0 \\ C_c A_c & 0 \end{bmatrix} x_{cc} + \begin{bmatrix} B_c \\ C_c B_c \end{bmatrix} y, \qquad x_{cc}(0) = \begin{bmatrix} I_q \\ C_c \end{bmatrix} x_c(0), \\ u &= \begin{bmatrix} 0 & I_m \end{bmatrix} x_{cc},\end{aligned}$$

which is of the form of (5.27). However, as stated in Remark 2.23, GPAW compensation with $\Gamma = I_{q+m}$ on this equivalent controller yields effectively *no anti-windup compensation*. Clearly, controller state augmentation for GPAW compensation needs further study, which we leave as future work (see Section 7.1.10).

On another aspect, observe that the GPAW scheme can be applied to the nominal controller (5.2) without any transformation to the form of (5.27). Doing so in the preceding sections is only to yield simplifications in the final controller description. When C_c is row rank deficient, i.e. $\text{rank}(C_c) = \alpha < m$, direct application of the GPAW scheme on (5.2) results in a *closed-loop* unsaturated region of the form $\mathbb{R}^{n+q-\alpha} \times \tilde{\mathbb{D}}^{\alpha}$ (or can be transformed to such a form), where $\tilde{\mathbb{D}}^{\alpha}$ is a *subset* of the hypercube $\mathbb{D}^{\alpha}$ in general (in fact $\tilde{\mathbb{D}}^{\alpha}$ would be a polyhedron with possibly more faces than the hypercube $\mathbb{D}^{\alpha}$). In other words, there could be *redundant* saturation constraints. This means that the available control authority may not be exploited to the fullest, i.e. some control outputs that may not have reached the saturation limits may be inadvertently limited by other linearly dependent saturated control outputs. With the transformation in Section 5.5.2, the corresponding unsaturated region is of the form $\mathbb{R}^{n+\tilde{q}-m} \times \mathbb{D}^m = \mathbb{R}^{n+q-\alpha} \times \mathbb{D}^m$, which is the type encountered in linear systems with partial state constraints [103–106]. Doing so also allows each control output to be constrained independently of other (originally linearly dependent) control outputs. We leave the study on effective handling of redundant saturation constraints as future work (see Section 7.1.10).

5.6 A Relation between GPAW-Compensated LTI Systems and Linear Systems with Partial State Constraints

Here, we present a link between GPAW-compensated LTI systems and *linear systems with partial state constraints* [103] described in Section 5.4. As mentioned in Section 5.4, this link is strategic in nature and allows cross utilization of ideas and methods in GPAW compensation and existing literature. We show that under a (non-unique) choice of the GPAW parameter, the GPAW-compensated system (5.9) (or one obtained by GPAW compensation on a non-minimal equivalent realization of (5.2), e.g. the augmented controller (5.31)) can be transformed into a *linear system with partial state constraints* of the form (5.25), which has been studied in [103–106]. Clearly, stability/performance properties of system (5.25) then carry over to system (5.9), and vice versa. This allows existing results in [103–106] to be applied to such GPAW-compensated systems.

We start from the $\tilde{q}$-th order controller (5.27), which is a (possibly non-minimal) realization of the nominal controller (5.2). First, we need to make the assumption that the saturation limits are *symmetric*,[7] i.e. $u_{\max,i} = -u_{\min,i}$ for all $i \in \mathcal{I}_m$ (see (1.2)), and perform a final similarity transformation of (5.27) by defining $\tilde{x}_c := Tx_{cc}$, where $T = \begin{bmatrix} I_{\tilde{q}-m} & 0 \\ 0 & \Lambda \end{bmatrix} \in \mathbb{R}^{\tilde{q}\times\tilde{q}}$ and $\Lambda = \operatorname{diag}(u_{\max,1}^{-1}, u_{\max,2}^{-1}, \ldots, u_{\max,m}^{-1}) \in \mathbb{R}^{m\times m}$ is the diagonal matrix with entries comprising the ordered reciprocals of the saturation limits. The transformed nominal controller is then described by

$$\begin{aligned} \dot{\tilde{x}}_c &= \tilde{A}_c\tilde{x}_c + \tilde{B}_c y, \qquad \tilde{x}_c(0) = Tx_{cc}(0), \\ u_c &= \tilde{C}_c\tilde{x}_c, \end{aligned} \tag{5.32}$$

where

$$\tilde{A}_c = TA_{cc}T^{-1}, \qquad \tilde{B}_c = TB_{cc}, \qquad \tilde{C}_c = C_{cc}T^{-1}. \tag{5.33}$$

The resulting $\tilde{q}$-th order GPAW-compensated controller is defined by the preceding matrices and has the form

$$\begin{aligned} \dot{\tilde{x}}_g &= \tilde{R}_{\mathcal{I}^*}(\tilde{x}_g, y)(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y), \qquad \tilde{x}_g(0) = Tx_{cc}(0), \\ u_g &= \tilde{C}_c\tilde{x}_g. \end{aligned} \tag{5.34}$$

Application of GPAW compensation then proceeds as in Section 5.1, with some simplifications to be described.

Due to the special structure of $C_{cc} = \begin{bmatrix} 0 & I_m \end{bmatrix}$ (see (5.27)), it can be verified that each row of $\tilde{C}_c = C_{cc}T^{-1} = \begin{bmatrix} 0 & \Lambda^{-1} \end{bmatrix} = [\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_m]^{\mathrm{T}}$ has the form[8]

$$\tilde{c}_i^{\mathrm{T}} = u_{\max,i}[\vec{0}_{\tilde{q}-m}^{\mathrm{T}}, e_i^{\mathrm{T}}] = u_{\max,i}[\vec{0}_{\tilde{q}-m}^{\mathrm{T}}, \underbrace{0,\ldots,0}_{(i-1)\text{ zeros}}, 1, \underbrace{0,\ldots,0}_{(m-i)\text{ zeros}}] = u_{\max,i}\tilde{e}_{\tilde{q}-m+i}^{\mathrm{T}}, \qquad \forall i \in \mathcal{I}_m,$$

where $\vec{0}_{\tilde{q}-m}$ is the zero vector in $\mathbb{R}^{\tilde{q}-m}$, $e_j \in \mathbb{R}^m$ for $j \in \mathcal{I}_m$ is the j-th standard basis vector in $\mathbb{R}^m$ with 1 as its j-th element and 0's elsewhere, and $\tilde{e}_k \in \mathbb{R}^{\tilde{q}}$ for $k \in \mathcal{I}_{\tilde{q}}$ is the k-th standard basis vector in $\mathbb{R}^{\tilde{q}}$. Partitioning the GPAW controller state in (5.34) as $\tilde{x}_g = (\tilde{x}_{gf}, \phi) \in \mathbb{R}^{\tilde{q}}$ with $\tilde{x}_{gf} \in \mathbb{R}^{\tilde{q}-m}$ and $\phi = [\phi_1, \phi_2, \ldots, \phi_m]^{\mathrm{T}} \in \mathbb{R}^m$ as the partial states, the saturation constraint functions corresponding to (5.20) simplify to

$$\begin{aligned} \tilde{h}_i(\tilde{x}_g) &= \tilde{c}_i^{\mathrm{T}}\tilde{x}_g - u_{\max,i} = u_{\max,i}[\vec{0}_{\tilde{q}-m}^{\mathrm{T}}, e_i^{\mathrm{T}}]\begin{bmatrix} \tilde{x}_{gf} \\ \phi \end{bmatrix} - u_{\max,i} = u_{\max,i}(\phi_i - 1), \\ \tilde{h}_{i+m}(\tilde{x}_g) &= -\tilde{c}_i^{\mathrm{T}}\tilde{x}_g + u_{\min,i} = -u_{\max,i}[\vec{0}_{\tilde{q}-m}^{\mathrm{T}}, e_i^{\mathrm{T}}]\begin{bmatrix} \tilde{x}_{gf} \\ \phi \end{bmatrix} - u_{\max,i} = -u_{\max,i}(\phi_i + 1), \end{aligned} \tag{5.35}$$

for all $i \in \mathcal{I}_m$. The corresponding constant gradient vectors are

$$\nabla\tilde{h}_i = -\nabla\tilde{h}_{i+m} = \tilde{c}_i = u_{\max,i}[\vec{0}_{\tilde{q}-m}^{\mathrm{T}}, e_i^{\mathrm{T}}]^{\mathrm{T}}, \qquad \forall i \in \mathcal{I}_m.$$

Define the maps $\varsigma\colon \mathcal{I}_{2m} \to \mathcal{I}_m$ and $\gamma\colon \mathcal{I}_{2m} \to \{-1, +1\}$ by

$$\varsigma(i) = ((i-1) \bmod m) + 1, \qquad \gamma(j) = \begin{cases} +1, & \text{if } j \in \mathcal{I}_m, \\ -1, & \text{otherwise.} \end{cases}$$

It can be seen that ς "wraps" any index in $\mathcal{I}_{2m}$ to an index in $\mathcal{I}_m$, γ is a "sign" function in

[7] See also Section 3.6.

[8] For clarity, we use the symbol $\vec{0}_{\tilde{q}-m}$ to denote the $(\tilde{q}-m)$ dimensional zero vector, only in this section.

some sense, and that $\nabla\tilde{h}_i = \gamma(i)\tilde{c}_{\varsigma(i)}$ for any $i \in \mathcal{I}_{2m}$. With these maps, for any $\mathcal{I} \subset \mathcal{I}_{2m}$ such that $|\mathcal{I}| = s > 0$, and any $i \in \mathcal{I}_s$, we have

$$\nabla\tilde{h}_{\sigma_{\mathcal{I}}(i)} = \gamma(\sigma_{\mathcal{I}}(i))\tilde{c}_{\varsigma(\sigma_{\mathcal{I}}(i))} = \gamma(\sigma_{\mathcal{I}}(i))u_{\max,\varsigma(\sigma_{\mathcal{I}}(i))}[\vec{0}^{\mathrm{T}}_{\tilde{q}-m}, e^{\mathrm{T}}_{\varsigma(\sigma_{\mathcal{I}}(i))}]^{\mathrm{T}},$$

so that the matrix $\tilde{N}_{\mathcal{I}}$ corresponding to (5.21) can be written as

$$\tilde{N}_{\mathcal{I}} = \begin{cases} [\nabla\tilde{h}_{\sigma_{\mathcal{I}}(1)}, \nabla\tilde{h}_{\sigma_{\mathcal{I}}(2)}, \ldots, \nabla\tilde{h}_{\sigma_{\mathcal{I}}(s)}] = \Omega_{\mathcal{I}}\Lambda_{\mathcal{I}}, & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases} \tag{5.36}$$

where[9]

$$\Omega_{\mathcal{I}} = \begin{bmatrix} & 0_{(\tilde{q}-m)\times s} & & \\ e_{\varsigma(\sigma_{\mathcal{I}}(1))} & e_{\varsigma(\sigma_{\mathcal{I}}(2))} & \cdots & e_{\varsigma(\sigma_{\mathcal{I}}(s))} \end{bmatrix} =\in \mathbb{R}^{\tilde{q}\times s},$$
$$\Lambda_{\mathcal{I}} = \mathrm{diag}(\gamma(\sigma_{\mathcal{I}}(1))u_{\max,\varsigma(\sigma_{\mathcal{I}}(1))}, \ldots, \gamma(\sigma_{\mathcal{I}}(s))u_{\max,\varsigma(\sigma_{\mathcal{I}}(s))}) \in \mathbb{R}^{s\times s},$$

and $0_{(\tilde{q}-m)\times s}$ is the $(\tilde{q}-m)\times s$ zero matrix.

Remark 5.4. Observe that $\Lambda_{\mathcal{I}}$ must be nonsingular because $u_{\max,i} \neq 0$ for all $i \in \mathcal{I}_m$. □

For the GPAW parameter $\tilde{\Gamma}$, we choose it to be of the form $\tilde{\Gamma} = \begin{bmatrix} \tilde{\Gamma}_1 & 0 \\ 0 & I_m \end{bmatrix}$, where $\tilde{\Gamma}_1 \in \mathbb{R}^{(\tilde{q}-m)\times(\tilde{q}-m)}$ is symmetric positive definite. It can be verified that $\tilde{\Gamma}\Omega_{\mathcal{I}} = \Omega_{\mathcal{I}}$ holds for this choice of $\tilde{\Gamma}$. Clearly the standard basis vectors satisfy $e_i^{\mathrm{T}}e_j = \delta_{ij}$ and $\tilde{e}_i^{\mathrm{T}}\tilde{e}_j = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta function [125, p. 544]. Then when $|\mathcal{I}| = s > 0$ and $\tilde{N}_{\mathcal{I}} = \Omega_{\mathcal{I}}\Lambda_{\mathcal{I}}$ is full rank, it can be verified that $\Omega_{\mathcal{I}}^{\mathrm{T}}\tilde{\Gamma}\Omega_{\mathcal{I}} = \Omega_{\mathcal{I}}^{\mathrm{T}}\Omega_{\mathcal{I}} = I_s$, so that the projection matrix corresponding to (5.22) can be written as

$$\begin{aligned}
\tilde{R}_{\mathcal{I}} &= I_{\tilde{q}} - \tilde{\Gamma}\tilde{N}_{\mathcal{I}}(\tilde{N}_{\mathcal{I}}^{\mathrm{T}}\tilde{\Gamma}\tilde{N}_{\mathcal{I}})^{-1}\tilde{N}_{\mathcal{I}}^{\mathrm{T}} = I_{\tilde{q}} - \tilde{\Gamma}\Omega_{\mathcal{I}}\Lambda_{\mathcal{I}}(\Lambda_{\mathcal{I}}^{\mathrm{T}}\Omega_{\mathcal{I}}^{\mathrm{T}}\tilde{\Gamma}\Omega_{\mathcal{I}}\Lambda_{\mathcal{I}})^{-1}\Lambda_{\mathcal{I}}^{\mathrm{T}}\Omega_{\mathcal{I}}^{\mathrm{T}}, \\
&= I_{\tilde{q}} - \Omega_{\mathcal{I}}(\Omega_{\mathcal{I}}^{\mathrm{T}}\Omega_{\mathcal{I}})^{-1}\Omega_{\mathcal{I}}^{\mathrm{T}} = I_{\tilde{q}} - \Omega_{\mathcal{I}}\Omega_{\mathcal{I}}^{\mathrm{T}} = I_{\tilde{q}} - \sum_{i=1}^{s} \tilde{e}_{\tilde{q}-m+\varsigma(\sigma_{\mathcal{I}}(i))}\tilde{e}^{\mathrm{T}}_{\tilde{q}-m+\varsigma(\sigma_{\mathcal{I}}(i))}, \\
&= I_{\tilde{q}} - \sum_{j\in\mathcal{I}} \tilde{e}_{\tilde{q}-m+\varsigma(j)}\tilde{e}^{\mathrm{T}}_{\tilde{q}-m+\varsigma(j)} = I_{\tilde{q}} - \sum_{j\in\mathcal{I}} \begin{bmatrix} 0 & 0 \\ 0 & e_{\varsigma(j)}e^{\mathrm{T}}_{\varsigma(j)} \end{bmatrix} = \begin{bmatrix} I_{\tilde{q}-m} & 0 \\ 0 & I_{\mathcal{I}} \end{bmatrix},
\end{aligned} \tag{5.37}$$

where

$$I_{\mathcal{I}} = I_m - \sum_{i\in\mathcal{I}} e_{\varsigma(i)}e^{\mathrm{T}}_{\varsigma(i)}. \tag{5.38}$$

Notice that the definition of $\tilde{R}_{\mathcal{I}}$ in (5.37) with (5.38) is also valid when $\mathcal{I} = \emptyset$ since in this case, the summation in (5.38) will be over an empty set.

Next, we express the GPAW-compensated system (with the GPAW-compensated controller (5.34) derived from the transformed nominal controller (5.32)) in a form closer to the description of the linear system with partial state constraints (5.25). To do this, partition the matrices A_{cc} and B_{cc} in (5.27) appropriately as $A_{cc} = \begin{bmatrix} A_{cc11} & A_{cc12} \\ A_{cc21} & A_{cc22} \end{bmatrix}$ and $B_{cc} = \begin{bmatrix} B_{cc1} \\ B_{cc2} \end{bmatrix}$ respectively. Then, using the definition of $T = \begin{bmatrix} I_{\tilde{q}-m} & 0 \\ 0 & \Lambda \end{bmatrix}$, it can be verified that (see (5.33))

$$\tilde{A}_c = \begin{bmatrix} A_{cc11} & A_{cc12}\Lambda^{-1} \\ \Lambda A_{cc21} & \Lambda A_{cc22}\Lambda^{-1} \end{bmatrix}, \qquad \tilde{B}_c = \begin{bmatrix} B_{cc1} \\ \Lambda B_{cc2} \end{bmatrix}, \qquad \tilde{C}_c = \begin{bmatrix} 0 & \Lambda^{-1} \end{bmatrix}. \tag{5.39}$$

[9] For clarity, we use the symbol $0_{i\times j}$ to denote the $i \times j$ zero matrix, only in this section.

Using (5.37), the GPAW-compensated system (5.1), (5.34), with $u = u_g$, is described by

$$\begin{aligned}
\dot{x} &= Ax + B\,\mathrm{sat}(\Lambda^{-1}\phi),\\
\dot{\tilde{x}}_{gf} &= B_{cc1}Cx + A_{cc11}\tilde{x}_{gf} + A_{cc12}\Lambda^{-1}\phi + B_{cc1}D\,\mathrm{sat}(\Lambda^{-1}\phi),\\
\dot{\phi} &= I_{\mathcal{I}^*}(\Lambda B_{cc2}Cx + \Lambda A_{cc21}\tilde{x}_{gf} + \Lambda A_{cc22}\Lambda^{-1}\phi + \Lambda B_{cc2}D\,\mathrm{sat}(\Lambda^{-1}\phi)),
\end{aligned}$$

where $I_{\mathcal{I}^*}$ in the last equation is defined by (5.38) and a solution $\mathcal{I}^*$ to subproblem (5.23). By initializing the partial controller state ϕ in the unit hypercube, i.e. $\phi(0) \in \mathbb{D}^m$, we have $\mathrm{sat}(\Lambda^{-1}\phi(0)) = \Lambda^{-1}\phi(0)$. Then Theorem 2.5.3 yields $\mathrm{sat}(\Lambda^{-1}\phi(t)) = \Lambda^{-1}\phi(t)$ for all $t \geq 0$, which implies $\phi(t) \in \mathbb{D}^m$ and $|\phi_i(t)| \leq 1$ for all $t \geq 0$ and all $i \in \mathcal{I}_m$ (see (5.24)). Defining $n_\theta := n + \tilde{q} - m$ and $\theta := (x, \tilde{x}_{gf}) \in \mathbb{R}^{n_\theta}$, the preceding system can be written as

$$\begin{aligned}
\dot{\theta} &= A_g\theta + B_g\phi, &\qquad \theta(0) &= \theta_0,\\
\dot{\phi} &= I_{\mathcal{I}^*}(C_g\theta + D_g\phi), &\qquad \phi(0) &= \phi_0,
\end{aligned} \tag{5.40}$$

where $\theta_0 = (x(0), T_1x_{cc}(0))$ and $\phi_0 = T_2x_{cc}(0)$ are the initial states, $T_1 = \begin{bmatrix} I_{\tilde{q}-m} & 0 \end{bmatrix}$ and $T_2 = \begin{bmatrix} 0 & \Lambda \end{bmatrix}$ are the top and bottom partitions of the transformation matrix $T = \left[\begin{smallmatrix} T_1 \\ T_2 \end{smallmatrix}\right]$, and

$$\begin{aligned}
A_g &= \begin{bmatrix} A & 0 \\ B_{cc1}C & A_{cc11} \end{bmatrix}, &\qquad B_g &= \begin{bmatrix} B \\ A_{cc12} + B_{cc1}D \end{bmatrix}\Lambda^{-1},\\
C_g &= \Lambda\begin{bmatrix} B_{cc2}C & A_{cc21} \end{bmatrix}, &\qquad D_g &= \Lambda(A_{cc22} + B_{cc2}D)\Lambda^{-1}.
\end{aligned}$$

It can be seen that system (5.40) has similarities with the linear system with partial state constraints (5.25). In fact, by comparing (5.40) with (5.25), we see that they are *equivalent* if and only if $I_{\mathcal{I}^*}(C_g\theta + D_g\phi) \equiv h(C_g\theta + D_g\phi)$ holds, or equivalently, if and only if

$$e_i^{\mathrm{T}} I_{\mathcal{I}^*}(C_g\theta + D_g\phi) = e_i^{\mathrm{T}} h(C_g\theta + D_g\phi), \qquad \forall i \in \mathcal{I}_m, \forall(\theta, \phi) \in \mathbb{R}^{n_\theta} \times \mathbb{D}^m, \tag{5.41}$$

holds, i.e. the individual elements are equivalent. We show this next.

Using the definition of $I_{\mathcal{I}}$ (see (5.38)), the left-hand-side of the equality in (5.41) is[10]

$$\begin{aligned}
e_i^{\mathrm{T}} I_{\mathcal{I}^*}(C_g\theta + D_g\phi) &= e_i^{\mathrm{T}}\Big(I_m - \sum_{j\in\mathcal{I}^*} e_{\varsigma(j)}e_{\varsigma(j)}^{\mathrm{T}}\Big)(C_g\theta + D_g\phi),\\
&= \Big(e_i^{\mathrm{T}} - \sum_{j\in\mathcal{I}^*} e_i^{\mathrm{T}} e_{\varsigma(j)}e_{\varsigma(j)}^{\mathrm{T}}\Big)(C_g\theta + D_g\phi),\\
&= \begin{cases} 0, & \text{if } (i \in \mathcal{I}^*) \vee ((i+m) \in \mathcal{I}^*),\\ e_i^{\mathrm{T}}(C_g\theta + D_g\phi), & \text{otherwise,} \end{cases}
\end{aligned} \tag{5.42}$$

where we have used $e_i^{\mathrm{T}} e_{\varsigma(j)} = \delta_{i\varsigma(j)}$ with δ_{ij} being the Kronecker delta function [125, p. 544]. From the definitions of h and $\tilde{h}$ in (5.26), the right-hand-side of the equality in (5.41) is

$$\begin{aligned}
e_i^{\mathrm{T}} h(C_g\theta + D_g\phi) &= \tilde{h}(e_i^{\mathrm{T}}(C_g\theta + D_g\phi), \phi_i),\\
&= \begin{cases} 0, & \text{if } (|\phi_i| = 1) \wedge (\phi_i e_i^{\mathrm{T}}(C_g\theta + D_g\phi) \geq 0),\\ e_i^{\mathrm{T}}(C_g\theta + D_g\phi), & \text{otherwise.} \end{cases}
\end{aligned} \tag{5.43}$$

[10] Note that $\vee$ denotes the logical OR operator.

To show that (5.41) holds, we need to show that (5.42) is equal to (5.43) for all $i \in \mathcal{I}_m$ and all $(\theta, \phi) \in \mathbb{R}^{n_\theta} \times \mathbb{D}^m$. Fix any $(\theta, \phi) \in \mathbb{R}^{n_\theta} \times \mathbb{D}^m$. For any $i \in \mathcal{I}_m$ such that $e_i^{\mathrm{T}}(C_g\theta + D_g\phi) = 0$, we see from (5.42) and (5.43) that the function values must be both zero regardless of the associated logical conditions, and hence the equality in (5.41) holds. It remains to show that the equality in (5.41) holds for all $i \in \mathcal{I}_m$ such that $e_i^{\mathrm{T}}(C_g\theta + D_g\phi) \neq 0$. From (5.42) and (5.43), we see that the equality in (5.41) holds when[11]

$$((|\phi_i| = 1) \wedge (\phi_i e_i^{\mathrm{T}}(C_g\theta + D_g\phi) > 0)) \Leftrightarrow ((i \in \mathcal{I}^*) \vee ((i+m) \in \mathcal{I}^*)). \tag{5.44}$$

We will analyze subproblem (5.23) to show that (5.44) holds for all $i \in \mathcal{I}_m$. Since (θ, ϕ) is arbitrary, this implies that (5.41) holds, which in turn implies the equivalence of systems (5.40) and (5.25).

First, partition the matrices A_g and B_g in (5.40) as $A_g = \begin{bmatrix} A_{g1} \\ A_{g2} \end{bmatrix}$ and $B_g = \begin{bmatrix} B_{g1} \\ B_{g2} \end{bmatrix}$ where $A_{g2} = \begin{bmatrix} B_{cc1}C & A_{cc11} \end{bmatrix}$ and $B_{g2} = (A_{cc12} + B_{cc1}D)\Lambda^{-1}$. Using (5.39), (5.1), $\tilde{x}_g = (\tilde{x}_{gf}, \phi)$, and $\theta = (x, \tilde{x}_{gf})$, it can be verified that

$$\begin{aligned}
\tilde{A}_c\tilde{x}_g + \tilde{B}_c y &= \begin{bmatrix} A_{cc11} & A_{cc12}\Lambda^{-1} \\ \Lambda A_{cc21} & \Lambda A_{cc22}\Lambda^{-1} \end{bmatrix} \begin{bmatrix} \tilde{x}_{gf} \\ \phi \end{bmatrix} + \begin{bmatrix} B_{cc1} \\ \Lambda B_{cc2} \end{bmatrix} (Cx + D\Lambda^{-1}\phi), \\
&= \begin{bmatrix} B_{cc1}Cx + A_{cc11}\tilde{x}_{gf} + (A_{cc12} + B_{cc1}D)\Lambda^{-1}\phi \\ \Lambda B_{cc2}Cx + \Lambda A_{cc21}\tilde{x}_{gf} + \Lambda(A_{cc22} + B_{cc2}D)\Lambda^{-1}\phi \end{bmatrix}, \\
&= \begin{bmatrix} A_{g2}\theta + B_{g2}\phi \\ C_g\theta + D_g\phi \end{bmatrix}.
\end{aligned} \tag{5.45}$$

Using (5.37), the objective function of subproblem (5.23) can be written as

$$\tilde{F}(\mathcal{I}) = (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}}\tilde{\Gamma}^{-1}\tilde{R}_{\mathcal{I}}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) = (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}}\tilde{\Gamma}^{-1}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) - \tilde{F}_1(\mathcal{I}),$$

where

$$\begin{aligned}
\tilde{F}_1(\mathcal{I}) &= \sum_{j \in \mathcal{I}} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}}\tilde{\Gamma}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & e_{\varsigma(j)}e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y), \\
&= \sum_{j \in \mathcal{I}} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}} \begin{bmatrix} \tilde{\Gamma}_1^{-1} & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & e_{\varsigma(j)}e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y), \\
&= \sum_{j \in \mathcal{I}} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y)^{\mathrm{T}} \begin{bmatrix} 0 & 0 \\ 0 & e_{\varsigma(j)}e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) = \sum_{j \in \mathcal{I}} \left| \begin{bmatrix} 0 & e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix} (\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) \right|^2, \\
&= \sum_{j \in \mathcal{I}} \left| \begin{bmatrix} 0 & e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A_{g2}\theta + B_{g2}\phi \\ C_g\theta + D_g\phi \end{bmatrix} \right|^2 = \sum_{j \in \mathcal{I}} \left| e_{\varsigma(j)}^{\mathrm{T}}(C_g\theta + D_g\phi) \right|^2,
\end{aligned}$$

and we have used $\begin{bmatrix} 0 \\ e_{\varsigma(j)} \end{bmatrix} \begin{bmatrix} 0 & e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & e_{\varsigma(j)}e_{\varsigma(j)}^{\mathrm{T}} \end{bmatrix}$ and (5.45). Since $\tilde{F}_1(\mathcal{I})$ is the only component of $\tilde{F}(\mathcal{I})$ that varies with $\mathcal{I}$, it is clear that maximizing $\tilde{F}(\mathcal{I})$ is equivalent to *minimizing* $\tilde{F}_1(\mathcal{I})$.

Next, we show that the rank condition of subproblem (5.23) always holds. Observe from (5.35) that the constraints $\tilde{h}_i(\tilde{x}_g) \leq 0$ and $\tilde{h}_{i+m}(\tilde{x}_g) \leq 0$ cannot be simultaneously

[11] Observe that when $\phi_i \neq 0$ and $e_i^{\mathrm{T}}(C_g\theta + D_g\phi) \neq 0$, the condition $\phi_i e_i^{\mathrm{T}}(C_g\theta + D_g\phi) \geq 0$ implies $\phi_i e_i^{\mathrm{T}}(C_g\theta + D_g\phi) > 0$ holds with *strict* inequality.

active, i.e. $(\tilde{h}_i(\tilde{x}_g) \geq 0) \Rightarrow (\tilde{h}_{i+m}(\tilde{x}_g) < 0)$ and $(\tilde{h}_{i+m}(\tilde{x}_g) \geq 0) \Rightarrow (\tilde{h}_i(\tilde{x}_g) < 0)$. In other words, the constraints indexed by $i \in \mathcal{I}_m$ and $i + m \in (\mathcal{I}_{2m} \setminus \mathcal{I}_m)$ cannot be simultaneously active. Hence for any $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$, $\mathcal{I} \neq \emptyset$, the matrix $\Omega_{\mathcal{I}} \in \mathbb{R}^{\tilde{q} \times |\mathcal{I}|}$ (see (5.36)) must be composed of linearly independent standard basis vectors and hence full rank, i.e. $\text{rank}(\Omega_{\mathcal{I}}) = \min\{\tilde{q}, |\mathcal{I}|\} = |\mathcal{I}|$. Since $\Lambda_{\mathcal{I}}$ is nonsingular for any $\mathcal{I} \neq \emptyset$ (see Remark 5.4), we have from (5.36) that $\text{rank}(\tilde{N}_{\mathcal{I}}) = \text{rank}(\Omega_{\mathcal{I}}\Lambda_{\mathcal{I}}) = \text{rank}(\Omega_{\mathcal{I}}) = |\mathcal{I}|$ [124, p. 13]. When $\mathcal{I} = \emptyset \subset \mathcal{I}_{\text{sat}}$, it is clear that $\text{rank}(\tilde{N}_{\mathcal{I}}) = 0 = |\mathcal{I}|$, so that the rank condition of subproblem (5.23) always holds.

When $|\mathcal{I}_{\text{sat}}| = s > 0$, the remaining constraint function of subproblem (5.23) reduces to

$$\begin{aligned}
\tilde{N}_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \tilde{R}_{\mathcal{I}}(\tilde{A}_c\tilde{x}_g + \tilde{B}_c y) &= \Lambda_{\mathcal{I}_{\text{sat}}} \Omega_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} \begin{bmatrix} I_{\tilde{q}-m} & 0 \\ 0 & I_{\mathcal{I}} \end{bmatrix} \begin{bmatrix} A_{g2}\theta + B_{g2}\phi \\ C_g\theta + D_g\phi \end{bmatrix}, \\
&= \Lambda_{\mathcal{I}_{\text{sat}}} \begin{bmatrix} 0_{s\times(\tilde{q}-m)} & \begin{matrix} e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(1))}^{\mathrm{T}} \\ \vdots \\ e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(s))}^{\mathrm{T}} \end{matrix} \end{bmatrix} \begin{bmatrix} I_{\tilde{q}-m} & 0 \\ 0 & I_{\mathcal{I}} \end{bmatrix} \begin{bmatrix} A_{g2}\theta + B_{g2}\phi \\ C_g\theta + D_g\phi \end{bmatrix}, \\
&= \Lambda_{\mathcal{I}_{\text{sat}}} \begin{bmatrix} 0_{s\times(\tilde{q}-m)} & \begin{matrix} e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(1))}^{\mathrm{T}} I_{\mathcal{I}} \\ \vdots \\ e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(s))}^{\mathrm{T}} I_{\mathcal{I}} \end{matrix} \end{bmatrix} \begin{bmatrix} A_{g2}\theta + B_{g2}\phi \\ C_g\theta + D_g\phi \end{bmatrix}, \\
&= \Lambda_{\mathcal{I}_{\text{sat}}} [e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(1))}, \ldots, e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(s))}]^{\mathrm{T}} I_{\mathcal{I}}(C_g\theta + D_g\phi),
\end{aligned}$$

where we have used (5.36), (5.37), and (5.45). Then subproblem (5.23) is equivalent to[12]

$$\begin{aligned}
&\min_{\mathcal{I} \in \mathcal{J}} \tilde{F}_1(\mathcal{I}) = \sum_{j \in \mathcal{I}} \left| e_{\varsigma(j)}^{\mathrm{T}}(C_g\theta + D_g\phi) \right|^2, \\
&\text{subject to} \quad \Lambda_{\mathcal{I}_{\text{sat}}} [e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(1))}, \ldots, e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(s))}]^{\mathrm{T}} I_{\mathcal{I}}(C_g\theta + D_g\phi) \leq 0,
\end{aligned} \tag{5.46}$$

when $|\mathcal{I}_{\text{sat}}| = s > 0$. Observe from the objective function of subproblem (5.46) that $\tilde{F}_1(\mathcal{I}) \geq 0$ for all $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$, $\tilde{F}_1(\emptyset) = 0$, and $\tilde{F}_1(\mathcal{I})$ will be minimized when $\mathcal{I}$ is of least cardinality such that the associated constraints are satisfied (see also Proposition 2.5.2). Since for every $j \in \mathcal{I}_s$, there is a unique $i \in \mathcal{I}_{\text{sat}}$ such that $\sigma_{\mathcal{I}_{\text{sat}}}(j) = i$ (see Remark 2.5), the j-th component of the constraint function in subproblem (5.46) can be rewritten as

$$\begin{aligned}
\beta(i, \mathcal{I}) &:= e_j^{\mathrm{T}} \Lambda_{\mathcal{I}_{\text{sat}}} [e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(1))}, \ldots, e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(s))}]^{\mathrm{T}} I_{\mathcal{I}}(C_g\theta + D_g\phi), \\
&= \gamma(\sigma_{\mathcal{I}_{\text{sat}}}(j)) u_{\max,\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(j))} e_{\varsigma(\sigma_{\mathcal{I}_{\text{sat}}}(j))}^{\mathrm{T}} I_{\mathcal{I}}(C_g\theta + D_g\phi), \\
&= \gamma(i) u_{\max,\varsigma(i)} e_{\varsigma(i)}^{\mathrm{T}} \Big(I_m - \sum_{k \in \mathcal{I}} e_{\varsigma(k)} e_{\varsigma(k)}^{\mathrm{T}} \Big)(C_g\theta + D_g\phi), \\
&= \gamma(i) u_{\max,\varsigma(i)} \Big(e_{\varsigma(i)}^{\mathrm{T}} - \sum_{k \in \mathcal{I}} e_{\varsigma(i)}^{\mathrm{T}} e_{\varsigma(k)} e_{\varsigma(k)}^{\mathrm{T}} \Big)(C_g\theta + D_g\phi), \\
&= \begin{cases} 0, & \text{if } \exists k \in \mathcal{I}, \varsigma(k) = \varsigma(i), \\ \gamma(i) u_{\max,\varsigma(i)} e_{\varsigma(i)}^{\mathrm{T}}(C_g\theta + D_g\phi), & \text{otherwise}, \end{cases}
\end{aligned} \tag{5.47}$$

where we have used the definition of $I_{\mathcal{I}}$ in (5.38) and $\Lambda_{\mathcal{I}}$ in (5.36). Now, recall that

[12] We will not need the equivalent of subproblem (5.46) when $|\mathcal{I}_{\text{sat}}| = 0$, i.e. $\mathcal{I}_{\text{sat}} = \emptyset$.

constraints indexed by $j \in \mathcal{I}_m$ and $j + m \in (\mathcal{I}_{2m} \setminus \mathcal{I}_m)$ cannot be simultaneously active. Since $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$ and $i \in \mathcal{I}_{\text{sat}}$, the condition $(\exists k \in \mathcal{I}, \varsigma(k) = \varsigma(i))$ in (5.47) can be verified to be equivalent to $(\exists k \in \mathcal{I}, k = i)$, and hence equivalent to $i \in \mathcal{I}$. Then the constraint $\beta(i, \mathcal{I}) \leq 0$ holds when $i \in \mathcal{I}$ or $\gamma(i) e_{\varsigma(i)}^{\mathrm{T}} (C_g \theta + D_g \phi) \leq 0$. Since the constraint must hold for the optimal solution $\mathcal{I}^*$, i.e. $\beta(i, \mathcal{I}^*) \leq 0$ for all $i \in \mathcal{I}_{\text{sat}}$, we have

$$i \in \mathcal{I}^*, \qquad \text{or} \qquad \gamma(i) e_{\varsigma(i)}^{\mathrm{T}} (C_g \theta + D_g \phi) \leq 0, \qquad \forall i \in \mathcal{I}_{\text{sat}}. \tag{5.48}$$

Recall that we want to establish the equivalence (5.44) and we only need to consider those $i \in \mathcal{I}_m$ for which $e_i^{\mathrm{T}} (C_g \theta + D_g \phi) \neq 0$. First, assume that $|\phi_i| = 1$ and $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) > 0$ for some $i \in \mathcal{I}_m$. This corresponds to two cases:

(i) $\phi_i = 1$ and $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) > 0$; or
(ii) $\phi_i = -1$ and $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) > 0$.

In case (i), we have $i \in \mathcal{I}_m \cap \mathcal{I}_{\text{sat}}$ (see (5.35)) and

$$\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) = e_i^{\mathrm{T}} (C_g \theta + D_g \phi) = \gamma(i) e_{\varsigma(i)}^{\mathrm{T}} (C_g \theta + D_g \phi) > 0.$$

Then $i \in \mathcal{I}^*$ follows from (5.48). In case (ii), we have $(i + m) \in (\mathcal{I}_{2m} \setminus \mathcal{I}_m) \cap \mathcal{I}_{\text{sat}}$ and

$$\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) = -e_i^{\mathrm{T}} (C_g \theta + D_g \phi) = \gamma(i + m) e_{\varsigma(i+m)}^{\mathrm{T}} (C_g \theta + D_g \phi) > 0.$$

Then $i + m \in \mathcal{I}^*$ follows from (5.48). Together, these establish one direction of the equivalence (5.44), namely

$$((|\phi_i| = 1) \wedge (\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) > 0)) \Rightarrow ((i \in \mathcal{I}^*) \vee ((i + m) \in \mathcal{I}^*)).$$

It remains to show the converse.

We will show that if $|\phi_i| < 1$ or $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) \leq 0$, then $i, (i + m) \notin \mathcal{I}^*$. When $|\phi_i| < 1$, we have $\mathcal{I}_{\text{sat}} = \emptyset$, so that $i, (i+m) \notin \mathcal{I}^* \subset \mathcal{I}_{\text{sat}} = \emptyset$. Finally, consider when $|\phi_i| = 1$ and $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) \leq 0$. Then $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) < 0$ holds with *strict* inequality due to $e_i^{\mathrm{T}} (C_g \theta + D_g \phi) \neq 0$. As before, this corresponds to two cases:

(iii) $\phi_i = 1$ and $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) < 0$; or
(iv) $\phi_i = -1$ and $\phi_i e_i^{\mathrm{T}} (C_g \theta + D_g \phi) < 0$.

In case (iii), we have $i \in \mathcal{I}_m \cap \mathcal{I}_{\text{sat}}$ and $\gamma(i) e_{\varsigma(i)}^{\mathrm{T}} (C_g \theta + D_g \phi) < 0$, which implies $(i+m) \notin \mathcal{I}_{\text{sat}}$, $(i + m) \notin \mathcal{I}^* \subset \mathcal{I}_{\text{sat}}$, and $e_{\varsigma(i)}^{\mathrm{T}} (C_g \theta + D_g \phi) \neq 0$. From (5.47), it can be seen that $\beta(i, \mathcal{I}) \leq 0$ for any $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$. Due to $e_{\varsigma(i)}^{\mathrm{T}} (C_g \theta + D_g \phi) \neq 0$, we must have $i \notin \mathcal{I}^*$ to minimize the objective function of subproblem (5.46), and hence $i, (i + m) \notin \mathcal{I}^*$. Finally, in case (iv), we have $(i + m) \in (\mathcal{I}_{2m} \setminus \mathcal{I}_m) \cap \mathcal{I}_{\text{sat}}$ and $\gamma(i + m) e_{\varsigma(i+m)}^{\mathrm{T}} (C_g \theta + D_g \phi) < 0$, which implies $i \notin \mathcal{I}_{\text{sat}}$, $i \notin \mathcal{I}^* \subset \mathcal{I}_{\text{sat}}$, and $e_{\varsigma(i+m)}^{\mathrm{T}} (C_g \theta + D_g \phi) \neq 0$. The same argument then shows that $(i + m) \notin \mathcal{I}^*$ and hence $i, (i + m) \notin \mathcal{I}^*$. These establish the equivalence (5.44) and show that the GPAW-compensated system (5.40) is indeed equivalent to the linear system with partial state constraint (5.25).

System (5.40) is obtained by GPAW compensation with parameter $\tilde{\Gamma}$ on the transformed nominal controller (5.32). Using Proposition 5.3.1, we can recover the GPAW parameter in the original coordinates as $\Gamma = T^{-1} \tilde{\Gamma} T^{-\mathrm{T}} = (T^{\mathrm{T}} \tilde{\Gamma}^{-1} T)^{-1}$. We summarize this result when C_c has full row rank below, which uses relations in (5.32) and Section 5.5.1.

Theorem 5.6.1 (GPAW-Compensated LTI System as Linear System with Partial State Constraints)**.** *If* C_c *in* (5.2) *satisfies* $\operatorname{rank}(C_c) = m$*, then GPAW compensation applied to the nominal controller* (5.2) *yields a closed-loop system* (5.9) *that can be transformed to a linear system with partial state constraints* (5.25) *when the GPAW parameter* Γ *has the form*

$$\Gamma = T^{-1}\tilde{\Gamma}T^{-\mathrm{T}} = (\tilde{T}^{\mathrm{T}}\tilde{\Gamma}_1^{-1}\tilde{T} + C_c^{\mathrm{T}}\Lambda^2 C_c)^{-1},$$

where

$$T = \begin{bmatrix} \tilde{T} \\ \Lambda C_c \end{bmatrix} \in \mathbb{R}^{q\times q}, \qquad \tilde{\Gamma} = \begin{bmatrix} \tilde{\Gamma}_1 & 0 \\ 0 & I_m \end{bmatrix} \in \mathbb{R}^{q\times q},$$
$$\Lambda = \operatorname{diag}(u_{\max,1}^{-1}, u_{\max,2}^{-1}, \ldots, u_{\max,m}^{-1}) \in \mathbb{R}^{m\times m},$$

$\tilde{T} \in \mathbb{R}^{(q-m)\times q}$ *is chosen such that* T *is nonsingular, and* $\tilde{\Gamma}_1 \in \mathbb{R}^{(q-m)\times(q-m)}$ *is symmetric positive definite. Moreover, the transformation is defined by* $\tilde{x}_g = Tx_g$*, where* $\tilde{x}_g$ *and* x_g *are the states of the transformed and original GPAW-compensated controllers respectively.*

Remark 5.5. When C_c is row rank deficient, an analogous form of Theorem 5.6.1 where the GPAW-compensated controller is derived from the augmented nominal controller (5.31) can be stated, but is omitted for brevity. □

For this class of GPAW-compensated systems, it may be more efficient to implement the associated GPAW-compensated controller as the transformed linear system with partial state constraints (5.25), since these are closed-form expressions and no optimization problem needs to be solved online (see also Section B.1 in Appendix B).

5.6.1 Illustration of an Existing Result

With Theorem 5.6.1, results from [103–106] can be applied to this class of GPAW-compensated systems. As an illustration, we state the following which is a direct translation of a stability result in [103].

Theorem 5.6.2 (Hou et al. [103, Theorem 3])**.** *Consider the transformed GPAW-compensated system* (5.40)*. If the matrix* $\left[\begin{smallmatrix} A_g & B_g \\ C_g & D_g \end{smallmatrix}\right]$ *is Hurwitz stable and there exist symmetric positive definite matrices* $P_1 \in \mathbb{R}^{n_\theta\times n_\theta}$*,* $P_2 \in \mathbb{R}^{m\times m}$*, and* $Q \in \mathbb{R}^{(n_\theta+m)\times(n_\theta+m)}$ *such that* $P_2 = [p_{ij}]$ *satisfies*

$$p_{ii} \geq \sum_{j=1, j\neq i}^{m} |p_{ji}|, \qquad \forall i \in \mathcal{I}_m,$$

and

$$\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}\begin{bmatrix} A_g & B_g \\ C_g & D_g \end{bmatrix} + \begin{bmatrix} A_g & B_g \\ C_g & D_g \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} = -Q,$$

then the region of attraction of the origin of system (5.40) *contains* $\mathbb{R}^{n_\theta}\times\mathbb{D}^m$*, i.e. the origin is* globally asymptotically stable *in the sense of [103].*

Proof. See [103]. ■

Under the similarity transformation defined by $\tilde{x}_g = Tx_g$ for some nonsingular $T \in \mathbb{R}^{\tilde{q}\times\tilde{q}}$, the corresponding unsaturated region in the original coordinates is $X := \mathbb{R}^n \times T^{-1}(\mathbb{R}^{\tilde{q}-m}\times\mathbb{D}^m)$. Hence in the original coordinates, the region of attraction of the origin for the GPAW-compensated system (5.9) must contain X. See [104–106] for other applicable results.

5.7 Chapter Summary

In this chapter, we restrict consideration to regulatory GPAW-compensated systems comprising input-constrained MIMO LTI plants driven by MIMO LTI controllers. A stability result that is specialized from the nonlinear case was presented, which is readily verified by solving a system of linear matrix inequalities. Even though global asymptotic stability is assured, numerical experience suggests this result to be conservative.

Similarity transformations are fundamental operations for LTI systems. Even though the GPAW-compensated controller is defined by the online solution to an optimization problem, similarity transformations can still be performed without much difficulty. We showed how to transform the nominal LTI controller into a canonical form that is more convenient for GPAW compensation. This canonical form is then used to show that under some non-unique choice of the GPAW parameter, the GPAW-compensated system can be transformed into a *linear system with partial state constraints*, a topic that has been previously studied. This allows results in existing literature to be applied to this class of GPAW-compensated systems, and vice versa.

Chapter 6

Numerical Comparisons

In this chapter, we compare the GPAW scheme against the nonlinear anti-windup scheme for Euler-Lagrange systems proposed in [24], the anti-windup scheme for feedback linearizable nonlinear systems proposed in [65], and the LMI-based anti-windup scheme for stable LTI systems proposed in [128]. As our stability results obtained thus far are somewhat conservative, we will not establish analytical stability of the GPAW-compensated systems. These numerical comparisons only show that GPAW compensation can yield *numerically* comparable performance for the cases studied. Note also that the stability results of [24, 65, 128] are of the "absolute" sense, the drawback of which was discussed in Section 3.7.

6.1 Nonlinear Anti-windup Scheme

The nonlinear anti-windup scheme of [24] was briefly described in Section 1.4.4. As mentioned, the method was first proposed in [74] and extended to Euler-Lagrange systems in [24], where two simulation studies were presented. Here, we compare the GPAW scheme against the nonlinear anti-windup scheme using one of the two simulation studies, which is on a double integrator plant driven by a PID controller. The main aspect to note from this study is the effect of controller state initialization (see the statement of Problem 1 in Section 1.3).

Remark 6.1. Attempts have been made to compare the GPAW scheme against the nonlinear anti-windup scheme using the four-link robot example in [24]. One difficulty is that the expressions for the generalized inertia matrix and Coriolis forces/torques were not provided. Clarifications with Prof. Luca Zaccarian, co-author of [24], indicates that the expressions are available in [184, pp. 105 – 106]. However, even using these expressions, we have been unable to reproduce the results in [24] for the robot example. □

The double integrator plant is described by

$$\begin{aligned} \dot{x} &= Ax + B\,\mathrm{sat}(u), \qquad x(0) = x_0, \\ y &= x, \end{aligned} \qquad A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{6.1}$$

where x is the state, $y = [y_1, y_2]^{\mathrm{T}}$ is the measurement, u is the control, and the saturation function (1.2) is defined by $u_{\max} = -u_{\min} = 0.25$. The nominal PID controller is

$$\begin{aligned} \dot{e}_i &= r - y_1, & e_i(0) &= e_{i0}, \\ u_c &= k_p(r - y_1) + k_i e_i - k_d y_2, \end{aligned} \tag{6.2}$$

where $(k_p, k_i, k_d) = (8, 4, 4)$. Here, the reference input r is a constant so that $\dot{r} \equiv 0$. The closed-loop system comprising (6.1) and (6.2) with $u = u_c$ is the *nominal system* and denoted by Σ_n. When the plant is unconstrained, i.e. $u_{\max} = -u_{\min} = \infty$, the same closed-loop system (6.1), (6.2) will be called the *unconstrained system* and denoted by Σ_u.

6.1.1 Nonlinear Anti-windup Compensated System

The nonlinear anti-windup *compensator*[1] is described by [24]

$$\begin{aligned} \dot{x}_e &= Ax_e + B(\mathrm{sat}(\mathrm{sat}(u_c) + v_1) - u_c), \qquad x_e(0) = x_{e0}, \\ v_1 &= -K_g \,\mathrm{sat}(K_g^{-1} K_q x_{e1}) - \gamma_E(x_{e1}, x_{e2}) K_0 x_{e2}, \\ \gamma_E(x_{e1}, x_{e2}) &= \begin{cases} 1, & \text{if } x_{e1}x_{e2} \geq 0, \\ \frac{K_g \,\mathrm{sat}(K_g^{-1} K_q x_{e1})}{K_q x_{e1}}, & \text{otherwise,} \end{cases} \end{aligned} \tag{6.3}$$

where $x_e = [x_{e1}, x_{e2}]^{\mathrm{T}}$ is the state and $(K_g, K_q, K_0) = (0.99, 80, 100)$ are the anti-windup gains. To incorporate the anti-windup compensator (6.3), the nominal controller (6.2) is modified by replacing measurement y with $y - x_e$, i.e.

$$\begin{aligned} \dot{e}_i &= r - (y_1 - x_{e1}), & e_i(0) = e_{i0}, \\ u_c &= k_p(r - (y_1 - x_{e1})) + k_i e_i - k_d(y_2 - x_{e2}), \end{aligned} \tag{6.4}$$

while the final *anti-windup compensated controller* output becomes[2]

$$u = \mathrm{sat}(u_c) + v_1. \tag{6.5}$$

The nonlinear anti-windup *compensated controller* induced by the nonlinear anti-windup *compensator* (6.3) is then defined by (6.3), (6.4), and (6.5). The closed-loop system (6.1), (6.3), (6.4), and (6.5), will be called the *nonlinear anti-windup compensated system* and denoted by Σ_{aw}.

6.1.2 Approximate Nominal System

It can be seen from the description of the nominal controller (6.2) that the output equation is dependent on measurements and reference input, hence is of the form (2.24). We use the technique of Section 2.6 (see also Section 2.8.1 and Section A.2 in Appendix A) to obtain an approximate nominal controller of the form (2.26) described by

$$\begin{aligned} \dot{x}_{c1} &= r - y_1, & x_{c1}(0) &= e_{i0}, \\ \dot{x}_{c2} &= a(k_p(r - y_1) - k_d y_2 - x_{c2}), & x_{c2}(0) &= k_p(r - y_1(0)) - k_d y_2(0), \\ u_c &= k_i x_{c1} + x_{c2}, \end{aligned}$$

[1] See Sections 1.2 and 1.3.1 for discussions on the distinction between an *anti-windup compensator* and the associated *anti-windup compensated controller*.

[2] In [24], the control signal u is defined with an additional saturation, i.e. $u = \mathrm{sat}(\mathrm{sat}(u_c) + v_1)$. Because the input to the plant (6.1) is already constrained, this final saturation is *redundant*. If the convention of [24] is adopted, the control signal will be artificially constrained and may give the false appearance of achieving controller state-output consistency (see Theorem 2.5.3).

where $a > 0$ is the approximation parameter. Notice that the augmented state x_{c2} can approximate the signal $k_p(r - y_1) - k_d y_2$ arbitrarily well when $a > 0$ is chosen sufficiently large. The preceding approximate nominal controller can be written as

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_{cy} y + B_{cr} r, \qquad x_c(0) = x_{c0}, \\ u_c &= C_c x_c, \end{aligned} \tag{6.6}$$

$$A_c = \begin{bmatrix} 0 & 0 \\ 0 & -a \end{bmatrix}, \qquad B_{cy} = \begin{bmatrix} -1 & 0 \\ -ak_p & -ak_d \end{bmatrix}, \qquad B_{cr} = \begin{bmatrix} 1 \\ ak_p \end{bmatrix}, \qquad C_c = \begin{bmatrix} k_i & 1 \end{bmatrix},$$

where $x_{c0} = [e_{i0}, k_p(r - y_1(0)) - k_d y_2(0)]^{\mathrm{T}}$.

The closed-loop system comprising (6.1) and (6.6) with $u = u_c$ will be called the *nominal approximate system* and denoted by Σ_{na}. When $u_{\max} = -u_{\min} = \infty$, the same closed-loop system (6.1) and (6.6) will be called the *unconstrained approximate system* and denoted by Σ_{ua}. System Σ_{ua} can written as

$$\dot{x}_{ua} = A_{ua} x_{ua} + B_{ua} r, \qquad A_{ua} = \begin{bmatrix} A & BC_c \\ B_{cy} & A_c \end{bmatrix}, \qquad B_{ua} = \begin{bmatrix} 0 \\ B_{cr} \end{bmatrix},$$

where $x_{ua} := (x, x_c)$. Clearly, for the nominal approximate system Σ_{na} to be at least locally stable, it is necessary for A_{ua} to be Hurwitz. By applying the Routh Criterion [185, p. 177] to the characteristic polynomial of A_{ua}, it can be shown that A_{ua} is Hurwitz when $a > 2.6559$ for $(k_p, k_i, k_d) = (8, 4, 4)$. By comparing the time responses for different values of a as in Section 2.8.1, it was found that setting $a = 50$ gives a good approximation of systems Σ_{na}, Σ_{ua} to systems Σ_n, Σ_u respectively. Hence we fix $a = 50$.

6.1.3 GPAW-Compensated System

Applying GPAW compensation to the approximate nominal controller (6.6) yields the GPAW-compensated controller (2.27), which, using the closed-form expressions (A.7) in Appendix A, can be written as

$$\begin{aligned} \dot{x}_g &= \begin{cases} \big(I - \frac{1}{C_c \Gamma C_c^{\mathrm{T}}} \Gamma C_c^{\mathrm{T}} C_c\big)(A_c x_g + B_{cy} y + B_{cr} r), & \text{if } A_{\max} \vee A_{\min}, \\ A_c x_g + B_{cy} y + B_{cr} r, & \text{otherwise}, \end{cases} \\ u_g &= C_c x_g, \qquad x_g(0) = x_{g0}, \\ A_{\max} &= (u_g \geq u_{\max}) \wedge \big(C_c(A_c x_g + B_{cy} y + B_{cr} r) > 0\big), \\ A_{\min} &= (u_g \leq u_{\min}) \wedge \big(C_c(A_c x_g + B_{cy} y + B_{cr} r) < 0\big). \end{aligned} \tag{6.7}$$

We fix the GPAW parameter as the identity matrix $\Gamma = I$. The closed-loop system comprising (6.1) and (6.7) with $u = u_g$ and $x_{g0} = x_{c0}$ will be called the *GPAW-compensated system* and denoted by Σ_g.

For reasons that will become obvious in the next section, we note that the controller state can be arbitrarily initialized. To improve the transient response, we will initialize it within the unsaturated region, i.e. $x_{g0} \in K := \{\bar{x} \in \mathbb{R}^2 \mid \mathrm{sat}(\bar{x}) = \bar{x}\}$. Decompose the controller state and initial state as $x_g = [x_{g1}, x_{g2}]^{\mathrm{T}}$ and $x_{g0} = [x_{g01}, x_{g02}]^{\mathrm{T}}$ respectively. For x_{g2} to be a good approximation of the signal $k_p(r - y_1) - k_d y_2$ starting from time $t = 0$, we need $x_{g2}(0) = x_{g02} = k_p(r - y_1(0)) - k_d y_2(0)$. Then it can be verified that setting[3]

$$x_{g1}(0) = x_{g01} = \frac{\text{sat}(C_c x_{c0}) - x_{g02}}{k_i},$$

where x_{c0} is the initial controller state for the approximate nominal controller (6.6), ensures $x_g(0) = x_{g0} \in K$. As will be seen in the next section, this initialization yields significant improvements in the transient response. We denote the GPAW-compensated system with the described controller state initialization by Σ_{gi}.

6.1.4 Numerical Results

The unconstrained system Σ_u, nominal system Σ_n, nonlinear anti-windup compensated system Σ_{aw}, and GPAW-compensated systems Σ_g, Σ_{gi}, are simulated for r being a unit step input. The time responses are shown in Fig. 6-1. Observe that system Σ_g exhibits a stable response with a sluggish transient. The controller state initialization $x_{g0} \in K$ improves the transient response of system Σ_{gi} significantly.

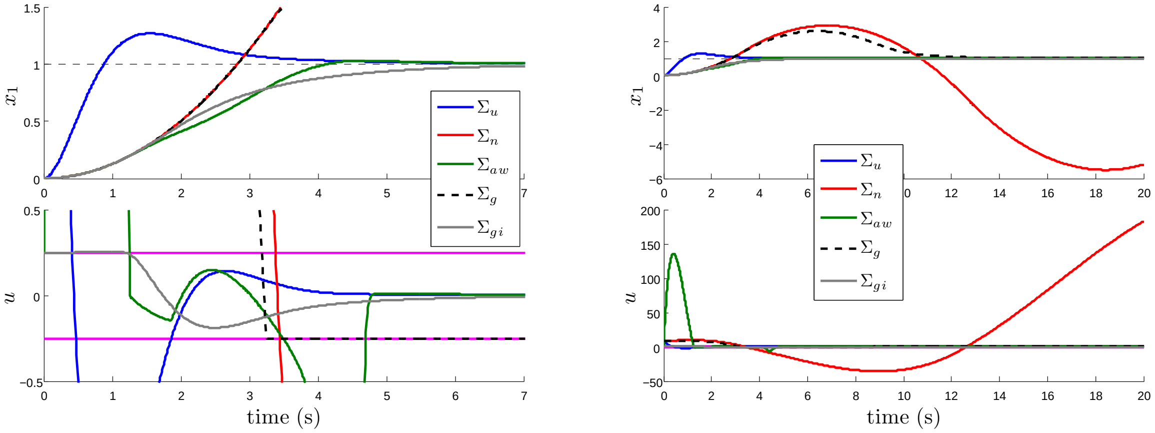


Figure 6-1: Time responses for double integrator plant. The left plot shows a close-up view, while the right plot shows a macro view. Observe that system Σ_g exhibits a stable response despite having a sluggish transient. The controller state initialization for system Σ_{gi} improves the transient response significantly.

Now, compare the responses of systems Σ_{aw} and Σ_{gi}. Observe that the GPAW-compensated system Σ_{gi} expends significantly less control effort in comparison to the nonlinear anti-windup compensated system Σ_{aw}. Moreover, controller state-output consistency (Theorem 2.5.3) holds for the GPAW-compensated systems Σ_g and Σ_{gi}, but does not apply to the nonlinear anti-windup compensated system Σ_{aw}. In terms of the output response x_1, Σ_{gi} has an initially faster response but reaches steady state later. We see that the performance of system Σ_{gi} is comparable to that of system Σ_{aw}.

6.2 Two Anti-windup Schemes for Nonlinear Systems

In this section, we compare the GPAW scheme against two anti-windup schemes developed for nonlinear systems, using the example in [65]. The first is proposed in [65] and applicable

[3] Note that this is similar to back-calculation [53], but applied only to controller state initialization.

to feedback linearizable nonlinear systems. This method has the attractive feature that the anti-windup compensator is fully defined independent of any parameters. The second is the nonlinear anti-windup scheme for Euler-Lagrange systems [24] that was also compared in Section 6.1.

The second order saturated feedback linearizable nonlinear plant is described by [65]

$$\begin{aligned} \dot{x} &= \begin{bmatrix} x_2 \\ \frac{-10x_1 - 0.1x_1^3 - 48.54x_2 - w + \mathrm{sat}(u)}{6.67(1+0.1\sin x_1)} \end{bmatrix}, \\ y &= x, \end{aligned} \tag{6.8}$$

where $x = [x_1, x_2]^{\mathrm{T}} \in \mathbb{R}^2$ is the state, $u \in \mathbb{R}$ is the control input, $w \in \mathbb{R}$ is a disturbance input, and $y \in \mathbb{R}^2$ is the measurement. The saturation limits are $u_{\max} = -u_{\min} = 100$ [65]. A feedback linearizing controller was designed in [65] and described by

$$\begin{aligned} \dot{x}_c &= x_1 - \tilde{r}, \\ u_c &= 10x_1 + 0.1x_1^3 + 48.54x_2 \\ &\qquad + 6.67(1 + 0.1\sin x_1)(-k_p(x_1 - \tilde{r}) - 200(x_2 - \dot{\tilde{r}}) - 6400x_c - \ddot{\tilde{r}}), \end{aligned} \tag{6.9}$$

where $k_p > 0$ is the proportional gain, $x_c \in \mathbb{R}$ is the controller state, $y = x = [x_1, x_2]^{\mathrm{T}}$ is the measurement, $r := [\tilde{r}, \dot{\tilde{r}}, \ddot{\tilde{r}}]^{\mathrm{T}}$ is the reference input, and $u_c \in \mathbb{R}$ is the controller output.

Remark 6.2. We will show in Section 6.2.5 that the nominal design in [65] with $k_p = 400$ corresponds to a design with poorly damped dynamics for the *unconstrained* system. As will be shown, more reasonable designs with good damping can be obtained by simply increasing k_p. □

The closed-loop system comprising (6.8) and (6.9) with $u = u_c$ will be called the *nominal system* and denoted by Σ_n. The same closed-loop system with the plant unconstrained, i.e. $u_{\max} = -u_{\min} = \infty$, will be called the *unconstrained system* and denoted by Σ_u. It can be verified that the unconstrained system can be written as

$$\begin{aligned} \dot{x}_{cl} &= A_{cl}x_{cl} + \begin{bmatrix} 0 \\ -\frac{w}{6.67(1+0.1\sin x_1)} + k_p\tilde{r} + 200\dot{\tilde{r}} - \ddot{\tilde{r}} \\ -\tilde{r} \end{bmatrix}, \\ A_{cl} &= \begin{bmatrix} 0 & 1 & 0 \\ -k_p & -200 & -6400 \\ 1 & 0 & 0 \end{bmatrix}, \end{aligned} \tag{6.10}$$

where $x_{cl} := [x_1, x_2, x_c]^{\mathrm{T}}$ is the state of the closed-loop system.

6.2.1 Feedback Linearized Anti-windup Compensated System

For distinction and ease of reference, we will call the anti-windup scheme of [65] the *feedback linearized anti-windup scheme.*[4] As shown in [65], the associated *anti-windup compensator*[5]

[4] This is just a name for ease of distinguishing between the different anti-windup schemes introduced. It does not imply that the anti-windup compensator uses feedback linearization in any significant way.

[5] See Sections 1.2 and 1.3.1 for discussions on the distinction between an *anti-windup compensator* and the associated *anti-windup compensated controller.*

is given by

$$\begin{aligned}\dot{x}_{aw} &= \begin{bmatrix} 0 & 1 \\ -k_p & -200 \end{bmatrix} x_{aw} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \frac{u_c - \mathrm{sat}(u_c)}{6.67(1+0.1\sin x_1)}, \\ \xi &= -\begin{bmatrix} 1 & 0 \end{bmatrix} x_{aw}, \end{aligned} \tag{6.11}$$

where $\xi \in \mathbb{R}$ is the anti-windup compensator output. To incorporate the preceding anti-windup compensator, the state equation of the nominal controller (6.9) is modified with ξ, resulting in the modified nominal controller

$$\begin{aligned} \dot{x}_c &= x_1 - \tilde{r} - \xi, \\ u_c &= 10x_1 + 0.1x_1^3 + 48.54x_2 \\ &\qquad + 6.67(1+0.1\sin x_1)(-k_p(x_1 - \tilde{r}) - 200(x_2 - \dot{\tilde{r}}) - 6400x_c - \ddot{\tilde{r}}). \end{aligned} \tag{6.12}$$

The closed-loop system comprising (6.8), (6.11), and (6.12), with $u = u_c$, will be called the *feedback linearized anti-windup compensated system* and denoted by Σ_{awf}.

6.2.2 Nonlinear Anti-windup Compensated System

The nonlinear anti-windup compensator for the plant (6.8) can be verified to be [24]

$$\begin{aligned} \dot{x}_e &= \begin{bmatrix} x_{e2} \\ \tilde{f}(x, x_e, u, u_c) \end{bmatrix}, \\ v_1 &= \gamma(x, x - x_e), \end{aligned} \tag{6.13}$$

where $x_e = [x_{e1}, x_{e2}]^{\mathrm{T}}$ is the state, $v_1 \in \mathbb{R}$ is the anti-windup compensator output, and

$$\begin{aligned} \tilde{f}(x, x_e, u, u_c) &= \frac{-10x_1 - 0.1x_1^3 - 48.54x_2 + \mathrm{sat}(u)}{6.67(1+0.1\sin x_1)} \\ &\quad - \frac{-10(x_1 - x_{e1}) - 0.1(x_1 - x_{e1})^3 - 48.54(x_2 - x_{e2}) + u_c}{6.67(1+0.1\sin(x_1 - x_{e1}))}, \\ \gamma(x, x - x_e) &= 10x_1 + 0.1x_1^3 - 10(x_1 - x_{e1}) - 0.1(x_1 - x_{e1})^3 \\ &\quad - K_g \,\mathrm{sat}(K_g^{-1} K_q x_{e1}) - \gamma_E(x_{e1}, x_{e2}) K_0 x_{e2}, \\ &= 10x_{e1} + 0.1(x_1^3 - (x_1 - x_{e1})^3) - K_g \,\mathrm{sat}(K_g^{-1} K_q x_{e1}) - \gamma_E(x_{e1}, x_{e2}) K_0 x_{e2}, \\ \gamma_E(x_{e1}, x_{e2}) &= \begin{cases} 1, & \text{if } x_{e1}x_{e2} \geq 0, \\ \frac{K_g \,\mathrm{sat}(K_g^{-1} K_q x_{e1})}{K_q x_{e1}}, & \text{otherwise.} \end{cases} \end{aligned}$$

The output of the *anti-windup compensated controller* is given by[6]

$$u = \mathrm{sat}(u_c) + v_1. \tag{6.14}$$

In [65], the anti-windup gains were chosen to be $(K_g, K_q, K_0) = (0.99, 240, 100)$.

To incorporate the *nonlinear anti-windup compensator* (6.13), the measurement input

[6] In [24], the controller output u is with an extra saturation, i.e. $u = \mathrm{sat}(\mathrm{sat}(u_c) + v_1)$. This final saturation has been incorporated in the plant (6.8) and hence redundant.

of the nominal controller (6.9) is modified from y to $y - x_e$ to yield

$$\begin{aligned} \dot{x}_c &= x_1 - x_{e1} - \tilde{r}, \\ u_c &= 10(x_1 - x_{e1}) + 0.1(x_1 - x_{e1})^3 + 48.54(x_2 - x_{e2}) + 6.67(1 + 0.1\sin(x_1 - x_{e1})) \\ &\qquad \times (-k_p(x_1 - x_{e1} - \tilde{r}) - 200(x_2 - x_{e2} - \dot{\tilde{r}}) - 6400x_c - \ddot{\tilde{r}}). \end{aligned} \tag{6.15}$$

The closed-loop system comprising (6.8), (6.13), (6.15), and (6.14), will be called the *nonlinear anti-windup compensated system* and denoted by Σ_{awn}.

6.2.3 Approximate Nominal System

It can be seen from the description of the nominal controller (6.9) that the output equation is dependent on measurements and reference input, hence is of the form (2.24). We use the technique of Section 2.6 (see also Section 2.8.1) to obtain an approximate nominal controller of the form (2.26) described by

$$\begin{aligned} \dot{x}_{c1} &= x_1 - \tilde{r}, \\ \dot{x}_{c2} &= a(z_1(y) - x_{c2}), \\ \dot{x}_{c3} &= a(z_2(y, r) - x_{c3}), \\ u_c &= x_{c1}x_{c2} + x_{c3}, \end{aligned} \tag{6.16}$$

where $a > 0$ is the approximation parameter and

$$\begin{aligned} z_1(y) &:= -6400 \times 6.67(1 + 0.1\sin x_1), \\ z_2(y, r) &:= 10x_1 + 0.1x_1^3 + 48.54x_2 + 6.67(1 + 0.1\sin x_1)(-k_p(x_1 - \tilde{r}) - 200(x_2 - \dot{\tilde{r}}) - \ddot{\tilde{r}}). \end{aligned}$$

Notice that the augmented state variables (x_{c2}, x_{c3}) can approximate $(z_1(y), z_2(y, r))$ arbitrarily well when $a > 0$ is chosen sufficiently large. To ensure good approximation of these signals from $t = 0$ onwards, we initialize these augmented states according to

$$x_{c2}(0) = z_1(y(0)), \qquad x_{c3}(0) = z_2(y(0), r(0)). \tag{6.17}$$

Given the initial state x_{c0} of the nominal controller (6.9), the remaining state is initialized as $x_{c1}(0) = x_{c0}$.

The closed-loop system comprising (6.8) and (6.16) with $u = u_c$ will be called the *nominal approximate system* and denoted by Σ_{na}. When $u_{\max} = -u_{\min} = \infty$, the same closed-loop system (6.8), (6.16) will be called the *unconstrained approximate system* and denoted by Σ_{ua}. By comparing the time responses for different values of $a > 0$ as in Section 2.8.1, it was found that setting $a = 200$ gives a good approximation of systems Σ_{na}, Σ_{ua} to systems Σ_n, Σ_u respectively. The responses of these systems to the reference and disturbance input $w = w_1$ defined in (6.18) are shown in Fig. 6-2. We fix the approximation parameter as $a = 200$.

6.2.4 GPAW-Compensated System

The approximate nominal controller (6.16) can be written in the form (2.26), which defines the functions f_c and g_c. Applying GPAW compensation to the approximate nominal controller (6.16) then yields the GPAW-compensated controller (2.27), which can be realized

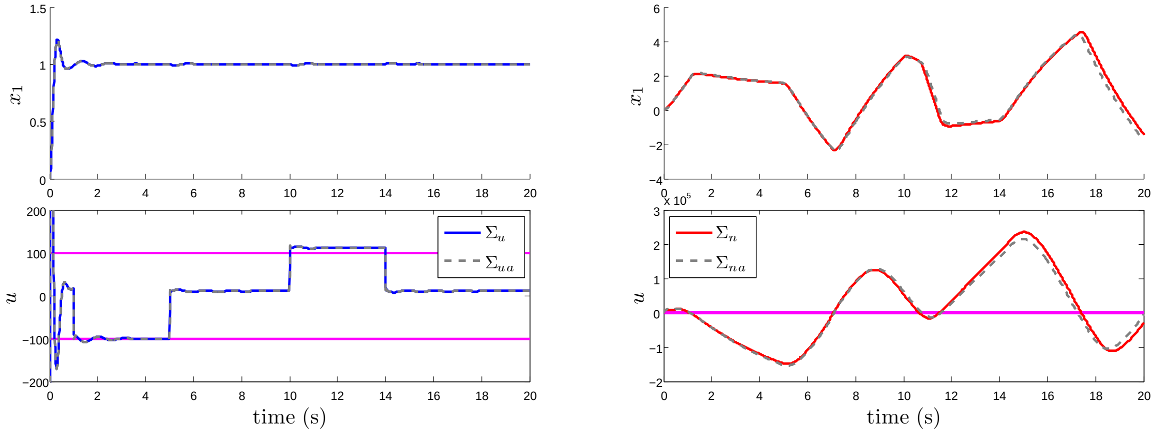


Figure 6-2: Approximation of systems with parameter $a = 200$.

using the closed-form expressions (A.5) in Appendix A. In the realization (A.5), we note that

$$g_c(x_g) = x_{g1}x_{g2} + x_{g3}, \qquad \nabla g_c(x_g) = \begin{bmatrix} x_{g2} & x_{g1} & 1 \end{bmatrix}^{\mathrm{T}},$$

where $x_g = [x_{g1}, x_{g2}, x_{g3}]^{\mathrm{T}}$ is the state of the GPAW-compensated controller (2.27).

We fix the GPAW parameter as the identity matrix $\Gamma = I$. The closed-loop system comprising (6.8) and (2.27) with $u = u_g$ will be called the *GPAW-compensated system* and denoted by Σ_g. For simplicity, we initialize the state of the GPAW-compensated controller to the initial state of the approximate nominal controller, i.e. $x_g(0) = x_c(0)$, where $x_c = [x_{c1}, x_{c2}, x_{c3}]^{\mathrm{T}}$ is the state of (6.16) (see also (6.17)).

6.2.5 Numerical Results

Define the reference r and disturbance inputs w_1, w_2 by

$$\begin{gathered} \ddot{\tilde{r}} + 11.4\dot{\tilde{r}} + 129.96\tilde{r} = 129.96, \qquad (\tilde{r}(0), \dot{\tilde{r}}(0)) = (0, 0), \\ w_1(t) = \begin{cases} -111, & \text{if } 1 \le t < 5, \\ 101, & \text{if } 10 \le t < 14, \\ 0, & \text{otherwise,} \end{cases} \qquad w_2(t) = \begin{cases} -111, & \text{if } 4 \le t < 8, \\ 101, & \text{if } 10 \le t < 14, \\ 0, & \text{otherwise.} \end{cases} \end{gathered} \tag{6.18}$$

The disturbance inputs are shown in Fig. 6-3. Notice that w_1 comprises a negative pulse followed by a positive pulse, while w_2 is obtained from w_1 by delaying the first negative pulse by 3 seconds.

With the proportional gain set at $k_p = 400$ (nominal design), we first simulate the responses of the unconstrained system Σ_u, nominal system Σ_n, feedback linearized anti-windup compensated system Σ_{awf}, nonlinear anti-windup compensated system Σ_{awn}, and GPAW-compensated system Σ_g, subject to the reference $r = [\tilde{r}, \dot{\tilde{r}}, \ddot{\tilde{r}}]^{\mathrm{T}}$ and disturbance $w = w_1$ (6.18). These conditions correspond to those used in [65]. The responses are shown in the left plot of Fig. 6-4, from which we see that all three anti-windup compensated systems Σ_{awf}, Σ_{awn}, and Σ_g have significantly superior responses compared to the uncompensated system Σ_n. The responses of systems Σ_{awf} and Σ_{awn} are comparable, while it appears that the GPAW-compensated system Σ_g exhibits some undesirable behavior, like the large overshoot at approximately $t = 1$ s followed by a slow decay, and the oscillations that follow

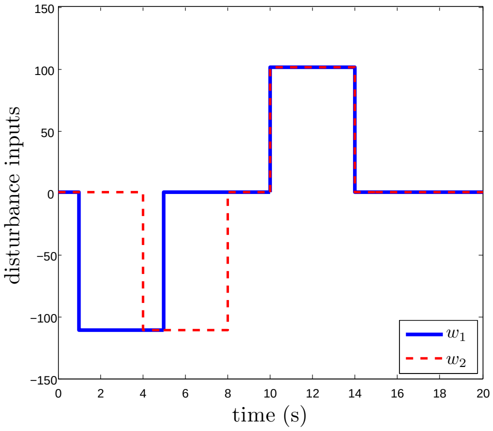


Figure 6-3: Two disturbance inputs.

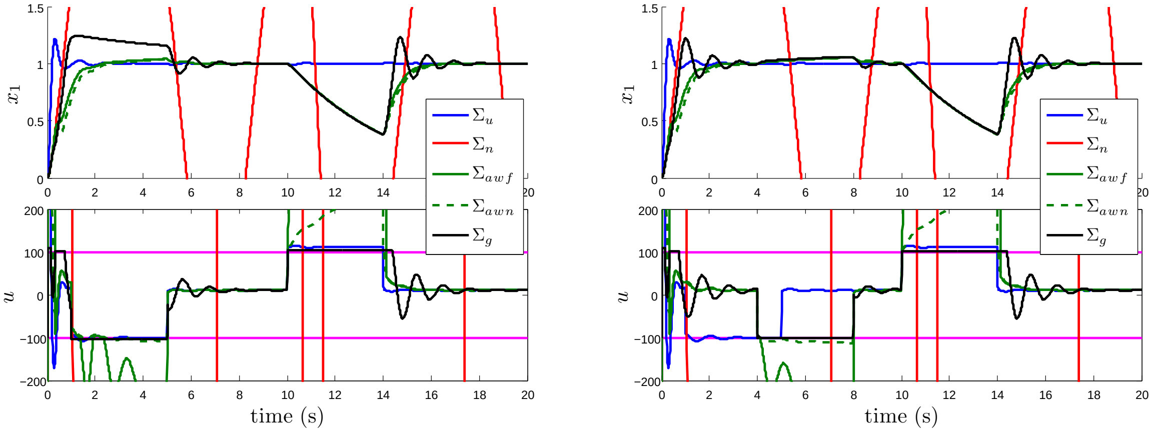


Figure 6-4: Disturbance response with nominal design, $k_p = 400$. The left plot shows the system responses with disturbance input $w = w_1$, while the right plot shows the case with $w = w_2$.

when the disturbance input switches to zero at times $t = 5$ s and $t = 14$ s. We will show that these undesirable phenomena are due to a poor design of the *unconstrained* system Σ_u coupled with the disturbance acting at an inopportune time, i.e. before transients subside.

First, we delay the negative pulse of the disturbance input by 3 seconds, i.e. using the disturbance input $w = w_2$ (6.18), for which the corresponding responses are shown in the right plot of Fig. 6-4. It can be seen that the slow decay (in the left plot of Fig. 6-4 for $t \in [1, 5]$) is no longer exhibited by system Σ_g, while the oscillations remain. This shows that the slow decay is due to the disturbance acting during transients and interacting with the oscillations at inopportune times.

These oscillations are the natural response of the *unconstrained* system, as seen from the left plot of Fig. 6-5, where the closed-loop system is subject to non-zero initial conditions but is *unforced*, i.e. $\tilde{r} \equiv 0$ and $w \equiv 0$. The oscillatory behavior is expected because the unconstrained system is poorly damped. This can be seen by inspection of the closed-loop poles. By inspecting the eigenvalues of A_{cl} (6.10), it can be verified that the unconstrained system has a real pole at -198 and a pair of complex conjugate poles at $-0.93 \pm 5.6i$. The dominant complex conjugate poles have natural undamped frequency of 5.7 rad/s and

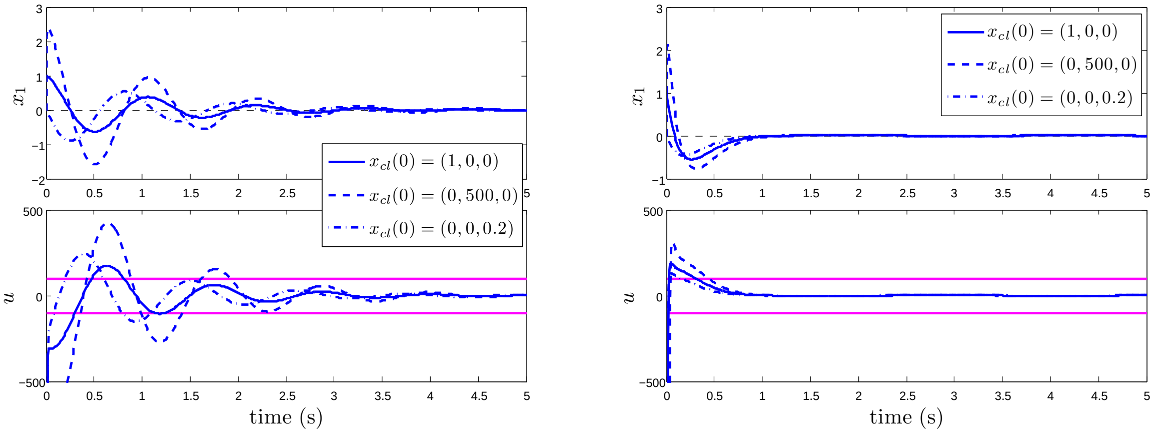


Figure 6-5: Initial response of unforced unconstrained system, Σ_u. The left plot shows the case of nominal design where $k_p = 400$, while the right plot shows the case where $k_p = 2000$.

damping ratio[7] of 0.16. Clearly, the unconstrained system Σ_u is poorly damped. We could argue that if such poorly damped and oscillatory behavior is acceptable for the *unconstrained* system, then it must be acceptable for the anti-windup compensated system, in particular, the GPAW-compensated system Σ_g.

However, the unconstrained system can be better designed by simply increasing the proportional gain k_p. In particular, setting $k_p = 2000$ results in superior damping. With $k_p = 2000$, the closed-loop poles are located at -190 and $-5.2 \pm 2.6i$, with associated natural undamped frequency of 5.8 rad/s and damping ratio of 0.89. This choice of k_p changes the real pole and natural undamped frequency of the complex poles marginally, while improving the damping ratio significantly. The unforced response of the resulting unconstrained system Σ_u with non-zero initial conditions is shown in the right plot of Fig. 6-5. The superiority of this design is clear, and would be a reasonable design for the *unconstrained* system.

With this new design ($k_p = 2000$), the system responses subjected to the same disturbance inputs are shown in Fig. 6-6. The undesirable phenomena exhibited by system Σ_g in Fig. 6-4 have been suppressed in Fig. 6-6. Moreover, the response of the GPAW-compensated system Σ_g has a much shorter rise time when compared to the responses of systems Σ_{awf} and Σ_{awn}. Clearly, the GPAW-compensated system Σ_g exhibits responses comparable (or even superior) to those of systems Σ_{awf} and Σ_{awn}.

Note that we are not advocating the redesign of the nominal controller, but simply showing that the relatively poor response of the GPAW-compensated system in Fig. 6-4 is due to a poor nominal controller design, rather than inherent weaknesses of GPAW compensation. The main point is that if the unconstrained system is designed to exhibit oscillatory responses, then it must be acceptable for the GPAW-compensated system to exhibit similar oscillatory responses.

6.3 LMI-based Anti-windup Scheme for Stable LTI Systems

The anti-windup scheme of [128] is a popular method based on the standard anti-windup framework depicted in Fig. 1-1, and applicable to saturated stable LTI systems. One feature

[7] Recall that for a pair of complex conjugate poles satisfying the equation $s^2 + 2\zeta\omega_n s + \omega_n^2 = 0$, the damping ratio is ζ and the natural undamped frequency is ω_n.

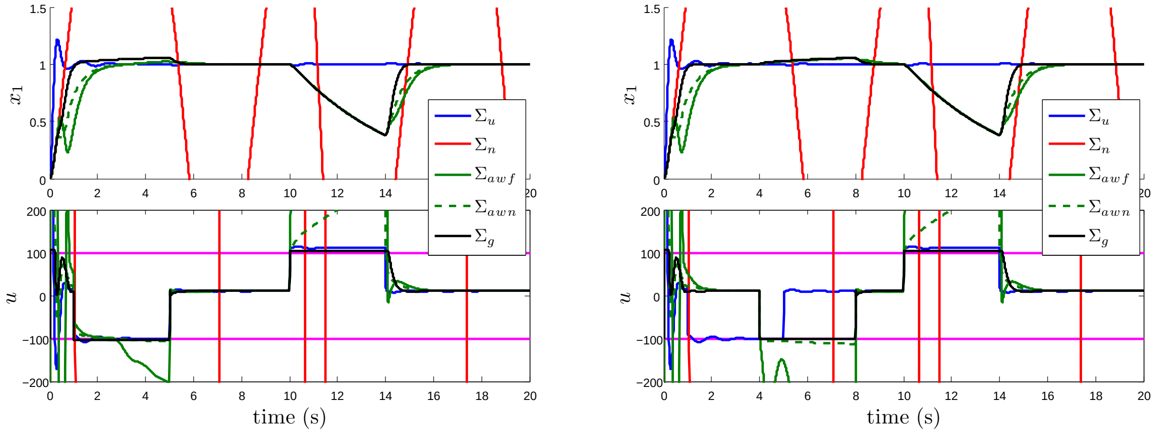


Figure 6-6: Disturbance response with improved design, $k_p = 2000$. The left plot shows the system responses with disturbance input $w = w_1$, while the right plot shows the case with $w = w_2$.

is that the anti-windup compensator is found by solving two LMI problems, which renders it numerically attractive. Here, we compare the method against the GPAW scheme using the two examples in [128]. The first is on the longitudinal dynamics of the F8 aircraft which was introduced in [83], and the second is on a cart-spring-pendulum system.[8]

The saturated LTI plant with input disturbance w is described by[9]

$$\begin{aligned} \dot{x} &= Ax + B\,\mathrm{sat}(u) + B_w w, \\ y &= Cx, \end{aligned} \tag{6.19}$$

while the nominal controller is described by

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_{cy} y + B_{cr} r, \\ u_c &= C_c x_c. \end{aligned} \tag{6.20}$$

The closed-loop system comprising (6.19) and (6.20) with $u = u_c$ is called the *nominal system* and denoted by Σ_n. The same system (6.19) and (6.20) with $u_{\max,i} = -u_{\min,i} = \infty$ will be called the *unconstrained system* and denoted by Σ_u.

For the LMI-based anti-windup compensated system, the nominal controller's state evolution and output are modified by signals v_1 and v_2 to have the form[9]

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_{cy} y + B_{cr} r + v_1, \\ u_c &= C_c x_c + v_2, \end{aligned} \tag{6.21}$$

with (v_1, v_2) being outputs of the anti-windup compensator

$$\begin{aligned} \dot{x}_{aw} &= A_{aw} x_{aw} + B_{aw}(u_c - \mathrm{sat}(u_c)), \\ \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} &= C_{aw} x_{aw} + D_{aw}(u_c - \mathrm{sat}(u_c)). \end{aligned} \tag{6.22}$$

[8] In [128], the cart-spring-pendulum system was an experimental setup. Here, we will only be performing simulation comparisons.

[9] The representations of the plant and controller in [128] are more general. We specialize them for the two examples by eliminating those parts that are not needed.

Setting $C_{aw} = 0$ in (6.22) results in a *static* anti-windup compensator, for which the state update equation in (6.22) becomes redundant. When $C_{aw} \neq 0$ and $B_{aw} \neq 0$, it is a *dynamic* anti-windup compensator. The closed-loop system comprising (6.19), (6.21), and (6.22) with $u = u_c$ will be called the *LMI-based anti-windup compensated system* and denoted by Σ_{aw}. The *performance output*

$$z = C_z x + D_{zu}\,\mathrm{sat}(u) + D_{zr} r, \tag{6.23}$$

will be needed for the LMI-based anti-windup design [128].

Applying GPAW compensation to the nominal controller (6.20) yields the GPAW-compensated controller (2.27), written as

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g, y, r)(A_c x_g + B_{cy} y + B_{cr} r), \\ u_g &= C_c x_g. \end{aligned} \tag{6.24}$$

Recall that the GPAW-compensated controller can be realized by solving a combinatorial optimization subproblem (2.31), a convex quadratic program (4.12), or a projection onto a convex polyhedral cone problem (see Section 4.1). Moreover, closed-form expressions are available when the controller output is of dimension one or two (see Appendices A and B). The closed-loop system (6.19) and (6.24) with $u = u_g$ will be called the *GPAW-compensated system* and denoted by Σ_g.

6.3.1 Longitudinal Dynamics of the F8 Aircraft

The saturated longitudinal dynamics of the F8 aircraft [83] is described by (6.19), saturation limits $u_{\max,i} = -u_{\min,i} = 25$ (deg) for $i \in \{1, 2\}$, and matrices

$$A = \begin{bmatrix} -0.8 & -0.0006 & -12 & 0 \\ 0 & -0.014 & -16.64 & -32.2 \\ 1 & -0.0001 & -1.5 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} -19 & -3 \\ -0.66 & -0.5 \\ -0.16 & -0.5 \\ 0 & 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \qquad B_w = 0.$$

The outputs $y = [y_1, y_2]^{\mathrm{T}} = Cx$ represent the aircraft pitch angle and flight path angle respectively, both in degrees. The inputs $u = [u_1, u_2]^{\mathrm{T}}$ represent the elevator angle and flaperon angle respectively, both also in degrees. The nominal controller is described by (6.20) and matrices [83, 128]

$$A_c = \begin{bmatrix} A_a + B_a G - H C_a & 0 \\ G & 0 \end{bmatrix}, \qquad B_{cy} = -B_{cr} = \begin{bmatrix} H \\ 0 \end{bmatrix}, \qquad C_c = \begin{bmatrix} 0 & I \end{bmatrix}, \tag{6.25}$$

$$A_a = \begin{bmatrix} 0 & 0 \\ B & A \end{bmatrix}, \qquad B_a = \begin{bmatrix} I \\ 0 \end{bmatrix}, \qquad C_a = \begin{bmatrix} 0 & C \end{bmatrix},$$

$$G = \begin{bmatrix} -52.23 & -3.36 & 73.1 & -0.0006 & -94.3 & 1072 \\ -3.36 & -29.7 & -2.19 & -0.006 & 908.9 & -921 \end{bmatrix},$$

$$H = \begin{bmatrix} -0.844 & -11.54 & -0.86 & -47.4 & 4.68 & 4.82 \\ 0.819 & 13.47 & 0.25 & 15 & -4.8 & 0.14 \end{bmatrix}^{\mathrm{T}}.$$

LMI-Based Anti-windup Compensated System

We use the anti-windup synthesis procedure [128, Procedure 1] to design four anti-windup compensators[10] as in the F8 aircraft example in [128]. The first two correspond to a static design and plant-order dynamic anti-windup design for the performance output $z = y - r$ (see (6.23)).

Remark 6.3. When following [128, Procedure 1], there are two somewhat arbitrary implementation decisions that may change the resultant anti-windup compensator gain matrices. Here, we state the choices made so that results can be reproduced in the same manner.

The first is on the decomposition of the matrix $\bar{N} := NN^{\mathrm{T}}$ where the rectangular matrix N is sought (see equation (16) in [128]). Since $\bar{N} = \bar{N}^{\mathrm{T}}$, it has a special eigendecomposition $\bar{N} = \bar{Q}\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\bar{Q}^{\mathrm{T}}$ [124, Corollary 2.5.14, pp. 107 – 108] where λ_i are the eigenvalues and $\bar{Q}$ has the associated eigenvectors as its columns and is orthogonal, i.e. $\bar{Q}^{\mathrm{T}} = \bar{Q}^{-1}$. This decomposition can be easily found by the MATLAB® function `eig`. We choose N to be composed of the columns of $\bar{Q}$ corresponding to nonzero eigenvalues, each multiplied by the square root of the associated eigenvalue, i.e. if $\lambda_i \neq 0$, then the i-th column of $\bar{Q}$ multiplied by $\sqrt{\lambda_i}$ is a column of N. It can be verified that N constructed as such is a full rank matrix satisfying $\bar{N} = NN^{\mathrm{T}}$.

The second implementation choice is on the parameters $\delta > 0$ and V_s [128, Procedure 1, Step 4]. We choose $\delta = 1$ and $V_s = I$, so that $U = \delta V_s^{-1} = I$. □

Following [128, Procedure 1] to design a static anti-windup compensator for the performance output $z = y - r$ yields the matrix $D_{aws1} \in \mathbb{R}^{10\times 2}$ with numeric values[11]

$$\begin{aligned} D_{aws1} = \left[\begin{matrix} -6.0630 & -10.002 & 5.2920 & 1277.8 & 0.18403 & -0.89916 & -6.2023 \\ 68.949 & -328.39 & 133.85 & 19621 & 8.4154 & -10.969 & 66.795 \end{matrix}\right. \\ \left.\begin{matrix} -10.954 & 0.90498 & -1.5567 \\ -344.53 & -2.0689 & -34.401 \end{matrix}\right]^{\mathrm{T}}. \end{aligned}$$

The static anti-windup compensator is (6.22) with $A_{aw} = 0$, $B_{aw} = 0$, $C_{aw} = 0$, and $D_{aw} = D_{aws1}$. The resulting LMI-based anti-windup compensated system will be denoted by Σ_{aws1}.

For the same performance output $z = y - r$, the plant-order dynamic anti-windup design is defined by matrices

$$A_{awd1} = \begin{bmatrix} -0.95591 & -139.50 & -136.67 & 231.64 \\ -2.4906 & -363.48 & -356.11 & 554.51 \\ -1.7471 & -255.25 & -250.17 & -824.85 \\ 0.15826 & 11.696 & 7.9575 & -46478 \end{bmatrix} \times 10^4,$$

[10] The numerical values for the matrices defining the dynamic anti-windup compensators were not listed in [128]. While the matrices Λ_4 given (at the bottom of [128, pp. 1517 – 1518]) are supposed to define the static anti-windup compensators, the responses obtained do not agree with those in [128, Figs. 6 and 7]. The designs obtained by following [128, Procedure 1] produces systems whose responses agree with those in [128, Fig. 6 and 7].

[11] The numeric values of the anti-windup gain matrices as presented are rounded to 5 significant digits, and listed only for verification purposes. The values used in simulations are in full resolution obtained through [128, Procedure 1].

$$B_{awd1} = \begin{bmatrix} -342.43 & -6143.4 \\ -892.43 & -16007 \\ -630.55 & -11219 \\ -118.73 & 1346.8 \end{bmatrix} \times 10^4,$$

$$C_{awd1} = \begin{bmatrix} -0.00099418 & -0.042736 & -0.042250 & 0.051971 \\ 0.013053 & -0.33842 & -0.38844 & 0.98543 \\ 0.019578 & -0.25296 & -0.34722 & 0.81876 \\ -0.010485 & 0.15023 & 0.56599 & -1.0497 \\ 0.0028824 & 0.024005 & -0.0045734 & -0.024250 \\ -0.0048033 & 0.028520 & 0.0050051 & -0.052220 \\ 0.000061898 & -0.038441 & -0.044896 & 0.061231 \\ 0.030087 & -0.26891 & -0.41181 & 1.0855 \\ -12.876 & -1883.4 & -1846.5 & -14849 \\ -231.53 & -33781 & -33093 & 81270 \end{bmatrix} \times 10^2,$$

$$D_{awd1} = \begin{bmatrix} -0.0015596 & -0.0089517 \\ -0.0068185 & -0.13679 \\ -0.0042489 & -0.077436 \\ 0.0067893 & 0.11487 \\ -0.000057559 & -0.00047552 \\ 0.00019815 & 0.0045669 \\ -0.0015658 & -0.0090705 \\ -0.0068035 & -0.13682 \\ -46.803 & -826.22 \\ -828.45 & -14882 \end{bmatrix} \times 10^4.$$

Specifically, it is given by (6.22) with $A_{aw} = A_{awd1}$, $B_{aw} = B_{awd1}$, $C_{aw} = C_{awd1}$, and $D_{aw} = D_{awd1}$. The resulting LMI-based anti-windup compensated system will be denoted by Σ_{awd1}.

To improve the performance, anti-windup compensators were designed for the performance output defined by (6.23) and matrices

$$C_z = \begin{bmatrix} 0 & 0 & 0 & \frac{3}{4} \\ -0.8 & -0.0006 & -12 & 0 \end{bmatrix}, \qquad D_{zu} = 0, \qquad D_{zr} = \begin{bmatrix} -\frac{3}{4} & 0 \\ 0 & 0 \end{bmatrix}. \tag{6.26}$$

The matrix $D_{aws2} \in \mathbb{R}^{10\times 2}$ that defines the static design is given by

$$D_{aws2} = \begin{bmatrix} -14.312 & 55.098 & 0.10642 & 4.8030 & -0.014923 & -0.30187 & -14.379 & 55.009 & 0.97391 & 0.023589 \\ 76.652 & -526.75 & -2.3909 & -30.136 & -0.52966 & -0.69344 & 76.354 & -525.43 & 0.13938 & 0.71856 \end{bmatrix}^{\mathrm{T}},$$

where the static anti-windup compensator is (6.22) with $A_{aw} = 0$, $B_{aw} = 0$, $C_{aw} = 0$, and $D_{aw} = D_{aws2}$. The resulting LMI-based anti-windup compensated system will be denoted by Σ_{aws2}.

For the same performance output defined by (6.23) and (6.26), the plant-order dynamic anti-windup design is defined by matrices

$$A_{awd2} = \begin{bmatrix} 0.067552 & 3.3508 & -358.69 & -162.76 \\ -2.6410 & -129.22 & 14044 & 3711.3 \\ 2.1302 & 99.123 & -11388 & 4604.2 \\ 0.0013977 & 33.404 & 386.35 & -49636 \end{bmatrix} \times 10^4,$$

$$B_{awd2} = \begin{bmatrix} 135.12 & -902.07 \\ -5278.1 & 35261 \\ 4244.6 & -28424 \\ 85.749 & -129.26 \end{bmatrix} \times 10^4,$$

$$C_{awd2} = \begin{bmatrix} 0.0000058254 & 0.00032522 & 0.0016350 & 0.0039970 \\ -0.00011178 & -0.0017618 & -0.0092329 & -0.023160 \\ -0.000062008 & -0.000068350 & 0.00011921 & -0.000079288 \\ -0.0037391 & 0.0040623 & -0.0093207 & -0.0051665 \\ 0.000052263 & -0.00019850 & 0.00016987 & 0.000020133 \\ 0.00046070 & -0.00037181 & -0.00049672 & -0.000097564 \\ 0.000019729 & 0.00030590 & 0.0016650 & 0.0039889 \\ -0.00010796 & -0.0019326 & -0.0084796 & -0.022410 \\ -0.54861 & -26.871 & 2917.0 & 814.51 \\ 3.6685 & 176.71 & -19541 & -1024.4 \end{bmatrix} \times 10^3,$$

$$D_{awd2} = \begin{bmatrix} -0.0072343 & -0.0069948 \\ 0.0058178 & -0.019080 \\ 0.00053019 & -0.0016356 \\ 0.00065605 & -0.026219 \\ -0.00056711 & 0.0013408 \\ -0.00046369 & 0.00012203 \\ -0.0072424 & -0.0071888 \\ 0.0058214 & -0.022481 \\ -1096.5 & 7324.7 \\ 7324.6 & -48969 \end{bmatrix} \times 10^3.$$

Specifically, it is given by (6.22) with $A_{aw} = A_{awd2}$, $B_{aw} = B_{awd2}$, $C_{aw} = C_{awd2}$, and $D_{aw} = D_{awd2}$. The resulting LMI-based anti-windup compensated system will be denoted by Σ_{awd2}.

GPAW-Compensated System

Because the stability results obtained thus far are too conservative to be applied to this system,[12] we use an ad hoc method (optimizing over the time response for a specific reference) to determine the GPAW parameter Γ. As shown in the next section, the resultant GPAW-compensated system exhibits reasonable responses for some reference inputs it was *not* optimized for.

First, we show that for the GPAW-compensated controller (6.24) defined by (6.25), only 15 elements among the 64 in $\Gamma \in \mathbb{R}^{8\times 8}$ are significant while the remaining 49 are redundant or implicitly defined. From (6.25), we have $C_c = \begin{bmatrix} 0 & I \end{bmatrix} = [e_7, e_8]^{\mathrm{T}}$, where

[12] Specifically, the nominal controller is only marginally stable and Theorem 5.2.1 requires both the plant and controller to be strictly stable (see Remark 5.1) and hence not applicable.

$e_i \in \mathbb{R}^8$ for $i \in \mathcal{I}_8$ are standard basis vectors in $\mathbb{R}^8$. Partition $\Gamma = \Gamma^{\mathrm{T}}$ as $\Gamma = \left[\begin{smallmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{12}^{\mathrm{T}} & \Gamma_{22} \end{smallmatrix}\right]$, where $\Gamma_{12} \in \mathbb{R}^{6\times 2}$ and $\Gamma_{22} \in \mathbb{R}^{2\times 2}$. From the closed-form expressions (B.5) in Appendix B, we see that Γ is always post-multiplied by either the gradient vectors $\nabla g_{c1} = e_7$, $\nabla g_{c2} = e_8$, or the matrix $\tilde{N} = C_c^{\mathrm{T}} = [e_7, e_8]$. Such multiplications eliminate the first 6 columns of Γ, so that $\Gamma_{11} \in \mathbb{R}^{6\times 6}$ can never affect the GPAW-compensated controller, rendering it redundant. Hence the GPAW-compensated controller is fully defined by Γ_{12} (12 elements) and $\Gamma_{22} = \Gamma_{22}^{\mathrm{T}}$ (3 independent elements), i.e. 15 elements in total.

Since Γ_{11} is redundant, we can choose it to be $\Gamma_{11} = \alpha I_6$ for some $\alpha > 0$, so that $\Gamma = \left[\begin{smallmatrix} \alpha I_6 & \Gamma_{12} \\ \Gamma_{12}^{\mathrm{T}} & \Gamma_{22} \end{smallmatrix}\right]$. Observing that $\Gamma > 0$ if and only if $\Gamma_{22} > \frac{1}{\alpha}\Gamma_{12}^{\mathrm{T}}\Gamma_{12}$ [124, Theorem 7.7.6, p. 472], we see that $\Gamma > 0$ when $\Gamma_{22} = \Gamma_{22}^{\mathrm{T}} > 0$ and $\alpha > 0$ is chosen sufficiently large. Hence to enforce $\Gamma > 0$, the only constraint we need to impose is $\Gamma_{22} > 0$. For simplicity, we fix $\Gamma_{22} = I_2 > 0$, so that the 12 elements of Γ_{12} can be chosen arbitrarily.

We determine Γ by optimizing the time response of the GPAW-compensated system to the step reference $r = [r_1, r_2]^{\mathrm{T}} = [15, 15]^{\mathrm{T}}$ for all $t \geq 0$ (see (6.24)), and with zero initial conditions. Specifically, we solve the *unconstrained* minimization problem[13]

$$\min_{\Gamma_{12}\in\mathbb{R}^{6\times 2}} F(\Gamma_{12}) = \int_0^3 50(y_1(t) - r_1)^2 + (y_2(t) - r_2)^2 \, dt,$$

where $y(t) = [y_1(t), y_2(t)]^{\mathrm{T}}$ for $t \in [0, 3]$ is the plant output (dependent on Γ) obtained from simulation, and we have fixed $\Gamma_{22} = I_2$ to simplify the optimization problem. As shown above, for any choice of Γ_{12}, there exists a Γ_{11} such that $\Gamma > 0$, which leads to an *unconstrained* optimization problem. The preceding optimization problem was solved using the nonlinear program solver `fminunc` of the MATLAB® Optimization Toolbox [186], which yields $\Gamma = \left[\begin{smallmatrix} \alpha I_6 & \Gamma_{12} \\ \Gamma_{12}^{\mathrm{T}} & I_2 \end{smallmatrix}\right]$ with

$$\Gamma_{12} = \begin{bmatrix} 0.0016116 & -0.0014137 & 0.027579 & -0.00000026455 & -0.054469 & 0.29149 \\ -0.32659 & -0.93800 & 0.77329 & -0.00018794 & 0.73616 & 0.052871 \end{bmatrix}^{\mathrm{T}}.$$

Numerical Results

The systems Σ_u, Σ_n, Σ_{aws1}, Σ_{awd1}, Σ_{aws2}, Σ_{awd2}, and Σ_g are first simulated for the step reference $r = [10, 10]^{\mathrm{T}}$ for all $t \geq 0$. Notice that the GPAW-compensated system was *not* optimized for this reference. The responses for systems Σ_u, Σ_n, Σ_{aws1}, Σ_{awd1}, and Σ_g are shown in Fig. 6-7, while those of Σ_u, Σ_n, Σ_{aws2}, Σ_{awd2}, and Σ_g are shown in Fig. 6-8.

Observe from Fig. 6-7 that when designed for the nominal performance output $z = y - r$, the LMI-based anti-windup compensated systems Σ_{aws1}, Σ_{awd1} exhibit some overshoot in their output y_1. In contrast, the GPAW-compensated system exhibits a superior well damped response. Notice that controller state-output consistency (see Theorem 2.5.3) holds only[14] for the GPAW-compensated system Σ_g, which is also evident in Figs. 6-8 and 6-9.

To reduce the overshoot of the LMI-based anti-windup designs, the performance output was modified to (6.23), (6.26), yielding systems Σ_{aws2} and Σ_{awd2} [128]. Fig. 6-8 shows this

[13] The weighting of 50 in the objective function was chosen to reduce overshoots, so that the GPAW-compensated system exhibits responses comparable to those of the LMI-based anti-windup scheme.

[14] Results shown in [128, Figs. 6 and 7] for the control signals are with *additional* saturation applied, which is conventional in the anti-windup literature. The control signals shown in Figs. 6-7, 6-8, and 6-9, are without the additional saturation.

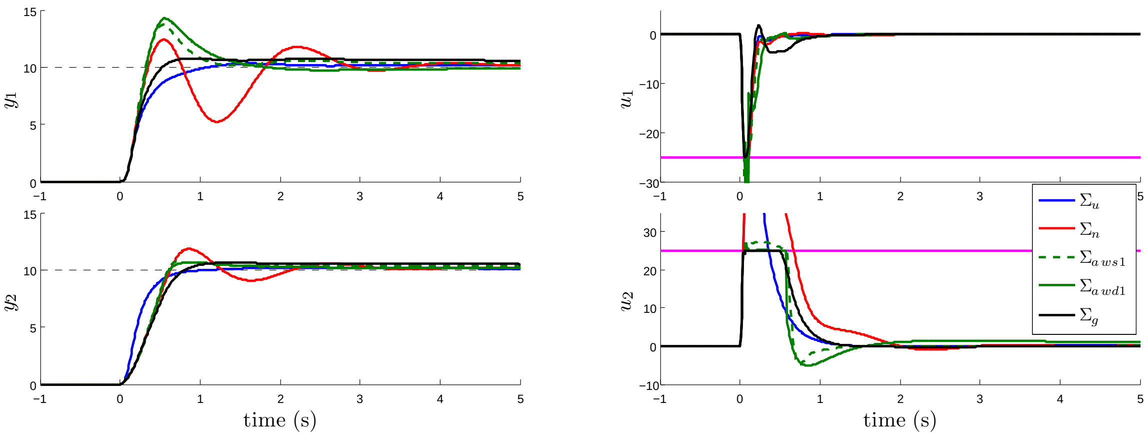


Figure 6-7: Step response for F8 aircraft longitudinal dynamics, nominal case.

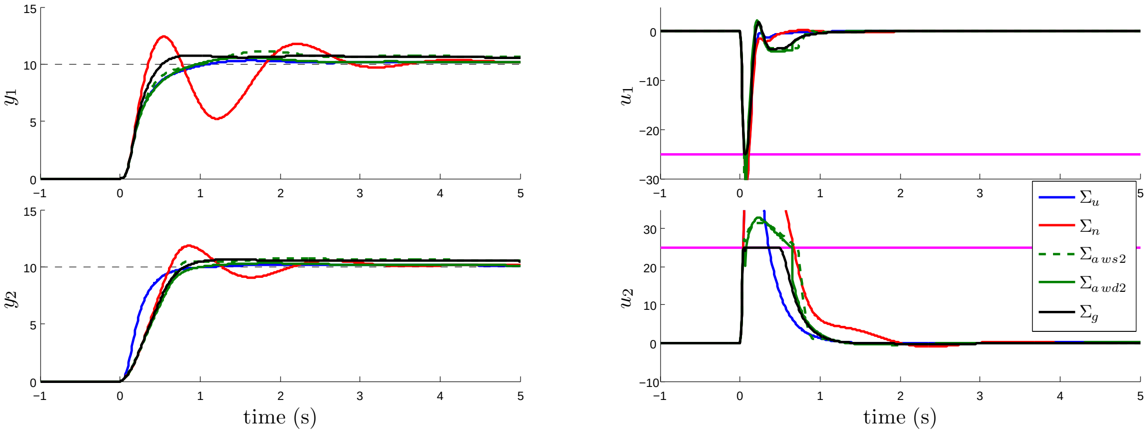


Figure 6-8: Step response for F8 aircraft longitudinal dynamics, with modified performance output for LMI-based anti-windup scheme.

modification to be effective in eliminating the overshoots in the output y_1. The response of the GPAW-compensated system Σ_g in Fig. 6-8 remains unchanged from Fig. 6-7. Clearly, responses of systems Σ_g and Σ_{aws2}, Σ_{awd2} are comparable (with the modified performance output), and are significant improvements over the uncompensated system Σ_n.

The GPAW parameter Γ was determined by optimizing for the reference input, $r = [15, 15]^{\mathrm{T}}$ for all $t \geq 0$, the response of which is shown in Fig. 6-9. Clearly, the response of the GPAW-compensated system is qualitatively similar to those in Figs. 6-7 and 6-8 (with reference input $r = [10, 10]^{\mathrm{T}}$ for all $t \geq 0$), cases for which it was *not* optimized for. This suggests that satisfactory responses can be expected of the GPAW-compensated system for at least a class of reference inputs.

6.3.2 Cart-Spring-Pendulum System

The other example used to demonstrate the effectiveness of the anti-windup scheme proposed in [128] is an experimental cart-spring-pendulum system, described by (6.19), satu-

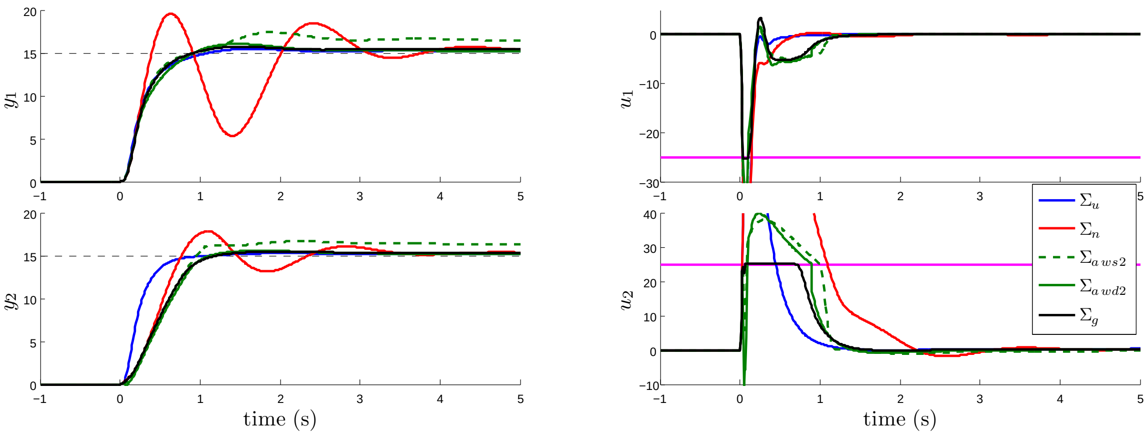


Figure 6-9: Step response for F8 aircraft longitudinal dynamics, with large step input.

ration limits $u_{\max} = -u_{\min} = 5$ (V), and matrices

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -330.46 & -12.15 & -2.44 & 0 \\ 0 & 0 & 0 & 1 \\ -812.61 & -29.87 & -30.10 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 2.71762 \\ 0 \\ 6.68268 \end{bmatrix}, \tag{6.27}$$
$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad B_w = \begin{bmatrix} 0 & 0 & 0 & 15.61 \end{bmatrix}^{\mathrm{T}}.$$

The outputs $y = [y_1, y_2]^{\mathrm{T}} = Cx$ represent the cart's displacement (in meters) from equilibrium and pendulum angle (in radians) from the downward vertical respectively. The input u represent the motor input in Volts. The nominal controller is described by (6.20) and matrices

$$A_c = A - BK - LC, \qquad B_{cy} = L, \qquad B_{cr} = 0, \qquad C_c = -K,$$
$$K = \begin{bmatrix} 64.81 & 213.12 & 1242.27 & 85.82 \end{bmatrix}, \qquad L = \begin{bmatrix} 64 & 2054 & -8 & -1432 \\ -8 & -280 & 142 & 10169 \end{bmatrix}^{\mathrm{T}}. \tag{6.28}$$

LMI-Based Anti-windup Compensated System

In contrast to Section 6.3.1, the LMI-based dynamic anti-windup compensator (6.22) defined by matrices

$$A_{aw} = \begin{bmatrix} -65.02 & 198.43 & 98.11 & -66.75 \\ 223.94 & -697.09 & -347.39 & 247.24 \\ 41.17 & -98.10 & -47.56 & 55.25 \\ -121.39 & 309.97 & 138.31 & -131.52 \end{bmatrix}, \qquad B_{aw} = \begin{bmatrix} 0.0688 \\ -0.2620 \\ -0.0637 \\ 0.1559 \end{bmatrix},$$
$$C_{aw} = \begin{bmatrix} 41.22 & -160.42 & -106.41 & 82.03 \\ -3469.09 & 8318.57 & 3423.87 & -2388.49 \\ -162.51 & 386.26 & -35.56 & 71.07 \\ -4584.37 & 9490.06 & -11350.16 & 11407.08 \\ 587.11 & -1687.16 & -821.25 & 632.86 \end{bmatrix}, \qquad D_{aw} = \begin{bmatrix} -0.0622 \\ 2.9070 \\ 0.2338 \\ 5.5623 \\ 0 \end{bmatrix},$$

in [128], does reproduce the results in [128, Figs. 9 and 10]. As mentioned in [128], the conditions to construct static anti-windup compensators are infeasible.

GPAW-Compensated System

It can be verified that both the open-loop plant (6.19), (6.27), and the nominal controller (6.20), (6.28), are strictly stable. However, condition (5.10) was found to be numerically infeasible, so that Theorem 5.2.1 cannot be applied. This again suggests conservatism of Theorem 5.2.1. Without a systematic mechanism to determine the GPAW parameter Γ, we use an ad hoc method as before.

We will be optimizing some measure of system performance over Γ for a specified disturbance input, similar to the approach in Section 6.3.1. First, we show that we can completely define the GPAW-compensated controller (6.24) using only 3 elements of $\Gamma = \Gamma^{\mathrm{T}} \in \mathbb{R}^{4\times 4}$. Using a modified Gram-Schmidt orthonormalization process [124, pp. 15 – 16], a nonsingular transformation matrix

$$T = \begin{bmatrix} C_c \\ \tilde{T} \end{bmatrix} = \begin{bmatrix} -64.81 & -213.12 & -1242.27 & -85.82 \\ 1263.3 & -10.933 & -63.729 & -4.4026 \\ 0 & 1246.9 & -212.89 & -14.707 \\ 0 & 0 & -87.182 & 1262.0 \end{bmatrix},$$

was found, that transforms the nominal controller (6.20), (6.28), into

$$\begin{aligned} \dot{\tilde{x}}_c &= \tilde{A}_c \tilde{x}_c + \tilde{B}_{cy} y = T A_c T^{-1} \tilde{x}_c + T B_{cy} y, \\ u_c &= \tilde{C}_c \tilde{x}_c = C_c T^{-1} \tilde{x}_c = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \tilde{x}_c, \end{aligned} \tag{6.29}$$

with the special form of $\tilde{C}_c$ (see also Section 5.5.1). Then from the closed-form expressions (A.7) (in Appendix A) for the GPAW-compensated controller *derived from the transformed nominal controller* (6.29), we see that $\Gamma = [\gamma_{ij}]$ appears only in the projection matrix, which, due to the special structure of $\tilde{C}_c$, can be simplified to

$$I - \frac{\Gamma \tilde{C}_c^{\mathrm{T}} \tilde{C}_c}{\tilde{C}_c \Gamma \tilde{C}_c^{\mathrm{T}}} = I - \frac{1}{\gamma_{11}} \begin{bmatrix} \gamma_{11} & 0 & 0 & 0 \\ \gamma_{21} & 0 & 0 & 0 \\ \gamma_{31} & 0 & 0 & 0 \\ \gamma_{41} & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -\frac{\gamma_{21}}{\gamma_{11}} & 1 & 0 & 0 \\ -\frac{\gamma_{31}}{\gamma_{11}} & 0 & 1 & 0 \\ -\frac{\gamma_{41}}{\gamma_{11}} & 0 & 0 & 1 \end{bmatrix} := \begin{bmatrix} 0 & 0 & 0 & 0 \\ \tilde{\gamma}_1 & 1 & 0 & 0 \\ \tilde{\gamma}_2 & 0 & 1 & 0 \\ \tilde{\gamma}_3 & 0 & 0 & 1 \end{bmatrix},$$

with $\tilde{\gamma}_1 := -\frac{\gamma_{21}}{\gamma_{11}}$, $\tilde{\gamma}_2 := -\frac{\gamma_{31}}{\gamma_{11}}$, and $\tilde{\gamma}_3 := -\frac{\gamma_{41}}{\gamma_{11}}$. Clearly, the resultant GPAW-compensated controller is fully defined by the 3 parameters in $\tilde{\gamma} := [\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\gamma}_3]^{\mathrm{T}}$. Note that the remaining elements of Γ can always be defined in a way to ensure Γ is symmetric positive definite, so that no restrictions need to be imposed on $\tilde{\gamma}$.

The GPAW-compensated controller derived from the transformed nominal controller (6.29) is given by

$$\begin{aligned} \dot{\tilde{x}}_g &= R_{\mathcal{I}^*}(\tilde{x}_g, y)(\tilde{A}_c \tilde{x}_g + \tilde{B}_{cy} y), \\ u_g &= \tilde{C}_c \tilde{x}_g. \end{aligned} \tag{6.30}$$

As before, we call the resultant closed-loop system comprising (6.19) and (6.30) with $u = u_g$ the GPAW-compensated system and denote it by Σ_g. To determine $\tilde{\gamma}$ (and hence Γ), we optimize (over $\tilde{\gamma}$) the time response of Σ_g to the disturbance input (see (6.19))

$$w = \begin{cases} 1.588, & \text{if } t \in [0, 0.01], \\ 0, & \text{otherwise.} \end{cases} \tag{6.31}$$

Specifically, we solve the *unconstrained* minimization problem

$$\min_{\tilde{\gamma} \in \mathbb{R}^3} F(\tilde{\gamma}) = \int_0^8 y_1^2(t)\, dt,$$

where $y(t) = [y_1(t), y_2(t)]^{\mathrm{T}}$ for $t \in [0, 8]$ is the plant output (dependent on Γ) obtained from simulation. The preceding unconstrained optimization problem was solved using the nonlinear program solver `fminunc` of the MATLAB® Optimization Toolbox [186], which yields a satisfactory solution

$$\tilde{\gamma} = \begin{bmatrix} -0.21853 & -4.9972 & -3.9991 \end{bmatrix}^{\mathrm{T}},$$

when started with the initial guess $\tilde{\gamma}_{ig} = [-0.2, -5, -4]^{\mathrm{T}}$.

Remark 6.4. Numerical experience indicates that the posed optimization problem is poorly scaled, i.e. the objective is highly sensitive to $\tilde{\gamma}_1$ while being relatively insensitive to $\tilde{\gamma}_2$ and $\tilde{\gamma}_3$. The optimization was solved repeatedly with numerous initial guesses, and the initial guess $\tilde{\gamma}_{ig} = [-0.2, -5, -4]^{\mathrm{T}}$ was found to yield a satisfactory solution. □

Numerical Results

The systems Σ_u, Σ_n, Σ_{aw}, and Σ_g are first simulated for the nominal disturbance input (6.31). The responses are shown in the left plot of Fig. 6-10. It can be seen that without anti-windup compensation, the response of system Σ_n is highly oscillatory with a large settling time of approximately 10 s. The response of the GPAW-compensated system Σ_g is comparable to that of the LMI-based anti-windup compensated system Σ_{aw}, although Σ_{aw} exhibits a marginally superior response. Both responses of systems Σ_g and

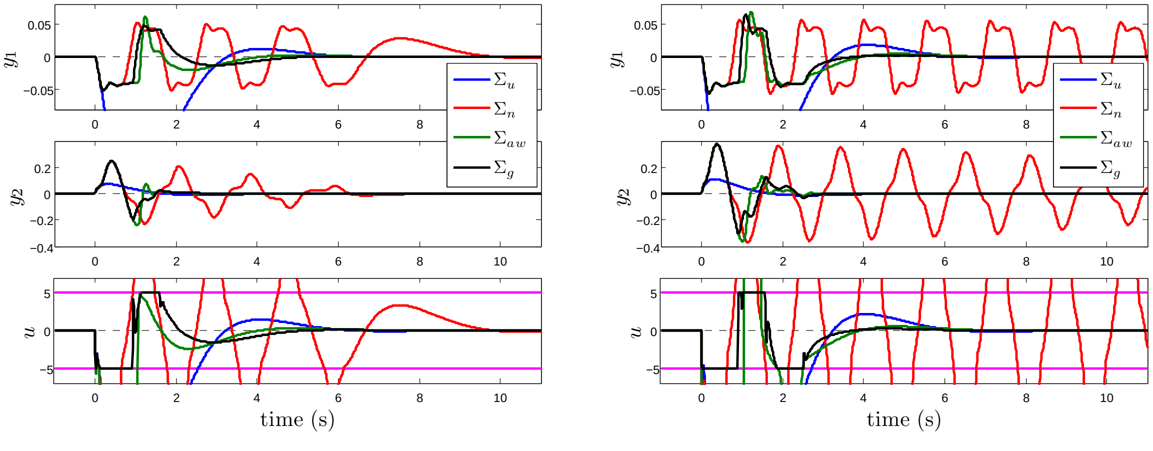


Figure 6-10: Response of cart-spring-pendulum system to disturbances. The left plot shows the case for nominal disturbance while the right plot shows the case where the nominal disturbance is magnified by 50%.

Σ_{aw} are significantly superior to the response of the uncompensated system Σ_n, reaching steady state at approximately 5 s. Observe also that controller state-output consistency (see Theorem 2.5.3) holds for the GPAW-compensated system Σ_g, but not for system Σ_{aw}.

Since the GPAW parameter Γ was determined by optimizing for a specific disturbance input, namely (6.31), we check its performance against a disturbance which is 50% larger in magnitude. The responses to the increased excitation are shown in the right plot of Fig. 6-10. Clearly, the GPAW-compensated system Σ_g exhibits satisfactory responses to disturbance inputs for which it was not optimized for.

6.4 Chapter Summary

We compared the GPAW scheme against three state-of-the-art anti-windup schemes using examples available from the literature. Because the stability results obtained thus far are too conservative to be applied, stability of the GPAW-compensated systems are not established. Even in the absence of stability results, ad hoc methods can be devised to design the GPAW-compensated controller. We showed that the GPAW scheme achieves comparable performance against these state-of-the-art anti-windup schemes in these examples. Where current stability results are not applicable, the GPAW scheme provides practitioners with a *candidate* anti-windup scheme where no candidates may be available otherwise.

Chapter 7

Conclusions and Future Work

Windup induced by control saturation remains one of the major problems affecting virtually all practical control systems, with adverse consequences. The *gradient projection anti-windup* (GPAW) scheme proposed in this dissertation is a general purpose anti-windup scheme constructed for saturated nonlinear systems driven by nonlinear controllers, a topic recognized as a largely open problem. The GPAW-compensated controller achieves *controller state-output consistency*, possesses clear geometric properties, and is characterized by a passive projection operator. It is defined by either the online solution to a combinatorial optimization subproblem, a convex quadratic program, or a projection onto a convex polyhedral cone problem. When the controller output has dimension one or two, closed-form expressions for the GPAW-compensated controller are available.

Strong results were obtained when GPAW compensation is applied to saturated first-order LTI plants driven by first-order LTI controllers. This simple system illustrates numerous attractive features of the GPAW scheme. For this system, GPAW compensation can only maintain/enlarge the ROA of the uncompensated system, a result *independent of any Lyapunov function.* Qualitative weaknesses of some existing anti-windup results were demonstrated. This motivates a new paradigm to address the anti-windup problem, where results *relative* to the uncompensated system are sought. Numerical results further suggest a need to consider *asymmetric* saturation constraints for general saturated systems, which has been largely ignored in the literature.

We derived some ROA comparison and stability results for GPAW-compensated MIMO nonlinear and MIMO LTI systems. These ROA comparison results represent the first steps consistent with the new anti-windup paradigm. We note that these ROA comparison results state explicit advantages of adopting GPAW compensation. This is in contrast to some existing anti-windup results where the purported advantages offered by the anti-windup scheme may be achieved by the *uncompensated* system.

Stability results obtained thus far are still fairly conservative. By means of non-trivial examples available in the literature, we showed that even in the absence of applicable stability results, ad hoc methods can still be devised to design GPAW-compensated controllers with performance comparable to some state-of-the-art anti-windup schemes. Where current stability results are not applicable, the general purpose GPAW scheme provides practitioners with a *candidate* anti-windup scheme endowed with some attractive properties.

The significance of the research presented herein has to be seen in a larger context. Consider the standard anti-windup structure depicted in Fig. 1-1. This anti-windup structure is essentially a generalization of that adopted in [187], which in turn is inspired by the back-calculation method [53]. Apart from the fact that this structure preserves the unconstrained response when the saturation constraints are not triggered,[1] it has no attractive inherent properties. All performance and stability properties of the anti-windup compensated system are intimately dictated by the anti-windup gains. Numerous anti-windup schemes have since been developed based on this anti-windup structure, differing only in the assumptions imposed and method of determining the anti-windup gains [14]. Even state-of-the-art anti-windup schemes for nonlinear systems like [24, 65] adopted variants of the standard anti-windup structure of Fig. 1-1.

In contrast to the standard anti-windup structure, the GPAW scheme has several additional "built-in" features (mentioned above) induced by the projection operator, and is defined by a single symmetric positive definite matrix parameter. In essence, the projection operator has endowed the GPAW-compensated controller much of its inherent properties, and the GPAW parameter is only meant to allow some "fine tuning".[2] The fact that the single unified structure of the GPAW scheme can achieve comparable performance as three state-of-the-art anti-windup schemes adopting variants of the standard anti-windup structure shows the versatility of the GPAW scheme. Moreover, these state-of-the-art anti-windup schemes cannot be applied to the simple system (4.42) used in Section 4.5.2 to demonstrate the application of the ROA comparison result, Theorem 4.4.1 (see Remark 4.11). These show the potential for the GPAW scheme to be developed into a truly general purpose anti-windup scheme with stability guarantees.

The research presented herein represents the first steps in the study of GPAW compensation. We invite the reader to join us in this quest to solve one of the most prevalent problems affecting control systems, in a journey that promises to be theoretically rich and practically important.

7.1 Future Work

Given the conservativeness of current stability results, there is a need to develop less conservative results for the GPAW scheme.[3] Apart from this obvious line of research, potential work that have been identified are listed below.

7.1.1 Robustness Issues due to Presence of Noise, Disturbances, Time Delays, and Unmodeled Dynamics

Apart from control saturation, practical control systems are adversely affected by the presence of noise, disturbances, time delays, and unmodeled dynamics. Recognizing that anti-windup compensation for nonlinear systems is still largely an open problem, we have chosen to ignore these important issues with regard to GPAW-compensated systems. Addressing these will become increasingly important as research in GPAW compensation matures.

[1] The state of the anti-windup compensator has to be appropriately initialized for this to hold.

[2] Indeed, for first order nominal controllers, the GPAW-compensated controller is *independent* of any GPAW parameters (see Remark B.1 in Appendix B).

[3] In this regard, we note that the GPAW parameter may be allowed to vary with the controller state, measurement, or reference input, and the objective function of the associated combinatorial optimization subproblem may be modified.

7.1.2 Robustness Issues due to Different Realizations of Anti-windup *Compensator* Derived from Anti-windup *Compensated Controller*

If an anti-windup *compensator* is needed, perhaps because the nominal controller cannot be modified, it was shown in Section 1.3.1 how it can be derived from an associated anti-windup *compensated controller*. There are numerous ways to do so, and each realization will vary in its robustness properties. One area of possible research is to study the methods to realize the anti-windup *compensator* with satisfactory robustness properties.

7.1.3 Verify if Reference Governors Solves General Anti-windup Problem

In Section 1.4.6, *reference governors* [23, 95–99] have been identified as a likely candidate to solve Problem 1 (see Section 1.3). At this point, due to the literature on reference governors adopting a different problem statement, it is uncertain whether it indeed solves Problem 1, or whether it can be modified such that it does so. There is a need to examine the literature on reference governors to confirm this.

7.1.4 Continuous Time Gradient Projection Method as a Valid Optimization Method

In Section 2.3, we extended Rosen's gradient projection method to continuous time with the intention of extracting only the resultant projection operator. We attempted to show that the extension is correct by means of a numerical example, without proving that it indeed solves the underlying optimization problem. Since continuous optimization is an active research field, study of the continuous time extension may contribute to this area.

7.1.5 Exploit Projection Operator

The projection operator was extracted from the continuous time gradient projection method in Section 2.4, for the purpose of constructing the GPAW-compensated controller. Clearly, it can be used for other purposes, e.g. to bound parameter estimates in adaptive control[4] as in [122, Sections 4.4, 8.4.2, and 8.5.5], and more general constrained control (see Section 2.8.3). Moreover, since the projection operator induces a "sliding" motion on the saturation constraint boundaries (e.g. see Figs. 2-2 and 3-4), it may possibly be used fruitfully in *sliding mode control* [38, Chapter 7, pp. 276 – 310].

7.1.6 GPAW-Compensated Systems in Relation to Projected Dynamical Systems

As shown in Section 3.2, the GPAW-compensated system obtained by applying GPAW compensation to the simple system considered is in fact a *projected dynamical system* (PDS) [107–110]. PDS is a significant line of independent research that has attracted the attention of mathematicians, physicists, and economists, among others. It is likely that GPAW-compensated systems are closely related to PDS. Establishing a concrete link between general GPAW-compensated systems and PDS will allow cross utilization of ideas and methods, among other strategic benefits. To this end, we note that results in [151] may be useful.

[4] In fact, the GPAW scheme was inspired by the projection mechanism used in adaptive control.

7.1.7 Passivity and Small-Gain Based Stability Results

We showed in Section 2.7 that the GPAW-compensated system is obtained by modifying the uncompensated system with a passive projection operator (with L_2-gain less than one) and two transformation matrices (see Fig. 2-5). Some possible research would be to exploit these properties to derive passivity and small-gain-based stability results for GPAW-compensated systems (see [37, Theorem 5.6, p. 218, Section 6.5, pp. 245 – 259]).

7.1.8 Show Performance Improvement and Prove Conjectures for Planar LTI Systems

In Section 3.8, we attempted to show that the GPAW scheme solves Problem 1 (in Section 1.3). As explained in Remark 3.20, verifying that performance is improved by GPAW compensation is a non-trivial task. This needs to be studied, as well as proving or disproving Conjectures 3.1 and 3.2. See possible hints following the statements of these conjectures.

7.1.9 Existence and Uniqueness of Solutions to General GPAW-Compensated Systems

Existence and uniqueness of solutions are fundamental properties required for a meaningful study of dynamic systems. For the planar LTI system studied in Chapter 3, these are assured by Proposition 3.3.3. For GPAW-compensated MIMO LTI systems that are equivalent to *linear systems with partial state constraints* (see Section 5.6), existence and uniqueness of solutions are established in [103]. We need to establish these properties for general GPAW-compensated systems. To this end, if GPAW-compensated systems can be shown to be equivalent to PDS [107–110], then [107, Theorem 2] can be used to assert existence and uniqueness of their solutions, as demonstrated in the proof of Proposition 3.3.3. A more general approach is to use results from the theory of *differential equations with discontinuous right-hand sides* [120]. In particular, results in [120, §2.7, pp. 75 – 86, and §2.10, pp. 106 – 117] may be useful.

7.1.10 Effects of Controller State Augmentation and Redundant Saturation Constraints

In Remark 2.23, we cautioned against indiscriminate controller state augmentation when attempting to obtain an approximation of the nominal controller. This is closely related to having *redundant* saturation constraints as discussed in Section 5.5.3. Observe that these issues are also present in the standard anti-windup framework of Fig. 1-1. However, to the best of our knowledge, these issues have not been addressed in the literature. These aspects are poorly understood and their study may allow more effective exploitation of available control authority.

Appendix A

Closed Form Expressions for Single-output GPAW-Compensated Controllers

Single-output controllers[1] abound in practice. Here, we show how closed-form expressions can be obtained for single-output GPAW-compensated controllers by analyzing the associated combinatorial optimization subproblem, which allows a highly efficient realization (see Section B.1 in Appendix B for computational results when the controller output is of dimension two). These demonstrations also show that such controllers reduces to those obtained by the conditional integration method (see Section 2.1). In fact, the hybrid nature of general GPAW-compensated controllers are made evident due to the switching structure exposed here (as well as in Appendix B). The closed-form expressions are first derived for single-output nonlinear controllers, and then specialized to LTI controllers. We conclude this appendix with the closed-form expressions for GPAW-compensated approximate PID controllers, since they represent an important subclass.

Consider the single-output nominal controller (2.26)

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c), \end{aligned} \tag{A.1}$$

with the restriction that u_c is scalar. Here, $(x_c, y, r) \in \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^{n_r}$ are of finite but otherwise arbitrary dimensions. For controllers of general structure (1.3), it was shown in Section 2.6 that an arbitrarily close approximating controller having the form of (A.1) can be constructed. Following the development in Section 2.5, the derived GPAW-compensated controller has the form (2.27)

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g), \end{aligned} \tag{A.2}$$

where $u_g \in \mathbb{R}$. The scalar saturation constraint $u_{\min} \leq u_g \leq u_{\max}$ is equivalent to the two constraints

$$h_1(x_g) = g_c(x_g) - u_{\max} \leq 0, \qquad h_2(x_g) = -g_c(x_g) + u_{\min} \leq 0,$$

[1] Clearly, single-output controllers include SISO controllers.

where the gradients of h_1 and h_2 are

$$\nabla h_1(x_g) = -\nabla h_2(x_g) = \nabla g_c(x_g).$$

Define the index set of active saturation constraints, and the candidate solution set[2]

$$\mathcal{I}_{\text{sat}} := \mathcal{I}_{\text{sat}}(x_g) = \{i \in \mathcal{I}_2 \mid h_i(x_g) \geq 0\}, \qquad \mathcal{J} := \{\mathcal{I} \subset \mathcal{I}_{\text{sat}} \mid |\mathcal{I}| \leq q\}.$$

Observe that the two constraints $h_1(x_g) \leq 0$, $h_2(x_g) \leq 0$, cannot be simultaneously active, so that $\mathcal{I}_{\text{sat}}$ can only be $\emptyset$, $\{1\}$, or $\{2\}$. Then the candidate solution set $\mathcal{J}$ for each of these cases maps according to[3]

$$\begin{aligned} (\mathcal{I}_{\text{sat}} = \emptyset) \Rightarrow (\mathcal{J} = \{\emptyset\}), \qquad\qquad & (\mathcal{I}_{\text{sat}} = \{1\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{1\}\}), \\ & (\mathcal{I}_{\text{sat}} = \{2\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{2\}\}). \end{aligned}$$

Since $\mathcal{I} \in \mathcal{J}$ implies $\mathcal{I} \in \{\emptyset, \{1\}, \{2\}\}$, we only need to define $N_{\mathcal{I}}(x_g)$ (2.29) as the $q \times 1$ matrix[4]

$$N_{\mathcal{I}}(x_g) = \begin{cases} \nabla h_1(x_g), & \text{if } \mathcal{I} = \{1\}, \\ \nabla h_2(x_g), & \text{if } \mathcal{I} = \{2\}, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{or} \qquad N_{\mathcal{I}}(x_g) = \begin{cases} \nabla g_c(x_g), & \text{if } \mathcal{I} = \{1\}, \\ -\nabla g_c(x_g), & \text{if } \mathcal{I} = \{2\}, \\ 0, & \text{otherwise}, \end{cases}$$

for any $\mathcal{I} \in \{\emptyset, \{1\}, \{2\}\}$. Choose some $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{q \times q}$ as the GPAW parameter. If $\nabla g_c(x_g) \neq 0 \in \mathbb{R}^q$ and $\mathcal{I} \in \{\emptyset, \{1\}, \{2\}\}$, the projection matrix (2.30) can be written as

$$R_{\mathcal{I}}(x_g) = \begin{cases} I - \frac{\Gamma \nabla g_c \nabla^{\mathrm{T}} g_c(x_g)}{\nabla^{\mathrm{T}} g_c \Gamma \nabla g_c(x_g)}, & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise}, \end{cases} \tag{A.3}$$

since $\Gamma N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}} \Gamma N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(x_g) = \Gamma \nabla g_c (\nabla^{\mathrm{T}} g_c \Gamma \nabla g_c)^{-1} \nabla^{\mathrm{T}} g_c(x_g) = \frac{\Gamma \nabla g_c \nabla^{\mathrm{T}} g_c(x_g)}{\nabla^{\mathrm{T}} g_c \Gamma \nabla g_c(x_g)}$, for $\mathcal{I} \in \{\{1\}, \{2\}\}$.

The single-output GPAW-compensated controller (A.2) is defined by (A.3) and a solution $\mathcal{I}^*$ to the combinatorial optimization subproblem (2.31), reproduced below

$$\begin{aligned} & \max_{\mathcal{I} \in \mathcal{J}} F(\mathcal{I}) = f_c^{\mathrm{T}}(x_g, y, r) \Gamma^{-1} R_{\mathcal{I}}(x_g) f_c(x_g, y, r), \\ & \text{subject to} \qquad \operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|, \\ & \qquad\qquad N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}}(x_g) R_{\mathcal{I}}(x_g) f_c(x_g, y, r) \leq 0. \end{aligned} \tag{A.4}$$

What follows is a detailed analysis of subproblem (A.4) to extract the closed form expresions for the single-output GPAW-compensated controller (A.2).

For notational convenience, we will drop all function arguments. Due to Proposition 2.5.2, $\mathcal{I} = \emptyset$ is an optimal solution whenever it is feasible (see also Remark 2.17). Clearly, $\mathcal{I} = \emptyset$ satisfies the first constraint of subproblem (A.4). It will be infeasible only when it fails to satisfy the second constraint. This happens only when

(i) $\mathcal{I}_{\text{sat}} = \{1\}$ and $N_{\mathcal{I}_{\text{sat}}}^{\mathrm{T}} R_{\emptyset} f_c = \nabla^{\mathrm{T}} g_c f_c > 0$; or

[2] Recall that $\mathcal{I}_i := \{1, 2, \ldots, i\}$ for any positive integer i.

[3] Note that $\Rightarrow$ denotes logical implication.

[4] Notice that the bijection $\sigma_{\mathcal{I}}$ described in Remark 2.5 is not needed for this case.

(ii) $\mathcal{I}_{\mathrm{sat}} = \{2\}$ and $N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}} R_{\emptyset} f_c = -\nabla^{\mathrm{T}} g_c f_c > 0$.

When condition (i) holds, we have $\nabla^{\mathrm{T}} g_c \neq 0$ (due to $\nabla^{\mathrm{T}} g_c f_c \neq 0$), which implies the first constraint of subproblem (A.4) is satisfied for $\mathcal{I} = \{1\}$. From (A.3), it can be verified that $N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}} R_{\mathcal{I}} = 0$ for $\mathcal{I} = \{1\}$, so that it is a feasible solution. Hence when condition (i) holds, the unique optimal solution is $\mathcal{I}^* = \{1\}$, since $\mathcal{J} = \{\emptyset, \{1\}\}$ admits only $\{1\}$ as the single feasible solution. In similar manner, it can be verified that when condition (ii) holds, the unique optimal solution is $\mathcal{I}^* = \{2\}$. For all other cases, an optimal solution[5] is $\mathcal{I}^* = \emptyset$.

Using (A.3), we collect these conditions together to write the single-output GPAW-compensated controller (A.2) as[6]

$$\dot{x}_g = \begin{cases} \left(I - \frac{\Gamma \nabla g_c \nabla^{\mathrm{T}} g_c(x_g)}{\nabla^{\mathrm{T}} g_c \Gamma \nabla g_c(x_g)}\right) f_c(x_g, y, r), & \text{if } (u_g \geq u_{\max}) \wedge \left(\nabla^{\mathrm{T}} g_c(x_g) f_c(x_g, y, r) > 0\right), \\ \left(I - \frac{\Gamma \nabla g_c \nabla^{\mathrm{T}} g_c(x_g)}{\nabla^{\mathrm{T}} g_c \Gamma \nabla g_c(x_g)}\right) f_c(x_g, y, r), & \text{if } (u_g \leq u_{\min}) \wedge \left(\nabla^{\mathrm{T}} g_c(x_g) f_c(x_g, y, r) < 0\right), \\ f_c(x_g, y, r), & \text{otherwise}, \end{cases} \tag{A.5}$$
$$u_g = g_c(x_g), \qquad x_g(0) = x_{c0}.$$

The preceding is the desired closed-form expressions for the single-output GPAW-compensated controller, which shows clearly that in general, GPAW-compensated controllers are *switched* controllers. The following sections specialize (A.5) for some common classes of nominal controllers.

Remark A.1. Observe the similarities between (A.5) and (2.2), especially the conditions for switching the state update. Note also that the projection matrix $\left(I - \frac{\Gamma \nabla g_c \nabla^{\mathrm{T}} g_c(x_g)}{\nabla^{\mathrm{T}} g_c \Gamma \nabla g_c(x_g)}\right)$ is similar to that in [122, Appendix B.4, pp. 788 – 791]. □

A.1 Output Equation Linear in States

Now, consider when the output equation of the nominal controller (A.1) is linear in the states, i.e.

$$u_c = g_c(x_c) = c_c^{\mathrm{T}} x_c,$$

where $c_c \in \mathbb{R}^q$ is a constant vector (which will be non-zero for any meaningful controller). Then the *constant* gradient of g_c is $\nabla g_c = c_c$, and (A.5) reduces to

$$\dot{x}_g = \begin{cases} \left(I - \frac{1}{c_c^{\mathrm{T}} \Gamma c_c} \Gamma c_c c_c^{\mathrm{T}}\right) f_c(x_g, y, r), & \text{if } (u_g \geq u_{\max}) \wedge \left(c_c^{\mathrm{T}} f_c(x_g, y, r) > 0\right), \\ \left(I - \frac{1}{c_c^{\mathrm{T}} \Gamma c_c} \Gamma c_c c_c^{\mathrm{T}}\right) f_c(x_g, y, r), & \text{if } (u_g \leq u_{\min}) \wedge \left(c_c^{\mathrm{T}} f_c(x_g, y, r) < 0\right), \\ f_c(x_g, y, r), & \text{otherwise}, \end{cases} \tag{A.6}$$
$$u_g = c_c^{\mathrm{T}} x_g, \qquad x_g(0) = x_{c0}.$$

The *constant* projection matrix $\left(I - \frac{1}{c_c^{\mathrm{T}} \Gamma c_c} \Gamma c_c c_c^{\mathrm{T}}\right)$ can be computed offline to reduce online computational demands.

The final specialization is for LTI nominal controllers, so that in addition, the vector

[5] No claim has been made here that $\mathcal{I}^* = \emptyset$ is the unique optimal solution, but only that it is one of possibly multiple optimal solutions. In particular, observe that when $f_c = 0$ (e.g. with $\mathcal{I}_{\mathrm{sat}} = \{1\}$ and $\nabla g_c \neq 0$), any $\mathcal{I} \in \mathcal{J}$ must be an optimal solution.

[6] Note that $\wedge$ denotes the logical AND operator.

field of the nominal controller (A.1) is

$$f_c(x_c, y, r) = A_c x_c + B_{cy} y + B_{cr} r.$$

This reduces (A.6) to

$$\dot{x}_g = \begin{cases} \big(I - \frac{1}{c_c^{\mathrm{T}} \Gamma c_c} \Gamma c_c c_c^{\mathrm{T}}\big)(A_c x_g + B_{cy} y + B_{cr} r), & \text{if } A_{\max} \vee A_{\min}, \\ A_c x_g + B_{cy} y + B_{cr} r, & \text{otherwise,} \end{cases} \tag{A.7}$$
$$u_g = c_c^{\mathrm{T}} x_g, \qquad x_g(0) = x_{c0},$$

where $\vee$ is the logical OR operator, and the logical statements $A_{\max}$ and $A_{\min}$ are

$$\begin{aligned} A_{\max} &= (u_g \geq u_{\max}) \wedge \big(c_c^{\mathrm{T}}(A_c x_g + B_{cy} y + B_{cr} r) > 0\big), \\ A_{\min} &= (u_g \leq u_{\min}) \wedge \big(c_c^{\mathrm{T}}(A_c x_g + B_{cy} y + B_{cr} r) < 0\big). \end{aligned}$$

The closed-form expressions in (A.7) represent single-output GPAW-compensated LTI controllers.

A.2 PID Controllers

PID controllers are among the most utilized controllers [188, p. 1], as evidenced by the numerous books devoted to its theory and applications [17, 188–192]. Since they represent an important subclass, we present the closed-form expressions for GPAW-compensated *approximate* PID controllers here. Specializing the obtained expressions to proportional-integral controllers is straightforward, which we omit. Consider the PID controller

$$\begin{aligned} \dot{e}_i &= e, & e_i(0) &= e_{i0}, \\ u_c &= K_p e + K_i e_i + K_d \dot{e}, \end{aligned} \tag{A.8}$$

where e is the input error signal, and (K_p, K_i, K_d) are the proportional, integral, and derivative gains respectively. Here, e_i is the scalar controller state, and $(e, \dot{e})$ are the measurement inputs. It can be seen that the output equation $u_c = K_p e + K_i e_i + K_d \dot{e}$ depends on the measurement, and hence is not of the form of (A.1).

To obtain an approximate PID controller having the structure of (A.1), we use the construction in Section 2.6. First, define the exponentially stable, unity DC-gain low-pass filter[7] parameterized by $a \in (0, \infty)$

$$\begin{aligned} \dot{e}_1 &= a(e - e_1), & e_1(0) &= e(0), \\ \dot{e}_2 &= a(\dot{e} - e_2), & e_2(0) &= \dot{e}(0). \end{aligned}$$

We see that (e_1, e_2) can approximate $(e, \dot{e})$ arbitrarily well as $a \to \infty$. By replacing $(e, \dot{e})$ with their approximations (e_1, e_2) in the output equation, we obtain the *approximate* PID

[7] Note that in constructing the low-pass filter, we have not exploited the known relation between e and its derivative $\dot{e}$. Exploiting this relation may yield an observer/filter with superior properties.

controller

$$\begin{aligned} \dot{e}_i &= e, & e_i(0) &= e_{i0}, \\ \dot{e}_1 &= a(e - e_1), & e_1(0) &= e(0), \\ \dot{e}_2 &= a(\dot{e} - e_2), & e_2(0) &= \dot{e}(0), \\ u_c &= K_p e_1 + K_i e_i + K_d e_2. \end{aligned}$$

Defining the augmented state $x_c := [e_i, e_1, e_2]^{\mathrm{T}}$, the preceding can be written as

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_{cy} y, \qquad x_c(0) = x_{c0}, \\ u_c &= c_c^{\mathrm{T}} x_c, \end{aligned} \tag{A.9}$$

where $y := [e, \dot{e}]^{\mathrm{T}}$, and

$$A_c = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -a & 0 \\ 0 & 0 & -a \end{bmatrix}, \qquad B_{cy} = \begin{bmatrix} 1 & 0 \\ a & 0 \\ 0 & a \end{bmatrix}, \qquad c_c = \begin{bmatrix} K_i \\ K_p \\ K_d \end{bmatrix}, \qquad x_{c0} = \begin{bmatrix} e_{i0} \\ e(0) \\ \dot{e}(0) \end{bmatrix}.$$

It was shown in Section 2.6 that the solution of the approximate PID controller (A.9) can be made arbitrarily close to the solution of the exact PID controller (A.8) by choosing $a > 0$ sufficiently large.

The closed-form expressions for the GPAW-compensated approximate PID controller is then given by (A.7), which simplifies to

$$\begin{aligned} \dot{x}_g &= \begin{cases} \left(I - \frac{\Gamma c_c c_c^{\mathrm{T}}}{c_c^{\mathrm{T}} \Gamma c_c}\right)(A_c x_g + B_{cy} y), & \text{if } (u_g \geq u_{\max}) \wedge \left(c_c^{\mathrm{T}}(A_c x_g + B_{cy} y) > 0\right), \\ \left(I - \frac{\Gamma c_c c_c^{\mathrm{T}}}{c_c^{\mathrm{T}} \Gamma c_c}\right)(A_c x_g + B_{cy} y), & \text{if } (u_g \leq u_{\min}) \wedge \left(c_c^{\mathrm{T}}(A_c x_g + B_{cy} y) < 0\right), \\ A_c x_g + B_{cy} y, & \text{otherwise}, \end{cases} \\ u_g &= c_c^{\mathrm{T}} x_g, \qquad x_g(0) = x_{c0}. \end{aligned} \tag{A.10}$$

Observe that there are two parameters introduced, $a \in (0, \infty)$ chosen sufficiently large for a good approximation of the nominal response, and the GPAW parameter $\Gamma = \Gamma^{\mathrm{T}} > 0 \in \mathbb{R}^{3 \times 3}$. Comparing (A.10) with the PID controller under conditional integration (2.1), we see that the approximation increases the controller order by two, among other differences. The GPAW scheme exploits the augmented state to ensure controller state-output consistency (see Theorem 2.5.3), not achieved for the conditionally integrated controller (2.1).

Appendix B

Closed Form Expressions for GPAW-Compensated Controllers with Output of Dimension Two

The development here is similar to that in the first part of Appendix A, but for the case of nominal controllers with output of dimension two. Consider the nominal controller (2.26)

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c), \end{aligned} \tag{B.1}$$

with the restriction that $u_c \in \mathbb{R}^2$ is of dimension two. Here, as in Appendix A, $(x_c, y, r) \in \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^{n_r}$ are of finite but otherwise arbitrary dimensions. Following the development in Section 2.5, the derived GPAW-compensated controller has the form (2.27)

$$\begin{aligned} \dot{x}_g &= R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r), \qquad x_g(0) = x_{c0}, \\ u_g &= g_c(x_g), \end{aligned} \tag{B.2}$$

where $u_g \in \mathbb{R}^2$. Let g_c be decomposed into its elements $g_c = [g_{c1}, g_{c2}]^{\mathrm{T}}$. The vector saturation function[1] (1.2)

$$\mathrm{sat}(u) = [\rho_1(u_1), \rho_2(u_2)]^{\mathrm{T}}, \qquad \rho_i(u_i) = \max\{\min\{u_i, u_{\max,i}\}, u_{\min,i}\}, \qquad \forall i \in \mathcal{I}_2,$$

induces 4 constraints

$$h_i(x_g) = g_{ci}(x_g) - u_{\max,i} \le 0, \qquad h_{i+2}(x_g) = -g_{ci}(x_g) + u_{\min,i} \le 0, \qquad \forall i \in \mathcal{I}_2,$$

where the gradients of h_i are

$$\nabla h_i(x_g) = -\nabla h_{i+2}(x_g) = \nabla g_{ci}(x_g), \qquad \forall i \in \mathcal{I}_2.$$

For any index set $\mathcal{I} \subset \mathcal{I}_4$, define the $q \times \max\{|\mathcal{I}|, 1\}$ matrix (2.29)

$$N_{\mathcal{I}}(x_g) = \begin{cases} [\nabla h_{\sigma_{\mathcal{I}}(1)}(x_g), \nabla h_{\sigma_{\mathcal{I}}(2)}(x_g), \ldots, \nabla h_{\sigma_{\mathcal{I}}(|\mathcal{I}|)}(x_g)], & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases}$$

[1] Recall that $\mathcal{I}_i := \{1, 2, \ldots, i\}$ for any positive integer i.

where $\sigma_{\mathcal{I}}$ is a chosen (non-unique) bijection described in Remark 2.5. For any $\mathcal{I} \subset \mathcal{I}_4$ such that $\mathrm{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$, define the projection matrix (2.30)

$$R_{\mathcal{I}}(x_g) = \begin{cases} I - \Gamma N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}} \Gamma N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ I, & \text{otherwise,} \end{cases} \tag{B.3}$$

where $\Gamma = \Gamma^{\mathrm{T}} > 0$ is the GPAW parameter.

Define the index set of active saturation constraints, and the candidate solution set

$$\mathcal{I}_{\mathrm{sat}} := \mathcal{I}_{\mathrm{sat}}(x_g) = \{i \in \mathcal{I}_4 \mid h_i(x_g) \geq 0\}, \qquad \mathcal{J} := \{\mathcal{I} \subset \mathcal{I}_{\mathrm{sat}} \mid |\mathcal{I}| \leq q\}.$$

The GPAW-compensated controller (B.2) is defined by (B.3) and a solution $\mathcal{I}^*$ to the combinatorial optimization subproblem (2.31)

$$\begin{aligned} &\max_{\mathcal{I} \in \mathcal{J}} F(\mathcal{I}) = f_c^{\mathrm{T}}(x_g, y, r)\Gamma^{-1} R_{\mathcal{I}}(x_g) f_c(x_g, y, r), \\ &\text{subject to} \qquad\qquad \mathrm{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|, \\ &\qquad\qquad\qquad N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}(x_g) R_{\mathcal{I}}(x_g) f_c(x_g, y, r) \leq 0. \end{aligned} \tag{B.4}$$

What follows is a detailed analysis of subproblem (B.4) to extract the closed form expressions for the GPAW-compensated controller (B.2) with two dimensional output, which is presented as (B.5).

First, observe that the constraints $h_1(x_g) \leq 0$ and $h_3(x_g) \leq 0$ cannot be simultaneously active. Similarly, the constraints $h_2(x_g) \leq 0$ and $h_4(x_g) \leq 0$ cannot be simultaneously active. Then all possible combinations of $\mathcal{I}_{\mathrm{sat}}$ and the associated candidate solution set $\mathcal{J}$ maps according to[2]

$$\begin{aligned} &(\mathcal{I}_{\mathrm{sat}} = \emptyset) \Rightarrow (\mathcal{J} = \{\emptyset\}), \\ &(\mathcal{I}_{\mathrm{sat}} = \{1\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{1\}\}), && (\mathcal{I}_{\mathrm{sat}} = \{1,2\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}), \\ &(\mathcal{I}_{\mathrm{sat}} = \{2\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{2\}\}), && (\mathcal{I}_{\mathrm{sat}} = \{1,4\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{1\}, \{4\}, \{1,4\}\}), \\ &(\mathcal{I}_{\mathrm{sat}} = \{3\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{3\}\}), && (\mathcal{I}_{\mathrm{sat}} = \{3,2\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{3\}, \{2\}, \{3,2\}\}), \\ &(\mathcal{I}_{\mathrm{sat}} = \{4\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{4\}\}), && (\mathcal{I}_{\mathrm{sat}} = \{3,4\}) \Rightarrow (\mathcal{J} = \{\emptyset, \{3\}, \{4\}, \{3,4\}\}). \end{aligned}$$

Here, we have assumed that the controller is at least second order, i.e. $q \geq 2$. The case of first order controllers,[3] i.e $q = 1$, is treated in Section B.2.

Consider when $\mathcal{I}_{\mathrm{sat}} = \{1,2\}$, so that $\mathcal{J} = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$. With the assumption that $\mathrm{rank}(N_{\mathcal{I}_{\mathrm{sat}}}(x_g)) = 2$, it can be verified that $\mathrm{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$ holds for any $\mathcal{I} \in \mathcal{J}$, and that $\mathcal{I} = \{1,2\}$ is always a feasible solution to subproblem (B.4) (due to $N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}(x_g) R_{\mathcal{I}}(x_g) = N_{\mathcal{I}}^{\mathrm{T}}(x_g) R_{\mathcal{I}}(x_g) = 0$ for $\mathcal{I} = \{1,2\}$). Proposition 2.5.2 shows that the objective function of subproblem (B.4) satisfies the two chained inequalities

$$F(\emptyset) \geq F(\{1\}) \geq F(\{1,2\}), \qquad F(\emptyset) \geq F(\{2\}) \geq F(\{1,2\}).$$

These inequalities show that:

- $\emptyset$ is an optimal solution whenever it is feasible;

[2] Note that $\Rightarrow$ denotes logical implication.

[3] It is clear that any meaningful controller of the form (B.1) must be at least first order, i.e. they cannot be static memoryless controllers.

- $\{1\}$ is an optimal solution when $\emptyset$ is infeasible, $\{1\}$ is feasible, and either
  - $\{2\}$ is infeasible; or
  - $\{2\}$ is feasible and $F(\{1\}) \geq F(\{2\})$;
- $\{2\}$ is an optimal solution when $\emptyset$ is infeasible, $\{2\}$ is feasible, and either
  - $\{1\}$ is infeasible; or
  - $\{1\}$ is feasible and $F(\{2\}) \geq F(\{1\})$;
- for all other cases, $\{1, 2\}$ must be an optimal solution due to Proposition 2.5.1 and the fact that it is the only remaining option (that as mentioned, is always feasible).

Observe that each of the two sets of sub-conditions can be simplified, so that the conditions become

- $\emptyset$ is an optimal solution whenever it is feasible;
- $\{1\}$ is an optimal solution when $\emptyset$ is infeasible, $\{1\}$ is feasible, and
  - if $\{2\}$ is feasible, then $F(\{1\}) \geq F(\{2\})$;
- $\{2\}$ is an optimal solution when $\emptyset$ is infeasible, $\{2\}$ is feasible, and
  - if $\{1\}$ is feasible, then $F(\{2\}) \geq F(\{1\})$;
- for all other cases, $\{1, 2\}$ must be an optimal solution.

We can write the preceding conditions in an if-then-else structure to avoid repeating some conditions as follows

- $\emptyset$ is an optimal solution whenever it is feasible;
- otherwise, $\{1\}$ is an optimal solution when it is feasible and, $\{2\}$ is feasible implies $F(\{1\}) \geq F(\{2\})$;
- otherwise, $\{2\}$ is an optimal solution when it is feasible;
- for all other cases, $\{1, 2\}$ must be an optimal solution.

Observe that if the third condition holds, i.e. $\{2\}$ is feasible, exclusion of the first two conditions imply $\emptyset$ is infeasible, and if $\{1\}$ is feasible, then $F(\{1\}) < F(\{2\})$.

The previous discussion is for the specialized case of $\mathcal{I}_{\text{sat}} = \{1, 2\}$, which we generalize next. Given an *arbitrary* $\mathcal{I}_{\text{sat}}$, the preceding if-then-else conditions translate to

(i) $\emptyset$ is an optimal solution whenever it is feasible;

(ii) otherwise, if $\{1, 3\} \cap \mathcal{I}_{\text{sat}} \neq \emptyset$ and $\mathcal{I} = \{1, 3\} \cap \mathcal{I}_{\text{sat}}$ is feasible, then $\mathcal{I}$ $(\in \{\{1\}, \{3\}\})$ is an optimal solution when
   (a) $2 \in \mathcal{I}_{\text{sat}}$ and $\{2\}$ is feasible, imply $F(\mathcal{I}) \geq F(\{2\})$; and
   (b) $4 \in \mathcal{I}_{\text{sat}}$ and $\{4\}$ is feasible, imply $F(\mathcal{I}) \geq F(\{4\})$;

(iii) otherwise, if $\{2, 4\} \cap \mathcal{I}_{\text{sat}} \neq \emptyset$ and $\mathcal{I} = \{2, 4\} \cap \mathcal{I}_{\text{sat}}$ is feasible, then $\mathcal{I}$ $(\in \{\{2\}, \{4\}\})$ is an optimal solution;

(iv) for all other cases, $\mathcal{I} = \mathcal{I}_{\text{sat}}$ must be an optimal solution.

We can express the preceding conditions in the form[4]

$$\mathcal{I}^* = \begin{cases} \emptyset, & \text{if } A_0, \\ \{1, 3\} \cap \mathcal{I}_{\text{sat}}, & \text{if } \neg A_0 \wedge A_1, \\ \{2, 4\} \cap \mathcal{I}_{\text{sat}}, & \text{if } \neg A_0 \wedge \neg A_1 \wedge A_2, \\ \mathcal{I}_{\text{sat}}, & \text{otherwise,} \end{cases}$$

where $\mathcal{I}^*$ is an optimal solution to subproblem (B.4) and the logical statements A_0, A_1, A_2

[4] Note that $\neg$ and $\wedge$ denote the logical NOT and logical AND operators respectively.

map directly to conditions (i), (ii), (iii) respectively.

For notational convenience, we will drop all function arguments. Observe that for any $\mathcal{I} \subset \mathcal{I}_{\text{sat}}$ to be feasible, we need $\operatorname{rank}(N_{\mathcal{I}}) = |\mathcal{I}|$ and $N^{\mathrm{T}}_{\mathcal{I}_{\text{sat}}} R_{\mathcal{I}} f_c \leq 0$ to hold. Since $N^{\mathrm{T}}_{\mathcal{I}} R_{\mathcal{I}} = 0$ for any well defined $R_{\mathcal{I}}$, feasibility requires only that $N^{\mathrm{T}}_{\mathcal{I}_{\text{sat}} \setminus \mathcal{I}} R_{\mathcal{I}} f_c \leq 0$, i.e. only those indices in $\mathcal{I}_{\text{sat}} \setminus \mathcal{I}$ need to be checked. Recall the basic logic operations [167, Fig. 7.11, p. 210] that will be used in the sequel[5]

$$\begin{aligned} &(A \Rightarrow B) \Leftrightarrow (\neg A \vee B), \\ &\neg(A \wedge B) \Leftrightarrow (\neg A \vee \neg B), \qquad (A \wedge (B \vee C)) \Leftrightarrow ((A \wedge B) \vee (A \wedge C)), \\ &\neg(A \vee B) \Leftrightarrow (\neg A \wedge \neg B), \qquad (A \vee (B \wedge C)) \Leftrightarrow ((A \vee B) \wedge (A \vee C)). \end{aligned}$$

Examination of condition (i) shows that the logical statement A_0, stating the feasibility conditions for $\mathcal{I}^* = \emptyset$, can be written concretely as[6]

$$\begin{aligned} A_0 \Leftrightarrow\ & (1 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_1 R_{\emptyset} f_c \leq 0) \wedge (3 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_3 R_{\emptyset} f_c \leq 0) \\ & \quad \wedge (2 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_2 R_{\emptyset} f_c \leq 0) \wedge (4 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_4 R_{\emptyset} f_c \leq 0), \\ \Leftrightarrow\ & (g_{c1} \geq u_{\max,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} f_c \leq 0) \wedge (g_{c1} \leq u_{\min,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} f_c \geq 0) \\ & \quad \wedge (g_{c2} \geq u_{\max,2} \Rightarrow \nabla^{\mathrm{T}} g_{c2} f_c \leq 0) \wedge (g_{c2} \leq u_{\min,2} \Rightarrow \nabla^{\mathrm{T}} g_{c2} f_c \geq 0), \\ \Leftrightarrow\ & \big(\neg(g_{c1} \geq u_{\max,1}) \vee \nabla^{\mathrm{T}} g_{c1} f_c \leq 0\big) \wedge \big(\neg(g_{c1} \leq u_{\min,1}) \vee \nabla^{\mathrm{T}} g_{c1} f_c \geq 0\big) \\ & \quad \wedge \big(\neg(g_{c2} \geq u_{\max,2}) \vee \nabla^{\mathrm{T}} g_{c2} f_c \leq 0\big) \wedge \big(\neg(g_{c2} \leq u_{\min,2}) \vee \nabla^{\mathrm{T}} g_{c2} f_c \geq 0\big), \\ \Leftrightarrow\ & (g_{c1} < u_{\max,1} \vee \nabla^{\mathrm{T}} g_{c1} f_c \leq 0) \wedge (g_{c1} > u_{\min,1} \vee \nabla^{\mathrm{T}} g_{c1} f_c \geq 0) \\ & \quad \wedge (g_{c2} < u_{\max,2} \vee \nabla^{\mathrm{T}} g_{c2} f_c \leq 0) \wedge (g_{c2} > u_{\min,2} \vee \nabla^{\mathrm{T}} g_{c2} f_c \geq 0). \end{aligned}$$

Similarly, examination of condition (ii) shows that the logical statement A_1 can be written as

$$\begin{aligned} A_1 \Leftrightarrow \Big(& 1 \in \mathcal{I}_{\text{sat}} \wedge \operatorname{rank}(N_{\{1\}}) = 1 \\ & \wedge (2 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_2 R_{\{1\}} f_c \leq 0) \wedge (4 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_4 R_{\{1\}} f_c \leq 0) \\ & \wedge \big((2 \in \mathcal{I}_{\text{sat}} \wedge \operatorname{rank}(N_{\{2\}}) = 1 \wedge \nabla^{\mathrm{T}} h_1 R_{\{2\}} f_c \leq 0) \Rightarrow F(\{1\}) \geq F(\{2\})\big) \\ & \wedge \big((4 \in \mathcal{I}_{\text{sat}} \wedge \operatorname{rank}(N_{\{4\}}) = 1 \wedge \nabla^{\mathrm{T}} h_1 R_{\{4\}} f_c \leq 0) \Rightarrow F(\{1\}) \geq F(\{4\})\big)\Big) \\ \vee \Big(& 3 \in \mathcal{I}_{\text{sat}} \wedge \operatorname{rank}(N_{\{3\}}) = 1 \\ & \wedge (2 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_2 R_{\{3\}} f_c \leq 0) \wedge (4 \in \mathcal{I}_{\text{sat}} \Rightarrow \nabla^{\mathrm{T}} h_4 R_{\{3\}} f_c \leq 0) \\ & \wedge \big((2 \in \mathcal{I}_{\text{sat}} \wedge \operatorname{rank}(N_{\{2\}}) = 1 \wedge \nabla^{\mathrm{T}} h_3 R_{\{2\}} f_c \leq 0) \Rightarrow F(\{3\}) \geq F(\{2\})\big) \\ & \wedge \big((4 \in \mathcal{I}_{\text{sat}} \wedge \operatorname{rank}(N_{\{4\}}) = 1 \wedge \nabla^{\mathrm{T}} h_3 R_{\{4\}} f_c \leq 0) \Rightarrow F(\{3\}) \geq F(\{4\})\big)\Big). \end{aligned}$$

Recognizing that $R_{\{1\}} = R_{\{3\}}$ and $R_{\{2\}} = R_{\{4\}}$ (see (B.3)), which implies $F(\{1\}) = F(\{3\})$ and $F(\{2\}) = F(\{4\})$ (see (B.4)), define

$$\begin{aligned} R_1 &:= R_{\{1\}} = R_{\{3\}}, \qquad & F_1 &:= F(\{1\}) = F(\{3\}), \\ R_2 &:= R_{\{2\}} = R_{\{4\}}, \qquad & F_2 &:= F(\{2\}) = F(\{4\}). \end{aligned}$$

[5] Note that $\vee$ denotes the logical OR operator, and $\Leftrightarrow$ denotes logical equivalence.

[6] Observe that $\operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$ holds trivially when $\mathcal{I} = \emptyset$.

Furthermore, we have $\operatorname{rank}(N_{\{1\}}) = \operatorname{rank}(N_{\{3\}})$ and $\operatorname{rank}(N_{\{2\}}) = \operatorname{rank}(N_{\{4\}})$. To simplify the subsequent manipulations of logic statements, let

$$\begin{gathered}
\alpha_1 \Leftrightarrow (1 \in \mathcal{I}_{\text{sat}}) \Leftrightarrow (g_{c1} \geq u_{\max,1}), \qquad \alpha_2 \Leftrightarrow (2 \in \mathcal{I}_{\text{sat}}) \Leftrightarrow (g_{c2} \geq u_{\max,2}),\\
\alpha_3 \Leftrightarrow (3 \in \mathcal{I}_{\text{sat}}) \Leftrightarrow (g_{c1} \leq u_{\min,1}), \qquad \alpha_4 \Leftrightarrow (4 \in \mathcal{I}_{\text{sat}}) \Leftrightarrow (g_{c2} \leq u_{\min,2}),\\
\beta_1 \Leftrightarrow (\operatorname{rank}(N_{\{1\}}) = \operatorname{rank}(N_{\{3\}}) = 1) \Leftrightarrow (\nabla g_{c1} \neq 0),\\
\beta_2 \Leftrightarrow (\operatorname{rank}(N_{\{2\}}) = \operatorname{rank}(N_{\{4\}}) = 1) \Leftrightarrow (\nabla g_{c2} \neq 0),\\
\delta_{12} \Leftrightarrow (\nabla^{\mathrm{T}} h_1 R_{\{2\}} f_c = \nabla^{\mathrm{T}} h_1 R_{\{4\}} f_c = \nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0),\\
\delta_{32} \Leftrightarrow (\nabla^{\mathrm{T}} h_3 R_{\{2\}} f_c = \nabla^{\mathrm{T}} h_3 R_{\{4\}} f_c = -\nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0) \Leftrightarrow (\nabla^{\mathrm{T}} g_{c1} R_2 f_c \geq 0),\\
\delta_{21} \Leftrightarrow (\nabla^{\mathrm{T}} h_2 R_{\{1\}} f_c = \nabla^{\mathrm{T}} h_2 R_{\{3\}} f_c = \nabla^{\mathrm{T}} g_{c2} R_1 f_c \leq 0),\\
\delta_{41} \Leftrightarrow (\nabla^{\mathrm{T}} h_4 R_{\{1\}} f_c = \nabla^{\mathrm{T}} h_4 R_{\{3\}} f_c = -\nabla^{\mathrm{T}} g_{c2} R_1 f_c \leq 0) \Leftrightarrow (\nabla^{\mathrm{T}} g_{c2} R_1 f_c \geq 0),\\
\eta \Leftrightarrow (F(\{1\}) = F(\{3\}) = F_1 \geq F_2 = F(\{2\}) = F(\{4\})).
\end{gathered}$$

Since $u_{\max,1} > u_{\min,1}$ and $u_{\max,2} > u_{\min,2}$, we have

$$\alpha_1 \Rightarrow \neg\alpha_3, \qquad \alpha_3 \Rightarrow \neg\alpha_1, \qquad \alpha_2 \Rightarrow \neg\alpha_4, \qquad \alpha_4 \Rightarrow \neg\alpha_2,$$

which gives $(\alpha_1 \wedge (\alpha_1 \vee \alpha_3)) \Leftrightarrow ((\alpha_1 \wedge \alpha_1) \vee (\alpha_1 \wedge \alpha_3)) \Leftrightarrow (\alpha_1 \vee \mathit{false}) \Leftrightarrow \alpha_1$. Similar operations yield

$$\alpha_1 \Leftrightarrow \alpha_1 \wedge (\alpha_1 \vee \alpha_3), \qquad \alpha_3 \Leftrightarrow \alpha_3 \wedge (\alpha_1 \vee \alpha_3), \qquad \alpha_2 \Leftrightarrow \alpha_2 \wedge (\alpha_2 \vee \alpha_4), \qquad \alpha_4 \Leftrightarrow \alpha_4 \wedge (\alpha_2 \vee \alpha_4).$$

With these, A_1 reduces according to

$$\begin{aligned}
A_1 \Leftrightarrow {} & \big(\alpha_1 \wedge \beta_1 \wedge (\alpha_2 \Rightarrow \delta_{21}) \wedge (\alpha_4 \Rightarrow \delta_{41})\\
&\qquad \wedge ((\alpha_2 \wedge \beta_2 \wedge \delta_{12}) \Rightarrow \eta) \wedge ((\alpha_4 \wedge \beta_2 \wedge \delta_{12}) \Rightarrow \eta)\big)\\
&\quad \vee \big(\alpha_3 \wedge \beta_1 \wedge (\alpha_2 \Rightarrow \delta_{21}) \wedge (\alpha_4 \Rightarrow \delta_{41})\\
&\qquad \wedge ((\alpha_2 \wedge \beta_2 \wedge \delta_{32}) \Rightarrow \eta) \wedge ((\alpha_4 \wedge \beta_2 \wedge \delta_{32}) \Rightarrow \eta)\big),\\
\Leftrightarrow {} & \beta_1 \wedge (\alpha_2 \Rightarrow \delta_{21}) \wedge (\alpha_4 \Rightarrow \delta_{41})\\
&\quad \wedge \Big(\big(\alpha_1 \wedge ((\alpha_2 \wedge \beta_2 \wedge \delta_{12}) \Rightarrow \eta) \wedge ((\alpha_4 \wedge \beta_2 \wedge \delta_{12}) \Rightarrow \eta)\big)\\
&\qquad \vee \big(\alpha_3 \wedge ((\alpha_2 \wedge \beta_2 \wedge \delta_{32}) \Rightarrow \eta) \wedge ((\alpha_4 \wedge \beta_2 \wedge \delta_{32}) \Rightarrow \eta)\big)\Big),\\
\Leftrightarrow {} & \beta_1 \wedge (\alpha_2 \Rightarrow \delta_{21}) \wedge (\alpha_4 \Rightarrow \delta_{41})\\
&\quad \wedge \Big(\big(\alpha_1 \wedge (\neg\alpha_2 \vee \neg\beta_2 \vee \neg\delta_{12} \vee \eta) \wedge (\neg\alpha_4 \vee \neg\beta_2 \vee \neg\delta_{12} \vee \eta)\big)\\
&\qquad \vee \big(\alpha_3 \wedge (\neg\alpha_2 \vee \neg\beta_2 \vee \neg\delta_{32} \vee \eta) \wedge (\neg\alpha_4 \vee \neg\beta_2 \vee \neg\delta_{32} \vee \eta)\big)\Big),\\
\Leftrightarrow {} & \beta_1 \wedge (\neg\alpha_2 \vee \delta_{21}) \wedge (\neg\alpha_4 \vee \delta_{41})\\
&\quad \wedge \Big(\big(\alpha_1 \wedge ((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \neg\delta_{12} \vee \eta)\big)\\
&\qquad \vee \big(\alpha_3 \wedge ((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \neg\delta_{32} \vee \eta)\big)\Big),\\
\Leftrightarrow {} & \beta_1 \wedge (\neg\alpha_2 \vee \delta_{21}) \wedge (\neg\alpha_4 \vee \delta_{41})\\
&\quad \wedge \Big(\big(\alpha_1 \wedge ((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \eta)\big) \vee (\alpha_1 \wedge \neg\delta_{12})
\end{aligned}$$

$$
\begin{aligned}
&\qquad\quad \vee \big(\alpha_3 \wedge ((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \eta)\big) \vee (\alpha_3 \wedge \neg\delta_{32})\Big),\\
&\Leftrightarrow \beta_1 \wedge (\neg\alpha_2 \vee \delta_{21}) \wedge (\neg\alpha_4 \vee \delta_{41})\\
&\qquad \wedge \Big(\big((\alpha_1 \vee \alpha_3) \wedge ((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \eta)\big) \vee (\alpha_1 \wedge \neg\delta_{12}) \vee (\alpha_3 \wedge \neg\delta_{32})\Big),\\
&\Leftrightarrow \beta_1 \wedge (\neg\alpha_2 \vee \delta_{21}) \wedge (\neg\alpha_4 \vee \delta_{41})\\
&\qquad \wedge \Big(\big((\alpha_1 \vee \alpha_3) \wedge ((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \eta)\big)\\
&\qquad\quad \vee ((\alpha_1 \vee \alpha_3) \wedge \alpha_1 \wedge \neg\delta_{12}) \vee ((\alpha_1 \vee \alpha_3) \wedge \alpha_3 \wedge \neg\delta_{32})\Big),\\
&\Leftrightarrow (\alpha_1 \vee \alpha_3) \wedge \beta_1 \wedge (\neg\alpha_2 \vee \delta_{21}) \wedge (\neg\alpha_4 \vee \delta_{41})\\
&\qquad \wedge \big((\neg\alpha_2 \wedge \neg\alpha_4) \vee \neg\beta_2 \vee \eta \vee (\alpha_1 \wedge \neg\delta_{12}) \vee (\alpha_3 \wedge \neg\delta_{32})\big).
\end{aligned}
$$

Written in its original variables, this is

$$
\begin{aligned}
A_1 \Leftrightarrow\ & (g_{c1} \geq u_{\max,1} \vee g_{c1} \leq u_{\min,1}) \wedge \nabla g_{c1} \neq 0 \wedge (g_{c2} < u_{\max,2} \vee \nabla^{\mathrm{T}} g_{c2} R_1 f_c \leq 0)\\
& \wedge (g_{c2} > u_{\min,2} \vee \nabla^{\mathrm{T}} g_{c2} R_1 f_c \geq 0) \wedge \big(u_{\min,2} < g_{c2} < u_{\max,2} \vee \nabla g_{c2} = 0 \vee F_1 \geq F_2\\
&\qquad \vee (g_{c1} \geq u_{\max,1} \wedge \nabla^{\mathrm{T}} g_{c1} R_2 f_c > 0) \vee (g_{c1} \leq u_{\min,1} \wedge \nabla^{\mathrm{T}} g_{c1} R_2 f_c < 0)\big).
\end{aligned}
$$

Finally, examination of condition (iii) yields A_2 as

$$
\begin{aligned}
A_2 \Leftrightarrow\ & \big(2 \in \mathcal{I}_{\mathrm{sat}} \wedge \operatorname{rank}(N_{\{2\}}) = 1 \wedge (1 \in \mathcal{I}_{\mathrm{sat}} \Rightarrow \nabla^{\mathrm{T}} h_1 R_{\{2\}} f_c \leq 0)\\
&\qquad \wedge (3 \in \mathcal{I}_{\mathrm{sat}} \Rightarrow \nabla^{\mathrm{T}} h_3 R_{\{2\}} f_c \leq 0)\big)\\
& \vee \big(4 \in \mathcal{I}_{\mathrm{sat}} \wedge \operatorname{rank}(N_{\{4\}}) = 1 \wedge (1 \in \mathcal{I}_{\mathrm{sat}} \Rightarrow \nabla^{\mathrm{T}} h_1 R_{\{4\}} f_c \leq 0)\\
&\qquad \wedge (3 \in \mathcal{I}_{\mathrm{sat}} \Rightarrow \nabla^{\mathrm{T}} h_3 R_{\{4\}} f_c \leq 0)\big),\\
\Leftrightarrow\ & \big(g_{c2} \geq u_{\max,2} \wedge \nabla g_{c2} \neq 0 \wedge (g_{c1} \geq u_{\max,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0)\\
&\qquad \wedge (g_{c1} \leq u_{\min,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} R_2 f_c \geq 0)\big)\\
& \vee \big(g_{c2} \leq u_{\min,2} \wedge \nabla g_{c2} \neq 0 \wedge (g_{c1} \geq u_{\max,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0)\\
&\qquad \wedge (g_{c1} \leq u_{\min,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} R_2 f_c \geq 0)\big),\\
\Leftrightarrow\ & (g_{c2} \geq u_{\max,2} \vee g_{c2} \leq u_{\min,2}) \wedge \nabla g_{c2} \neq 0 \wedge (g_{c1} \geq u_{\max,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0)\\
&\qquad \wedge (g_{c1} \leq u_{\min,1} \Rightarrow \nabla^{\mathrm{T}} g_{c1} R_2 f_c \geq 0),\\
\Leftrightarrow\ & (g_{c2} \geq u_{\max,2} \vee g_{c2} \leq u_{\min,2}) \wedge \nabla g_{c2} \neq 0 \wedge (g_{c1} < u_{\max,1} \vee \nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0)\\
&\qquad \wedge (g_{c1} > u_{\min,1} \vee \nabla^{\mathrm{T}} g_{c1} R_2 f_c \geq 0).
\end{aligned}
$$

With the logical statements A_0, A_1, and A_2 made explicit, the closed-form expressions for the GPAW-compensated controller (B.2) with two dimensional output are

$$
\dot{x}_g = \begin{cases}
f_c(x_g, y, r), & \text{if } A_0,\\
\big(I - \frac{1}{\nabla^{\mathrm{T}} g_{c1} \Gamma \nabla g_{c1}(x_g)} \Gamma \nabla g_{c1} \nabla^{\mathrm{T}} g_{c1}(x_g)\big) f_c(x_g, y, r), & \text{if } \neg A_0 \wedge A_1,\\
\big(I - \frac{1}{\nabla^{\mathrm{T}} g_{c2} \Gamma \nabla g_{c2}(x_g)} \Gamma \nabla g_{c2} \nabla^{\mathrm{T}} g_{c2}(x_g)\big) f_c(x_g, y, r), & \text{if } \neg A_0 \wedge \neg A_1 \wedge A_2,\\
\big(I - \Gamma \tilde{N} (\tilde{N}^{\mathrm{T}} \Gamma \tilde{N})^{-1} \tilde{N}^{\mathrm{T}}(x_g)\big) f_c(x_g, y, r), & \text{otherwise,}
\end{cases} \tag{B.5}
$$

$$
u_g = g_c(x_g), \qquad x_g(0) = x_{c0},
$$

where the switching conditions are collected below for ease of reference

$$
\begin{aligned}
A_0 \Leftrightarrow{} & (g_{c1} < u_{\max,1} \vee \nabla^{\mathrm{T}} g_{c1} f_c \leq 0) \wedge (g_{c1} > u_{\min,1} \vee \nabla^{\mathrm{T}} g_{c1} f_c \geq 0) \\
& \wedge (g_{c2} < u_{\max,2} \vee \nabla^{\mathrm{T}} g_{c2} f_c \leq 0) \wedge (g_{c2} > u_{\min,2} \vee \nabla^{\mathrm{T}} g_{c2} f_c \geq 0), \\
A_1 \Leftrightarrow{} & (g_{c1} \geq u_{\max,1} \vee g_{c1} \leq u_{\min,1}) \wedge \nabla g_{c1} \neq 0 \wedge (g_{c2} < u_{\max,2} \vee \nabla^{\mathrm{T}} g_{c2} R_1 f_c \leq 0) \\
& \wedge (g_{c2} > u_{\min,2} \vee \nabla^{\mathrm{T}} g_{c2} R_1 f_c \geq 0) \wedge \big(u_{\min,2} < g_{c2} < u_{\max,2} \vee \nabla g_{c2} = 0 \vee F_1 \geq F_2 \\
& \vee (g_{c1} \geq u_{\max,1} \wedge \nabla^{\mathrm{T}} g_{c1} R_2 f_c > 0) \vee (g_{c1} \leq u_{\min,1} \wedge \nabla^{\mathrm{T}} g_{c1} R_2 f_c < 0)\big), \\
A_2 \Leftrightarrow{} & (g_{c2} \geq u_{\max,2} \vee g_{c2} \leq u_{\min,2}) \wedge \nabla g_{c2} \neq 0 \wedge (g_{c1} < u_{\max,1} \vee \nabla^{\mathrm{T}} g_{c1} R_2 f_c \leq 0) \\
& \wedge (g_{c1} > u_{\min,1} \vee \nabla^{\mathrm{T}} g_{c1} R_2 f_c \geq 0),
\end{aligned}
$$

and

$$
\begin{aligned}
R_1 &= I - \frac{\Gamma \nabla g_{c1} \nabla^{\mathrm{T}} g_{c1}(x_g)}{\nabla^{\mathrm{T}} g_{c1} \Gamma \nabla g_{c1}(x_g)}, & \tilde{N} &= [\nabla g_{c1}(x_g), \nabla g_{c2}(x_g)], \\
& & F_1 &= f_c^{\mathrm{T}}(x_g, y, r) \Gamma^{-1} R_1 f_c(x_g, y, r), \\
R_2 &= I - \frac{\Gamma \nabla g_{c2} \nabla^{\mathrm{T}} g_{c2}(x_g)}{\nabla^{\mathrm{T}} g_{c2} \Gamma \nabla g_{c2}(x_g)}, & F_2 &= f_c^{\mathrm{T}}(x_g, y, r) \Gamma^{-1} R_2 f_c(x_g, y, r).
\end{aligned}
$$

While these closed-form expressions appear complicated, they can be easily implemented without the need to solve any optimization problem online. They allow a highly computationally efficient implementation of GPAW-compensated controllers with two dimensional output, as shown in the next section. If the output equation of the nominal controller is linear, i.e. $g_c(x_g) = c_c^{\mathrm{T}} x_g$ for some constant vector $c_c \in \mathbb{R}^q$, simplifications analogous to those in Section A.1 of Appendix A can be obtained. In particular, the condition $F_1 \geq F_2$ in the statement A_1 reduces to a condition on the positive definiteness of the *constant* matrix $\Gamma^{-1}(R_1 - R_2)$ that can be verified offline. Further simplifications are obtained when the nominal controller is LTI. We omit these for brevity.

B.1 Comparison of Computational Performance

Here, we demonstrate the efficiency of the computational procedure employing the closed-form expressions in (B.5), by comparing its computation times with those obtained by solving the quadratic program (4.12) in Section 4.1,

$$
\begin{aligned}
& \min_{x \in \mathbb{R}^q} \|\Phi^{-1} f_c - x\|^2, \\
& \text{subject to} \qquad N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}} \Phi x \leq 0,
\end{aligned}
$$

where $f_c := f_c(x_g, y, r)$, Φ is obtained from a decomposition of the GPAW parameter $\Gamma = \Phi\Phi^{\mathrm{T}}$ [124, Theorem 7.2.7, p. 406], and $N_{\mathcal{I}_{\mathrm{sat}}} := N_{\mathcal{I}_{\mathrm{sat}}}(x_g)$ is the vector of gradient vectors corresponding to active saturation constraints (see its definition before (B.3)). The preceding quadratic program can be written equivalently as

$$
\begin{aligned}
& \min_{x \in \mathbb{R}^q} x^{\mathrm{T}} x - 2 f_c^{\mathrm{T}} \Phi^{-\mathrm{T}} x, \\
& \text{subject to} \qquad N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}} \Phi x \leq 0,
\end{aligned}
\tag{B.6}
$$

where the constant term $f_c^{\mathrm{T}} \Phi^{-\mathrm{T}} \Phi f_c$ in the objective function has been dropped.

We will be comparing the vector $f_g := R_{\mathcal{I}^*}(x_g, y, r) f_c(x_g, y, r)$ obtained using the two methods. The closed-form expressions (B.5) yields the switching condition $\mathcal{I}^*$ together with

f_g. We denote this method by "CF", and its solution by f_{gCF}. Given the solution x^* to the quadratic program (B.6), f_g is also defined by $f_g = \Phi x^*$ (see (4.11)). Both these equivalent ways to obtain f_g were implemented in MATLAB®, where the MATLAB® function `quadprog` is used to solve (B.6). Recognizing that the initial guess x_{ig} supplied to the quadratic program solver may have a significant impact on the computation times, we solve it in two ways (see also Section 4.1.1):

(i) set the initial guess x_{ig} as $x_{ig} = f_c$. This is motivated by the fact that it yields the optimal solution immediately when the constraints are feasible. We refer to this solution method by "QP", and the solution obtained (with this initial guess) by f_{gQP};

(ii) set the initial guess x_{ig} as $x_{ig} = x^*$, where x^* is the optimal solution to the quadratic program (B.6) obtained in case (i). Observe that this is the best possible initialization, i.e. it is initialized *at the optimal solution*. This solution method will be denoted by "QPo", and its solution by f_{gQPo}.

The computations were carried out for controllers of orders between 2 and 10, i.e. $q \in \{2, 3, \ldots, 10\}$. For each controller order, 10000 computations were performed using the methods CF, QP, and QPo, for randomly generated data $(f_c, g_c, N_{\mathcal{I}_{\text{sat}}}, \Gamma, u_{\max}, u_{\min})$. For $f_c \in \mathbb{R}^q$, $g_c, u_{\max}, u_{\min} \in \mathbb{R}^2$, $N_{\mathcal{I}_{\text{sat}}} \in \mathbb{R}^{q \times |\mathcal{I}_{\text{sat}}|}$, each of its elements are uniformly distributed in the interval $(-1, 1)$. Additionally, the elements of $u_{\max}$, $u_{\min}$, and g_c are swapped whenever necessary to ensure $u_{\min,i} < u_{\max,i}$ for $i \in \mathcal{I}_2$, and that at least one of the four conditions

$$g_{c1} \geq u_{\max,1}, \qquad g_{c1} \leq u_{\min,1}, \qquad g_{c2} \geq u_{\max,2}, \qquad g_{c2} \leq u_{\min,2},$$

holds, i.e. restrict only to the saturated cases. The way this was done ensures that the probability of both g_{c1} and g_{c2} saturating is $\frac{2}{3}$, and g_{c1}, g_{c2} saturate alone with probability $\frac{1}{6}$ each. The GPAW parameter is computed from a randomly generated $\Phi \in \mathbb{R}^{q \times q}$ by $\Gamma = \Phi\Phi^{\mathrm{T}}$. Each element of Φ is again uniformly distributed in the interval $(-1, 1)$. The matrix Φ is regenerated whenever its condition number [124, p. 336] is greater than 10, to ensure numerical robustness (see also Remark 2.25).

For all these 90000 instances, the maximum errors are

$$\max\|f_{gCF} - f_{gQPo}\| \approx \max\|f_{gCF} - f_{gQP}\| = 3.3 \times 10^{-10},$$

and the normalized errors are

$$\max \frac{\|f_{gCF} - f_{gQPo}\|}{\|f_c\|} \approx \max \frac{\|f_{gCF} - f_{gQP}\|}{\|f_c\|} = 3.8 \times 10^{-10}.$$

The distribution of the switching conditions is shown in Fig. B-1 for controller order ranging between 2 and 10. This shows that all switching conditions are well tested.[7]

While the numerical errors involving f_g presented above are repeatable (when the random number generator has been seeded appropriately), note that computation times are

[7] It can be shown that when both g_{c1} and g_{c2} are saturated, the probability of each of the switching conditions ($\mathcal{I}^* = \emptyset$, $\mathcal{I}^* \in \{\{1\}, \{3\}\}$, $\mathcal{I}^* \in \{\{2\}, \{4\}\}$, $\mathcal{I}^* = \mathcal{I}_{\text{sat}}$) occurring is $\frac{1}{4}$. The randomly generated conditions were such that the probability of generating cases where *both* g_{c1} and g_{c2} are saturated is $\frac{2}{3}$, and g_{c1}, g_{c2} saturate alone with probability $\frac{1}{6}$ each. Then the probabilities of each of the switching conditions occurring can be shown to be $\{\frac{2}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{2}, \frac{2}{3} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{2}, \frac{2}{3} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{2}, \frac{2}{3} \cdot \frac{1}{4}\} = \{\frac{1}{3}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}\} \approx \{0.33, 0.25, 0.25, 0.17\}$ respectively. From Fig. B-1, it can be seen that the experimental distribution of the switching conditions agree well with the prediction.

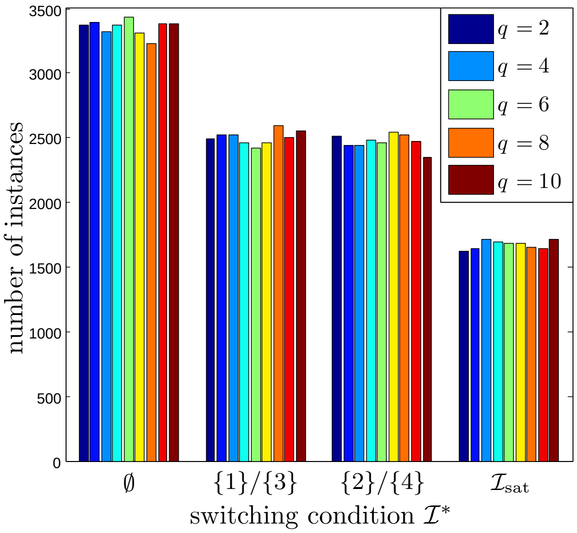


Figure B-1: Distribution of switching conditions for solutions where controller order q ranges from 2 to 10.

repeatable only on average. Repeating the computations will produce identical errors for f_g, but will be different for computation times in general. This is due to uncontrollable operating system demands that may have interrupted the computations, resulting in sporadic instances having larger than normal computation times. All computations are performed on the same computer, a dual 3.2 GHz Intel Pentium processor machine with 1 Gb memory, running on Ubuntu 10.04 (a Linux distribution).[8] All possible preprocessing and postprocessing are not included in the computation times. In particular, matrix inversions where necessary are performed beforehand, and diagnostic messages from the quadratic program solver are disabled to obtain more accurate computation times.

The statistical data of the computation times are shown in Fig. B-2. The left plot of Fig. B-2 shows the computation times with the controller order ranging between 2 and 10, while the right plot shows the normalized computation times. Here, $(t_{QP}, t_{QPo}, t_{CF})$ denote the computation times obtained with methods QP, QPo, and CF respectively, and $(t_{CF}/t_{QP}, t_{CF}/t_{QPo})$ denote the normalized computation times. Solid lines represent mean values while dashed lines represent maximum values. The vertical bars on the solid lines represent the standard deviation across all 10000 test cases for each controller order. These results are summarized in Table B.1 across all 90000 test cases. Note that the large maximum computation time for t_{QP} and t_{QPo} is likely due to sporadic operating system interrup-

Table B.1: Statistical Summary

Quantity	Mean	Maximum
t_{QP}	3.6 ms	565.7 ms
t_{QPo}	2.7 ms	97.2 ms
t_{CF}	0.17 ms	1.2 ms
t_{CF}/t_{QP}	0.046	0.55
t_{CF}/t_{QPo}	0.060	0.58

[8] See `http://www.ubuntu.com`.

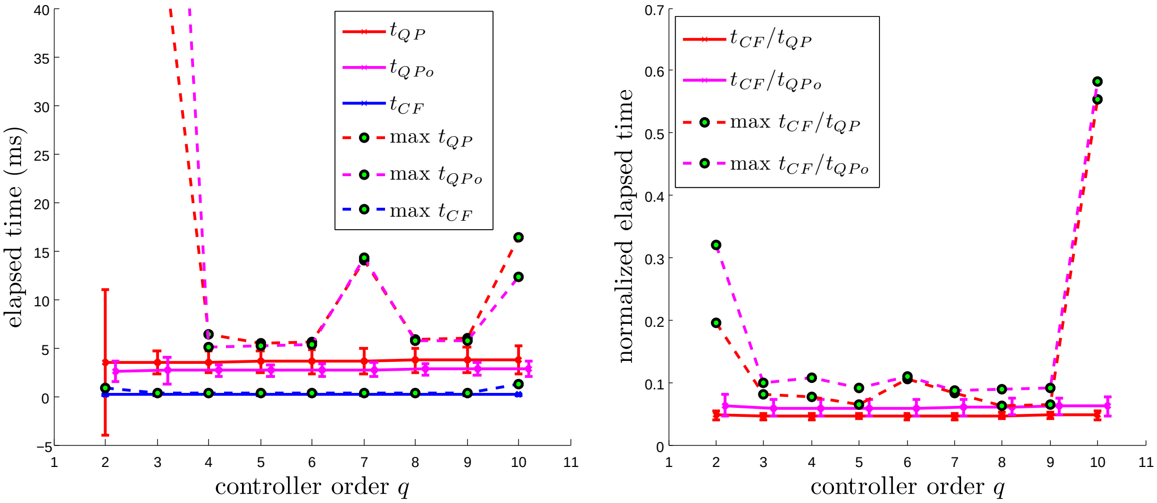


Figure B-2: Statistical data of computation times. The left plot shows the absolute computation times, while the normalized computation times are shown on the right. Dashed lines represent maximum values, and solid lines represent mean values, both over 10000 test cases for each controller order. The extents of the vertical bars on the solid lines represent the standard deviation over the test cases. It can be seen that on average, the solution obtained with the closed-form expressions take less than 10% of the computation time required for the solution obtained from the best case quadratic program.

tions mentioned previously. They occur rarely and are outliers, as seen by the significantly lower mean for t_{QP} and t_{QPo}. To support this view, the left plot of Fig. B-3 shows the individual computation times when the controller order is 2 and 10, while the right plot is when the controller order is 4 and 8. Observe that for instances near the 7000-th for $q = 2$, and near the 8000-th for $q = 10$, the computation times were increased significantly. These increases are sporadic and indicate interruptions by the operating system, which accounts for a large part of the variance in data. Nominal cases are shown on the right plot of Fig. B-3 where such interruptions are absent. Note that out of 90000 test cases, only 10 instances for method CF has computation times of more than 0.4 ms.

From Table B.1, it can be seen that for all 90000 test cases, the solution obtained using the closed-form expressions takes only a fraction of time needed to obtain the best case quadratic program solution. On average, it takes less than 7% of the computation time. From these, we can conclude that in general, adopting the closed-form expressions (B.5) can yield significant computational savings as compared to the *best case quadratic program initialized with the optimal solution.*

B.2 First Order Controllers

For completeness, we derive here the closed-form expressions for *first order* GPAW-compensated controllers (B.2) with output of dimension two. All the constructions up to the definition of subproblem (B.4) in the first part of this Appendix remains valid, with some simplifications to be shown.

First, observe that for first order controllers, the gradients $\nabla h_i(x_g) = -\nabla h_{i+2}(x_g) = \nabla g_{ci}(x_g)$ for $i \in \mathcal{I}_2$ reduces to *scalars*, and $\Gamma \in \mathbb{R}^{1\times 1}$ is a positive *scalar*. Then the projection

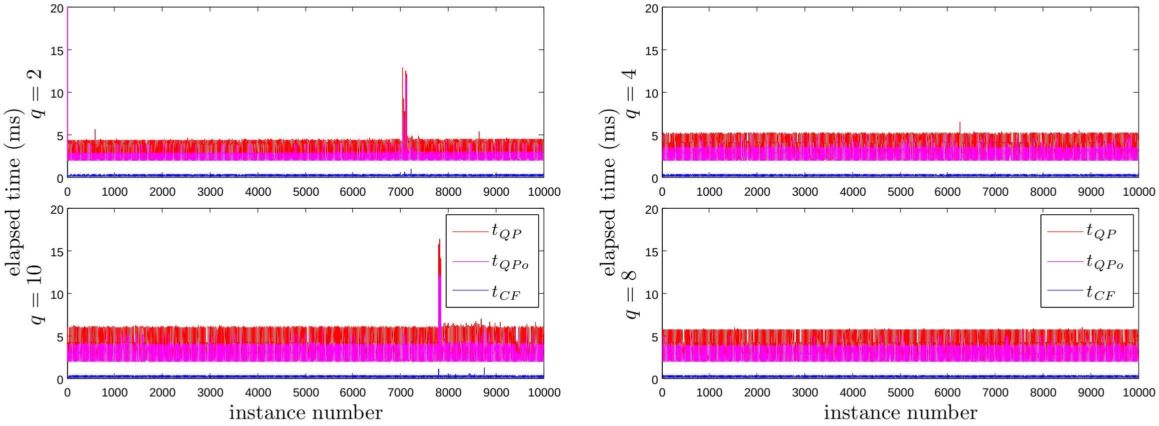


Figure B-3: Interruptions of computations by operating system. The left plot shows sporadic increases in computation times occurring when the controller order is 2 and 10, indicating operating system interruptions. The right plot shows the nominal cases where the controller order is 4 and 8.

matrix (B.3) reduces to

$$R_{\mathcal{I}}(x_g) = \begin{cases} 1 - N_{\mathcal{I}}(N_{\mathcal{I}}^{\mathrm{T}} N_{\mathcal{I}})^{-1} N_{\mathcal{I}}^{\mathrm{T}}(x_g), & \text{if } \mathcal{I} \neq \emptyset, \\ 1, & \text{otherwise,} \end{cases}$$

which is a *scalar, independent of* Γ. Since $q = 1$ for first order controllers, a full rank $N_{\mathcal{I}}(x_g) \in \mathbb{R}^{q \times \max\{|\mathcal{I}|,1\}} = \mathbb{R}^{1 \times \max\{|\mathcal{I}|,1\}}$ implies that it is *scalar*. Then for any $\mathcal{I} \subset \mathcal{I}_{\mathrm{sat}}$ such that $\operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|$, the preceding projection matrix reduces further to

$$R_{\mathcal{I}}(x_g) = \begin{cases} 0, & \text{if } \mathcal{I} \neq \emptyset, \\ 1, & \text{otherwise.} \end{cases} \tag{B.7}$$

Multiplying the objective function of subproblem (B.4) by the positive scalar Γ, we obtain

$$\tilde{F}(\mathcal{I}) = \Gamma F(\mathcal{I}) = f_c^{\mathrm{T}}(x_g, y, r) R_{\mathcal{I}}(x_g) f_c(x_g, y, r) = \begin{cases} 0, & \text{if } \mathcal{I} \neq \emptyset, \\ \|f_c(x_g, y, r)\|^2, & \text{otherwise.} \end{cases}$$

Using this as the equivalent objective function, we obtain the subproblem

$$\begin{aligned} \max_{\mathcal{I} \in \mathcal{J}} \tilde{F}(\mathcal{I}) &= \begin{cases} 0, & \text{if } \mathcal{I} \neq \emptyset, \\ \|f_c(x_g, y, r)\|^2, & \text{otherwise,} \end{cases} \\ \text{subject to} \quad & \operatorname{rank}(N_{\mathcal{I}}(x_g)) = |\mathcal{I}|, \\ & N_{\mathcal{I}_{\mathrm{sat}}}^{\mathrm{T}}(x_g) R_{\mathcal{I}}(x_g) f_c(x_g, y, r) \leq 0, \end{aligned} \tag{B.8}$$

which again is *independent of* Γ.

Remark B.1. These observations apply to general first order GPAW-compensated controllers, and not only those with output of dimension two. Since the GPAW-compensated controller is fully defined by $R_{\mathcal{I}^*}(x_g, y, r)$ with $\mathcal{I}^*$ a solution to the optimization subproblem (B.8), the fact that both quantities are independent of Γ show that first order GPAW-compensated controllers are fully defined *independent of any parameter*.

This is due to the fact that the GPAW parameter Γ only changes the way the controller state is modified on the boundary of the unsaturated region K. For first order controllers, there can only be two boundary *points* (in contrast to a higher dimensional surface) for which there is only one way to maintain controller state-output consistency, i.e. stop the scalar state evolution when the nominal update will cause constraint violations. □

It is clear from subproblem (B.8) (as well as Proposition 2.5.2) that the optimal solution is $\mathcal{I}^* = \emptyset$ whenever it is feasible. Recognizing that the rank condition always hold for $\mathcal{I} = \emptyset$, it is feasible when condition A_0 in (B.5) holds. For all other cases, we have $R_{\mathcal{I}^*}(x_g) = 0$ due to (B.7). Hence the closed-form expressions for the first order GPAW-compensated controller with output of dimension two is given by

$$\dot{x}_g = \begin{cases} f_c(x_g, y, r), & \text{if } A_0, \\ 0, & \text{otherwise,} \end{cases}$$
$$u_g = g_c(x_g), \qquad x_g(0) = x_{c0},$$

where the switching condition A_0 is identical to that in (B.5), reproduced here for ease of reference

$$\begin{aligned} A_0 \Leftrightarrow\ & (g_{c1} < u_{\max,1} \vee \nabla^{\mathrm{T}} g_{c1} f_c \leq 0) \wedge (g_{c1} > u_{\min,1} \vee \nabla^{\mathrm{T}} g_{c1} f_c \geq 0) \\ & \wedge (g_{c2} < u_{\max,2} \vee \nabla^{\mathrm{T}} g_{c2} f_c \leq 0) \wedge (g_{c2} > u_{\min,2} \vee \nabla^{\mathrm{T}} g_{c2} f_c \geq 0). \end{aligned}$$

Appendix C

Procedure to Apply GPAW Compensation

There are three different but equivalent ways (detailed in Sections 2.5 and 4.1) to realize *general* GPAW-compensated controllers, while closed-form expressions (detailed in Appendices A and B) are available when the GPAW-compensated controller (or nominal controller) has output of dimension one or two. For ease of reference, we summarize the procedure to apply GPAW compensation here.

Step I If the nominal controller is of the form (2.24),

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c, y, r), \end{aligned} \tag{2.24}$$

i.e. with output function g_c depending on measurement y and/or external reference signal r, obtain an approximating controller of the form (2.26),

$$\begin{aligned} \dot{x}_c &= f_c(x_c, y, r), \qquad x_c(0) = x_{c0}, \\ u_c &= g_c(x_c), \end{aligned} \tag{2.26}$$

with output function g_c depending only on the controller state x_c. One way to approximate the nominal controller is discussed in Section 2.6, examples of which are available in Sections 2.8.1, 6.1.2, 6.2.3, and A.2 (in Appendix A).

Step II Determine the GPAW parameter Γ. Note that any meaningful nominal controller of the form (2.26) (possibly approximated due to Step I above) must be *dynamic*, i.e. with controller state of dimension at least one. When the nominal controller is first order, this step can be bypassed (see Remark B.1). Theorems 4.4.3 and 5.2.1 yield a Γ when they are applicable. When they are not applicable, some trial and error may be needed to determine Γ. Once Γ is determined, the GPAW-compensated controller is fully defined.

Step III Implement the GPAW-compensated controller.

(i) For GPAW-compensated controllers with output of dimension one or two, the closed-form expressions (A.5) and (B.5) are recommended for computational efficiency (see Section B.1 in Appendix B). See Appendices A and B for specializations and possible simplifications.

(ii) For GPAW-compensated controllers with output of dimension greater than two, they can be implemented in one of three possible ways, namely:

   (a) the combinatorial optimization formulation detailed in Section 2.5. The GPAW-compensated controller is defined by (2.27), (2.28), (2.29), (2.30), and the online solution to the combinatorial optimization subproblem (2.31). Proposition 2.5.1 ensures the existence of solutions to subproblem (2.31), which can always be found by an exhaustive search algorithm (see Remark 2.8);

   (b) the quadratic program formulation detailed in Section 4.1. The GPAW-compensated controller is defined by (4.11), a matrix Φ obtained from a decomposition of the GPAW parameter $\Gamma = \Phi\Phi^{\mathrm{T}}$, and the online solution[1] to the convex quadratic program (4.12);

   (c) the projection onto a convex polyhedral cone formulation detailed in Section 4.1. This is identical to case (b), except that the quadratic program (4.12) is solved instead as a projection of the vector $\Phi^{-1}f_c(x_g, y, r)$ onto the convex polyhedral cone $\mathcal{K}^{\circ}$ (4.10). For algorithms to project onto convex polyhedral cones, see [170–175] and the references therein.

Step IV Determine controller state initialization. As shown in Section 6.1, an appropriate controller state initialization may yield significant improvements in the transient response of the GPAW-compensated system when the default choice $x_g(0) = x_c(0)$, i.e. taking the initial state of the nominal controller, is “too far” from the unsaturated region. An example was shown in Section 6.1.3 on how to determine a reasonable state to initialize the GPAW-compensated controller.

The preceding summarizes the procedure to apply GPAW compensation to a general saturated nonlinear plant (1.1) driven by a nonlinear controller (1.3).

[1] See `http://www.numerical.rl.ac.uk/qp/qp.html` for a list of available quadratic program solvers.

Bibliography

[1] D. S. Bernstein and A. N. Michel, “A chronological bibliography on saturating actuators,” *Int. J. Robust Nonlinear Control*, vol. 5, no. 5, pp. 375 – 380, 1995.

[2] D. Liu and A. N. Michel, *Dynamical Systems with Saturation Nonlinearities: Analysis and Design*, ser. Lect. Notes Control Inf. Sci. London, Great Britain: Springer, 1994, vol. 195.

[3] T. Hu and Z. Lin, *Control Systems with Actuator Saturation: Analysis and Design*, ser. Control Eng. Boston, MA: Birkhäuser, 2001.

[4] V. Kapila and K. M. Grigoriadis, Eds., *Actuator Saturation Control*, ser. Control Eng. New York, NY: Marcel Dekker, 2002.

[5] A. H. Glattfelder and W. Schaufelberger, *Control Systems with Input and Output Constraints*, ser. Adv. Textb. Control Signal Process. London, United Kingdom: Springer, 2003.

[6] G. C. Goodwin, M. M. Seron, and J. A. De Doná, *Constrained Control and Estimation: An Optimization Approach*, ser. Commun. Control Eng. London, United Kingdom: Springer, 2005.

[7] P. Hippe, *Windup in Control: Its Effects and Their Prevention*, ser. Adv. Ind. Control. London, United Kingdom: Springer, 2006.

[8] S. Tarbouriech, G. Garcia, and A. H. Glattfelder, Eds., *Advanced Strategies in Control Systems with Input and Output Constraints*, ser. Lect. Notes Control Inf. Sci. Berlin, Germany: Springer, 2007, vol. 346.

[9] *Special Issue on Saturating Actuators, Int. J. Robust Nonlinear Control*, vol. 5, no. 5, pp. 375 – 540, 1995.

[10] *Special Issue on Control Problems with Constraints, Int. J. Robust Nonlinear Control*, vol. 9, no. 10, pp. 583 – 734, Aug. 1999.

[11] *Special Issue on Anti-windup, Int. J. Syst. Sci.*, vol. 37, no. 2, pp. 65 – 139, Feb. 2006.

[12] M. V. Kothare, P. J. Campo, M. Morari, and C. N. Nett, “A unified framework for the study of anti-windup designs,” *Automatica*, vol. 30, no. 12, pp. 1869 – 1883, Dec. 1994.

[13] C. Edwards and I. Postlethwaite, “Anti-windup and bumpless-transfer schemes,” *Automatica*, vol. 34, no. 2, pp. 199 – 210, Feb. 1998.

[14] S. Tarbouriech and M. Turner, “Anti-windup design: an overview of some recent advances and open problems,” *IET Control Theory Appl.*, vol. 3, no. 1, pp. 1 – 19, Jan. 2009.

[15] J. C. Doyle, R. S. Smith, and D. F. Enns, “Control of plants with input saturation nonlinearities,” in *Proc. American Control Conf.*, Minneapolis, MN, Jun. 1987, pp. 1034 – 1039.

[16] D. A. Richardson, “Means for preventing reset wind-up in electronic control apparatus,” U.S. Patent 3 197 711, Jul. 27, 1965. [Online]. Available: http://www.google.com/patents?id=JOFJAAAAEBAJ

[17] A. Visioli, *Practical PID Control*, ser. Adv. Ind. Control. London, United Kingdom: Springer, 2006.

[18] K. S. Walgama, S. Rönnbäck, and J. Sternby, “Generalisation of conditioning technique for anti-windup compensators,” *IEE Proc.-Control Theory Appl.*, vol. 139, no. 2, pp. 109 – 118, Mar. 1992.

[19] M. A. Dornheim, “Report pinpoints factors leading to YF-22 crash,” *Aviat. Week Space Technol.*, vol. 137, no. 19, pp. 53 – 54, Nov. 1992.

[20] P. Butterworth-Hayes, “Gripen crash raises canard fears,” *Aerosp. Am.*, vol. 32, no. 2, pp. 10 – 11, Feb. 1994.

[21] G. Stein, “Respect the unstable,” *IEEE Control Syst. Mag.*, vol. 23, no. 4, pp. 12 – 25, Aug. 2003.

[22] J. C. Lozier, “A steady state approach to the theory of saturable servo systems,” *IRE Trans. Autom. Control*, vol. 1, no. 1, pp. 19 – 39, May 1956.

[23] E. Gilbert and I. Kolmanovsky, “Nonlinear tracking control in the presence of state and control constraints: a generalized reference governor,” *Automatica*, vol. 38, no. 12, pp. 2063 – 2073, Dec. 2002.

[24] F. Morabito, A. R. Teel, and L. Zaccarian, “Nonlinear antiwindup applied to Euler-Lagrange systems,” *IEEE Trans. Robot. Autom.*, vol. 20, no. 3, pp. 526 – 537, Jun. 2004.

[25] A. E. Bryson, Jr. and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control.* Levittown, PA: Taylor & Francis, 1975.

[26] J. Sofrony, M. C. Turner, I. Postlethwaite, O. Brieger, and D. Leiβling, “Anti-windup synthesis for PIO avoidance in an experimental aircraft,” in *Proc. 45th IEEE Conf. Decision and Control*, San Diego, CA, Dec. 2006, pp. 5412 – 5417.

[27] E. F. Mulder, P. Y. Tiwari, and M. V. Kothare, “Simultaneous linear and anti-windup controller synthesis using multiobjective convex optimization,” *Automatica*, vol. 45, no. 3, pp. 805 – 811, Mar. 2009.

[28] C. Burgat and S. Tarbouriech, “Intelligent anti-windup for systems with input magnitude saturation,” *Int. J. Robust Nonlinear Control*, vol. 8, no. 12, pp. 1085 – 1100, Oct. 1998.

[29] T. Kiyama and T. Iwasaki, "On the use of multi-loop circle criterion for saturating control synthesis," *Syst. Control Lett.*, vol. 41, no. 2, pp. 105 – 114, Oct. 2000.

[30] F. Wu, K. M. Grigoriadis, and A. Packard, "Anti-windup controller design using linear parameter-varying control methods," *Int. J. Control*, vol. 73, no. 12, pp. 1104 – 1114, Aug. 2000.

[31] T. Kiyama, S. Hara, and T. Iwasaki, "Effectiveness and limitation of circle criterion for lti robust control systems with control input nonlinearities of sector type," *Int. J. Robust Nonlinear Control*, vol. 15, no. 17, pp. 873 – 901, Nov. 2005.

[32] J. M. Gomes da Silvar Jr., D. Limon, T. Alamo, and E. F. Camacho, "Output feedback for discrete-time systems with amplitude and rate constrained actuators," in *Advanced Strategies in Control Systems with Input and Output Constraints*, ser. Lect. Notes Control Inf. Sci., S. Tarbouriech, G. Garcia, and A. H. Glattfelder, Eds. Berlin, Germany: Springer, 2007, vol. 346, pp. 369 – 396.

[33] L. A. Zadeh and C. A. Desoer, *Linear System Theory: The State Space Approach.* Mineola, NY: Dover, 1991.

[34] E. D. Sontag, "An algebraic approach to bounded controllability of linear systems," *Int. J. Control*, vol. 39, no. 1, pp. 181 – 188, Jan. 1984.

[35] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 49 – 95, Mar. 1996.

[36] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, ser. SIAM Stud. Appl. Math. Philadelphia, PA: SIAM, 1994, vol. 15. [Online]. Available: http://www.stanford.edu/~boyd/lmibook/index.html

[37] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[38] J.-J. Slotine and W. Li, *Applied Nonlinear Control.* Upper Saddle River, NJ: Prentice Hall, 1991.

[39] S. Sastry, *Nonlinear Systems: Analysis, Stability, and Control*, ser. Interdiscip. Appl. Math. New York, NY: Springer, 1999, vol. 10.

[40] R. Ortega, A. Loría, P. J. Nicklasson, and H. Sira-Ramírez, *Passivity-based Control of Euler-Lagrange Systems*, ser. Commun. Control Eng. London, Great Britain: Springer, 1998.

[41] A. van der Schaft, L_2-*Gain and Passivity Techniques in Nonlinear Control*, 2nd ed., ser. Commun. Control Eng. London, Great Britain: Springer, 2000.

[42] R. C. Nelson, *Flight Stability and Automatic Control*, 2nd ed., ser. Aerosp. Sci. Technol. Boston, MA: McGraw-Hill, 1998.

[43] S. M. LaValle, *Planning Algorithms.* New York, NY: Cambridge University Press, 2006. [Online]. Available: http://planning.cs.uiuc.edu

[44] A. Isidori, *Nonlinear Control Systems*, 3rd ed., ser. Commun. Control Eng. London, Great Britain: Springer, 1995.

[45] ——, *Nonlinear Control Systems II*, ser. Commun. Control Eng. London, Great Britain: Springer, 1999.

[46] W. M. Haddad and V. Chellaboina, *Nonlinear Dynamical Systems and Control: A Lyapunov-based Approach.* Princeton, NJ: Princeton University Press, 2008.

[47] T. Hu, A. N. Pitsillides, and Z. Lin, "Null controllability and stabilization of linear systems subject to asymmetric actuator saturation," in *Actuator Saturation Control*, ser. Control Eng., V. Kapila and K. M. Grigoriadis, Eds. New York, NY: Marcel Dekker, 2002, ch. 3, pp. 47 – 76.

[48] B. G. Romanchuk and M. C. Smith, "Incremental gain analysis of linear systems with bounded controls and its application to the antiwindup problem," in *Proc. 35th IEEE Conf. Decision and Control*, vol. 3, Kobe, Japan, Dec. 1996, pp. 2942 – 2947.

[49] N. Kapoor, A. R. Teel, and P. Daoutidis, "An anti-windup design for linear systems with input saturation," *Automatica*, vol. 34, no. 5, pp. 559 – 574, May 1998.

[50] H. J. Sussmann, "Existence and uniqueness of minimal realizations of nonlinear systems," *Math. Syst. Theory*, vol. 10, no. 1, pp. 263 – 284, 1977.

[51] R. Hanus, M. Kinnaert, and J.-L. Henrotte, "Conditioning technique, a general anti-windup and bumpless transfer method," *Automatica*, vol. 23, no. 6, pp. 729 – 739, Nov. 1987.

[52] R. Hanus and Y. Peng, "Conditioning technique for controllers with time delays," *IEEE Trans. Autom. Control*, vol. 37, no. 5, pp. 689 – 692, May 1992.

[53] H. A. Fertik and C. W. Ross, "Direct digital control algorithm with anti-windup feature," *ISA Trans.*, vol. 6, no. 4, pp. 317 – 328, 1967.

[54] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, ser. Classics Appl. Math. Philadelphia, PA: SIAM, 1996, vol. 16.

[55] J.-K. Park and C.-H. Choi, "Author's reply," *IEEE Trans. Autom. Control*, vol. 41, no. 10, pp. 1550 – 1551, Oct. 1996.

[56] J.-P. Calvet and Y. Arkun, "Feedforward and feedback linearization of nonlinear systems and its implementation using internal model control (IMC)," *Ind. Eng. Chem. Res.*, vol. 27, no. 10, pp. 1822 – 1831, Oct. 1988.

[57] T. A. Kendi and F. J. Doyle III, "An anti-windup scheme for multivariable nonlinear systems," *J. Process Control*, vol. 7, no. 5, pp. 329 – 343, Oct. 1997.

[58] F. J. Doyle III, "An anti-windup input-output linearization scheme for SISO systems," *J. Proc. Control*, vol. 9, no. 3, pp. 213 – 220, Jun. 1999.

[59] N. Kapoor and P. Daoutidis, "An observer-based anti-windup scheme for non-linear systems with input constraints," *Int. J. Control*, vol. 72, no. 1, pp. 18 – 29, Jan. 1999.

[60] W. Wu, "Anti-windup schemes for a constrained continuous stirred tank reactor process," *Ind. Eng. Chem. Res.*, vol. 41, no. 7, pp. 1796 – 1804, 2002.

[61] G. Herrmann, M. C. Turner, P. Menon, D. G. Bates, and I. Postlethwaite, "Anti-windup synthesis for nonlinear dynamic inversion controllers," in *Proc. 5th IFAC Symp. Robust Control Design*, vol. 5, no. 1, Toulouse, France, Jul. 2006.

[62] P. P. Menon, G. Herrmann, M. C. Turner, D. G. Bates, and I. Postlethwaite, "General anti-windup synthesis for input constrained nonlinear systems controlled using nonlinear dynamic inversion," in *Proc. 45th IEEE Conf. Decision and Control*, San Diego, CA, Dec. 2006, pp. 5435 – 5440.

[63] P. P. Menon, G. Herrmann, M. C. Turner, M. Lowenberg, D. Bates, and I. Postlethwaite, "Dynamic wind tunnel rig implementation of nonlinear dynamic inversion based anti-windup scheme," in *Proc. AIAA Guidance Navigation and Control Conf. and Exhibit*, Honolulu, HI, Aug. 2008, AIAA–2008–7166.

[64] P. P. Menon, G. Herrmann, M. Turner, M. Lowenberg, D. Bates, and I. Postlethwaite, "Nonlinear dynamic inversion based anti-windup - an aerospace application," in *Proc. 17th IFAC World Congress*, vol. 17, no. 1, Seoul, Korea, Jul. 2008, pp. 14 156 – 14 161.

[65] S.-S. Yoon, J.-K. Park, and T.-W. Yoon, "Dynamic anti-windup scheme for feedback linearizable nonlinear control systems with saturating inputs," *Automatica*, vol. 44, no. 12, pp. 3176 – 3180, Dec. 2008.

[66] G. Herrmann, P. P. Menon, M. C. Turner, D. G. Bates, and I. Postlethwaite, "Anti-windup synthesis for nonlinear dynamic inversion control schemes," *Int. J. Robust Nonlinear Control*, vol. 20, no. 13, pp. 1465 – 1482, Sep. 2010.

[67] C. I. Byrnes, "Remarks on nonlinear planar control systems which are linearizable by feedback," *Syst. Control Lett.*, vol. 5, no. 6, pp. 363 – 367, May 1985.

[68] P. Rouchon, "Necessary condition and genericity of dynamic feedback linearization," *J. Math. Syst. Estimat. Control*, vol. 4, no. 2, pp. 1 – 14, 1994.

[69] Q. Hu and G. P. Rangaiah, "Anti-windup schemes for uncertain nonlinear systems," *IET Control Theory Appl.*, vol. 147, no. 3, pp. 321 – 329, May 2000.

[70] E. N. Johnson and A. J. Calise, "Neural network adaptive control of systems with input saturation," in *Proc. American Control Conf.*, Arlington, VA, Jun. 2001, pp. 3527 – 3532.

[71] ——, "Limited authority adaptive flight control for reusable launch vehicles," *J. Guid. Control Dyn.*, vol. 26, no. 6, pp. 906 – 913, Nov. – Dec. 2003.

[72] H. M. Do, T. Başar, and J. Y. Choi, "An anti-windup design for single input adaptive control systems in strict feedback form," in *Proc. American Control Conf.*, vol. 3, Boston, MA, Jun./Jul. 2004, pp. 2551 – 2556.

[73] M. Yokoyama, G.-N. Kim, and M. Tsuchiya, "Integral sliding mode control with anti-windup compensation and its application to a power assist system," *J. Vib. Control*, vol. 16, no. 4, pp. 503 – 512, Apr. 2010.

[74] A. R. Teel and N. Kapoor, “The $\mathcal{L}_2$ anti-windup problem: Its definition and solution,” in *Proc. European Control Conf.*, Brussels, Belgium, Jul. 1997.

[75] M. Soroush and S. Valluri, “Calculation of optimal feasible controller output in multivariable processes with input constraints,” in *Proc. American Control Conf.*, vol. 5, Albuquerque, NM, Jun. 1997, pp. 3475 – 3479.

[76] S. Valluri and M. Soroush, “Analytical control of SISO nonlinear processes with input constraints,” *AIChE J.*, vol. 44, no. 1, pp. 116 – 130, Jan. 1998.

[77] M. Soroush and S. Valluri, “Optimal directionality compensation in processes with input saturation non-linearities,” *Int. J. Control*, vol. 72, no. 17, pp. 1555 – 1564, Nov. 1999.

[78] M. Soroush and N. Mehranbod, “Optimal compensation for directionality in processes with a saturating actuator,” *Comput. Chem. Eng.*, vol. 26, no. 11, pp. 1633 – 1641, Nov. 2002.

[79] M. Soroush and P. Daoutidis, “Optimal windup and directionality compensation in input-constrained nonlinear systems,” in *Actuator Saturation Control*, ser. Control Eng., V. Kapila and K. M. Grigoriadis, Eds. New York, NY: Marcel Dekker, 2002, ch. 9, pp. 227 – 246.

[80] S. Valluri and M. Soroush, “A non-linear controller design method for processes with saturating actuators,” *Int. J. Control*, vol. 76, no. 7, pp. 698 – 716, Jan. 2003.

[81] M. Soroush, S. Valluri, and N. Mehranbod, “Nonlinear control of input-constrained systems,” *Comput. Chem. Eng.*, vol. 30, no. 1, pp. 158 – 181, Nov. 2005.

[82] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, “Constrained model predictive control: Stability and optimality,” *Automatica*, vol. 36, no. 6, pp. 789 – 814, Jun. 2000.

[83] P. Kapasouris, M. Athans, and G. Stein, “Design of feedback control systems for stable plants with saturating actuators,” in *Proc. 27th IEEE Conf. Decision and Control*, vol. 1, Austin, TX, Dec. 1988, pp. 469 – 479.

[84] E. G. Gilbert and K. T. Tan, “Linear systems with state and control constraints: The theory and application of maximal output admissible sets,” *IEEE Trans. Autom. Control*, vol. 36, no. 9, pp. 1008 – 1020, Sep. 1991.

[85] A. Bemporad and E. Mosca, “Constraint fulfilment in feedback control via predictive reference management,” in *Proc. 3rd IEEE Conf. Control Applications*, vol. 3, Glasgow, Scotland, Aug. 1994, pp. 1909 – 1914.

[86] ——, “Constraint fulfilment in control systems via predictive reference management,” in *Proc. 33th IEEE Conf. Decision and Control*, vol. 3, Lake Buena Vista, FL, Dec. 1994, pp. 3017 – 3022.

[87] E. G. Gilbert, I. Kolmanovsky, and K. T. Tan, “Nonlinear control of discrete-time linear systems with state and control constraints: A reference governor with global convergence properties,” in *Proc. 33rd IEEE Conf. Decision and Control*, vol. 1, Lake Buena Vista, FL, Dec. 1994, pp. 144 – 149.

[88] A. Bemporad and E. Mosca, “Nonlinear predictive reference governor for constrained control systems,” in *Proc. 34th IEEE Conf. Decision and Control*, vol. 2, New Orleans, LA, Dec. 1995, pp. 1205 – 1210.

[89] E. G. Gilbert and I. Kolmanovsky, “Discrete-time reference governors for systems with state and control constraints and disturbance inputs,” in *Proc. 34th IEEE Conf. Decision and Control*, vol. 2, New Orleands, LA, Dec. 1995, pp. 1189 – 1194.

[90] E. G. Gilbert, I. Kolmanovsky, and K. T. Tan, “Discrete-time reference governors and the nonlinear control of systems with state and control constrains,” *Int. J. Robust Nonlinear Control*, vol. 5, no. 5, pp. 487 – 504, 1995.

[91] A. Casavola and E. Mosca, “Reference governor for constrained uncertain linear systems subject to bounded input disturbances,” in *Proc. 35th IEEE Conf. Decision and Control*, vol. 3, Koba, Japan, Dec. 1996, pp. 3531 – 3536.

[92] A. Bemporad, A. Casavola, and E. Mosca, “Nonlinear control of constrained linear systems via predictive reference management,” *IEEE Trans. Autom. Control*, vol. 42, no. 3, pp. 340 – 349, Mar. 1997.

[93] E. G. Gilbert and I. Kolmanovsky, “Fast reference governors for systems with state and control constraints and disturbance inputs,” *Int. J. Robust Nonlinear Control*, vol. 9, no. 15, pp. 1117 – 1141, Dec. 1999.

[94] K. Kogiso and K. Hirata, “Reference governor for constrained systems with time-varying references,” *Robot. Auton. Syst.*, vol. 57, no. 3, pp. 289 – 295, Mar. 2009.

[95] A. Bemporad, “Reference governor for constrained nonlinear systems,” *IEEE Trans. Autom. Control*, vol. 43, no. 3, pp. 415 – 419, Mar. 1998.

[96] D. Angeli and E. Mosca, “Command governors for constrained nonlinear systems,” *IEEE Trans. Autom. Control*, vol. 44, no. 4, pp. 816 – 820, Apr. 1999.

[97] E. G. Gilbert and I. V. Kolmanovsky, “Set-point control of nonlinear systems with state and control constraints: A Lyapunov-function, reference-governor approach,” in *Proc. 38th IEEE Conf. Decision and Control*, vol. 3, Phoenix, AZ, Dec. 1999, pp. 2507 – 2512.

[98] R. H. Miller, I. Komanovsky, E. G. Gilbert, and P. D. Washabaugh, “Control of constrained nonlinear systems: A case study,” *IEEE Control Syst. Mag.*, vol. 20, no. 1, pp. 23 – 32, Feb. 2000.

[99] T. Hatanaka and K. Takaba, “Output feedback reference governor for nonlinear systems,” in *Proc. 44th IEEE Conf. Decision and Control & European Control Conf.*, Seville, Spain, Dec. 2005, pp. 7558 – 7563.

[100] J. B. Rosen, “The gradient projection method for nonlinear programming. part I. linear constraints,” *J. Soc. Ind. Appl. Math.*, vol. 8, no. 1, pp. 181 – 217, Mar. 1960.

[101] ——, “The gradient projection method for nonlinear programming. part II. nonlinear constraints,” *J. Soc. Ind. Appl. Math.*, vol. 9, no. 4, pp. 514 – 532, Dec. 1961.

[102] J.-J. E. Slotine and J. A. Coetsee, “Adaptive sliding controller synthesis for non-linear systems,” *Int. J. Control*, vol. 43, no. 6, pp. 1631 – 1651, Jun. 1986.

[103] L. Hou and A. N. Michel, “Asymptotic stability of systems with saturation constraints,” *IEEE Trans. Autom. Control*, vol. 43, no. 8, pp. 1148 – 1154, Aug. 1998.

[104] H. Fang and Z. Lin, “Stability analysis for linear systems under state constraints,” *IEEE Trans. Autom. Control*, vol. 49, no. 6, pp. 950 – 955, Jun. 2004.

[105] X. Ji, T. Liu, and M. Ren, “Stability analysis for continuous-time planar linear systems with state saturation,” in *Proc. Chinese Conf. Decision and Control*, Shandong, China, Jul. 2008, pp. 4355 – 4359.

[106] W. Guan and G. H. Yang, “Analysis and design of output feedback control systems in the presence of state saturation,” in *Proc. American Control Conf.*, St. Louis, MO, Jun. 2009, pp. 5677 – 5682.

[107] P. Dupuis and A. Nagurney, “Dynamical systems and variational inequalities,” *Ann. Oper. Res.*, vol. 44, no. 1, pp. 7 – 42, Feb. 1993.

[108] D. Zhang and A. Nagurney, “On the stability of projected dynamical systems,” *J. Optim. Theory Appl.*, vol. 85, no. 1, pp. 97 – 124, Apr. 1995.

[109] A. Nagurney and D. Zhang, *Projected Dynamical Systems and Variational Inequalities with Applications*, ser. Int. Ser. Oper. Res. Manag. Sci. Norwell, MA: Kluwer, 1996.

[110] M.-G. Cojocaru and L. B. Jonker, “Existence of solutions to projected differential equations in Hilbert spaces,” *Proc. Amer. Math. Soc.*, vol. 132, no. 1, pp. 183 – 193, Jan. 2004.

[111] D. G. Eksten and G. F. Ohlson, “Automatic control circuit utilizing input and internal signals controlling reset for providing improved step response,” U.S. Patent 3 219 936, Nov. 23, 1965. [Online]. Available: http://www.google.com/patents?id=AQdnAAAAEBAJ

[112] J. O. Jacques, “Input for digital controller,” U.S. Patent 3 387 282, Jun. 4, 1968. [Online]. Available: http://www.google.com/patents?id=0yVnAAAAEBAJ

[113] J. O. Jacques, D. Montgomery, and P. A. Kuckein, “Digital controller with automatic balance and manually adjusted operating point,” U.S. Patent 3 479 493, Nov. 18, 1969. [Online]. Available: http://www.google.com/patents?id=1a1vAAAAEBAJ

[114] J. O. Jacques, “Digital control system with integral clamping,” U.S. Patent 3 495 074, Feb. 10, 1970. [Online]. Available: http://www.google.com/patents?id=l7JZAAAAEBAJ

[115] J. O. Jacques, P. A. Kuckein, and R. K. Oswald, “Closed loop controller having digital integrator with variable gain,” U.S. Patent 3 513 302, May 19, 1970. [Online]. Available: http://www.google.com/patents?id=0Pd0AAAAEBAJ

[116] J. W. Slover, “Leak preventing control for heat exchangers,” U.S. Patent 3 400 753, Sep. 10, 1968. [Online]. Available: http://www.google.com/patents?id=Yg9jAAAAEBAJ

[117] F. B. Davis III and C. W. Ross, "Load-frequency control system without proportional windup," U.S. Patent 3 525 857, Aug. 25, 1970. [Online]. Available: http://www.google.com/patents?id=zYd1AAAAEBAJ

[118] G. Labinaz, M. M. Bayoumi, and K. Rudie, "A survey of modeling and control of hybrid systems," *Annu. Rev. Control*, vol. 21, pp. 79 – 92, 1997.

[119] H. Lin and P. J. Antsaklis, "Stability and stabilization of switched linear systems: A survey of recent results," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 308 – 322, Feb. 2009.

[120] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, ser. Math. Appl. Dordrecht, Netherlands: Kluwer Academic Publishers, 1988.

[121] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., ser. Optim. Comput. Belmont, MA: Athena Scientific, 1999.

[122] P. A. Ioannou and J. Sun, *Robust Adaptive Control.* Upper Saddle River, NJ: Prentice Hall, 1996. [Online]. Available: http://www-rcf.usc.edu/~ioannou/Robust_Adaptive_Control.htm

[123] J.-B. Pomet and L. Praly, "Adaptive nonlinear regulation: Estimation from the Lyapunov equation," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 729 – 740, Jun. 1992.

[124] R. A. Horn and C. R. Johnson, *Matrix Analysis.* New York, NY: Cambridge University Press, 1985.

[125] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review.* Mineola, NY: Dover, 1968.

[126] J. Dattorro, *Convex Optimization & Euclidean Distance Geometry.* Palo Alto, CA: Meboo, 2005, version 2010.06.04. [Online]. Available: http://meboo.convexoptimization.com/Meboo.html

[127] J. Teo and J. P. How, "Anti-windup compensation for nonlinear systems via gradient projection: Application to adaptive control," in *Proc. 48th IEEE Conf. Decision and Control & 28th Chinese Control Conf.*, Shanghai, China, Dec. 2009, pp. 6910 – 6916.

[128] G. Grimm, J. Hatfield, I. Postlethwaite, A. R. Teel, M. C. Turner, and L. Zaccarian, "Antiwindup for stable linear systems with input saturation: An LMI-based synthesis," *IEEE Trans. Autom. Control*, vol. 48, no. 9, pp. 1509 – 1525, Sep. 2003.

[129] B. Brogliato, R. Lozano, B. Maschke, and O. Egeland, *Dissipative Systems Analysis and Control: Theory and Applications*, 2nd ed., ser. Commun. Control Eng. London, United Kingdom: Springer, 2007.

[130] J. A. De Doná, G. C. Goodwin, and M. M. Seron, "Anti-windup and model predictive control: Reflections and connections," *Eur. J. Control*, vol. 6, no. 5, pp. 467 – 477, 2000.

[131] L. Lu and Z. Lin, “A switching anti-windup design using multiple Lyapunov functions,” *IEEE Trans. Autom. Control*, vol. 55, no. 1, pp. 142 – 148, Jan. 2010.

[132] R. J. Mantz and H. De Battista, “Sliding mode compensation for windup and direction of control problems in two-input-two-output proportional-integral controllers,” *Ind. Eng. Chem. Res.*, vol. 41, no. 13, pp. 3179 – 3185, 2002.

[133] R. J. Mantz, H. De Battista, and F. D. Bianchi, “Sliding mode conditioning for constrained processes,” *Ind. Eng. Chem. Res.*, vol. 43, no. 26, pp. 8251 – 8256, 2004.

[134] F. Garelli, R. J. Mantz, and H. De Battista, “Collective sliding-mode technique for multivariable bumpless transfer,” *Ind. Eng. Chem. Res.*, vol. 41, no. 8, pp. 2721 – 2727, 2008.

[135] J. K. Hale, *Ordinary Differential Equations*, 2nd ed. Mineola, NY: Dover, 1997.

[136] A. A. Andronov, A. A. Vitt, and S. E. Khaikin, *Theory of Oscillators*, ser. Int. Ser. Monogr. Phys. Oxford, England: Pergamon Press, 1966, vol. 4.

[137] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1993.

[138] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, ser. Int. Ser. Pure Appl. Math. New York, NY: McGraw-Hill, 1955.

[139] O. Hájek, *Control Theory in the Plane*, 2nd ed., ser. Lect. Notes Control Inf. Sci. Berlin, Germany: Springer, 2009, vol. 153.

[140] R. Mantri, A. Saberi, and V. Venkatasubramanian, “Stability analysis of continuous time planar systems with state saturation nonlinearity,” *IEEE Trans. Circuits Syst. I*, vol. 45, no. 9, pp. 989 – 993, Sep. 1998.

[141] J. Alvarez, R. Suárez, and J. Alvarez, “Planar linear systems with single saturated feedback,” *Syst. Control Lett.*, vol. 20, no. 4, pp. 319 – 326, Apr. 1993.

[142] J.-Y. Favez, P. Mullhaupt, B. Srinivasan, and D. Bonvin, “Attraction region of planar linear systems with one unstable pole and saturated feedback,” *J. Dyn. Control Syst.*, vol. 12, no. 3, pp. 331 – 355, Jul. 2006.

[143] T. Hu, L. Qiu, and Z. Lin, “Stabilization of LTI systems with planar anti-stable dynamics using saturated linear feedback,” in *Proc. 37th IEEE Conf. Decision and Control*, Tampa, FL, Dec. 1998, pp. 389 – 394.

[144] T. Hu, Z. Lin, and L. Qiu, “Stabilization of exponentially unstable linear systems with saturating actuators,” *IEEE Trans. Autom. Control*, vol. 46, no. 6, pp. 973 – 979, Jun. 2001.

[145] M. L. Corradini, A. Cristofaro, and F. Giannoni, “Sharp estimates on the region of attraction of planar linear systems with bounded controls,” in *Proc. 48th IEEE Conf. Decision and Control & 28th Chinese Control Conf.*, Shanghai, China, Dec. 2009, pp. 5345 – 5350.

[146] F. Liu, G. T.-C. Chiu, E. S. Hamby, and Y. Eun, "Time maximum control for a class of single-input planar affine control systems and constraints," in *Proc. 48th IEEE Conf. Decision and Control & 28th Chinese Control Conf.*, Shanghai, China, Dec. 2009, pp. 5045 – 5050.

[147] J. A. Ball, M. V. Day, and P. Kachroo, "Robust feedback control of a single server queueing system," *Math. Control Signal Syst.*, vol. 12, no. 4, pp. 307 – 345, Nov. 1999.

[148] J. A. Ball, M. V. Day, T. Yu, and P. Kachroo, "Robust L_2-gain control for nonlinear systems with projection dynamics and input constraints: an example from traffic control," *Automatica*, vol. 35, no. 3, pp. 429 – 444, Mar. 1999.

[149] K. Kuhnen and P. Krejci, "Identification of linear error-models with projected dynamical systems," *Math. Comput. Model. Dyn. Syst.*, vol. 10, no. 1, pp. 59 – 91, Mar. 2004.

[150] K. Kuhnen and P. Krejcí, "An adaptive gradient law with projection for non-smooth convex boundaries," *Eur. J. Control*, vol. 12, no. 6, pp. 606 – 619, 2006.

[151] B. Brogliato, A. Daniilidis, C. Lemaréchal, and V. Acary, "On the equivalence between complementarity systems, projected systems and differential inclusions," *Syst. Control Lett.*, vol. 55, no. 1, pp. 45 – 51, Jan. 2006.

[152] A. van der Schaft and H. Schumacher, *An Introduction to Hybrid Dynamical Systems*, ser. Lect. Notes Control Inf. Sci. London, Great Britain: Springer, 2000, vol. 251.

[153] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control: Analysis and Design.* West Sussex, England: Wiley, 1996.

[154] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., ser. Int. Ser. Pure Appl. Math. New York, NY: McGraw-Hill, 1976.

[155] J. P. LaSalle, "Some extensions of Liapunov's second method," *IRE Trans. Circuit Theory*, vol. 7, no. 4, pp. 520 – 527, Dec. 1960.

[156] S. Weissenberger, "Stability regions of large-scale systems," *Automatica*, vol. 9, no. 6, pp. 653 – 663, Nov. 1973.

[157] R. Genesio, M. Tartaglia, and A. Vicino, "On the estimation of asymptotic stability regions: State of the art and new proposals," *IEEE Trans. Autom. Control*, vol. 30, no. 8, pp. 747 – 755, Aug. 1985.

[158] J. Zaborszky, G. Huang, B. Zheng, and T.-C. Leung, "On the phase portrait of a class of large nonlinear dynamic systems such as the power system," *IEEE Trans. Autom. Control*, vol. 33, no. 1, pp. 4 – 15, Jan. 1988.

[159] H.-D. Chiang, M. W. Hirsch, and F. F. Wu, "Stability regions of nonlinear autonomous dynamical systems," *IEEE Trans. Autom. Control*, vol. 33, no. 1, pp. 16 – 27, Jan. 1988.

[160] H.-D. Chiang and J. S. Thorp, "Stability regions of nonlinear dynamical systems: A constructive methodology," *IEEE Trans. Autom. Control*, vol. 34, no. 12, pp. 1229 – 1241, Dec. 1989.

[161] A. Levin, “An analytical method of estimating the domain of attraction for polynomial differential equations,” *IEEE Trans. Autom. Control*, vol. 39, no. 12, pp. 2471 – 2475, Dec. 1994.

[162] H.-D. Chiang and L. Fekih-Ahmed, “Quasi-stability regions of nonlinear dynamical systems: Optimal estimations,” *IEEE Trans. Circuits Syst. I*, vol. 43, no. 8, pp. 636 – 643, Aug. 1996.

[163] L. Gruyitch, J.-P. Richard, P. Borne, and J.-C. Gentina, *Stability Domains*, ser. Nonlinear Syst. Aviat. Aerosp. Aeronaut. Astronaut. Boca Raton, FL: Chapman & Hall, 2004, vol. 1.

[164] W. Tan and A. Packard, “Stability region analysis using polynomial and composite polynomial Lyapunov functions and sum-of-squares programming,” *IEEE Trans. Autom. Control*, vol. 53, no. 2, pp. 565 – 571, Mar. 2008.

[165] G. Chesi, “Estimating the domain of attraction for non-polynomial systems via LMI optimizations,” *Automatica*, vol. 45, no. 6, pp. 1536 – 1541, Jun. 2009.

[166] A. I. Zečević and D. D. Šiljak, “Estimating the region of attraction for large-scale systems with uncertainties,” *Automatica*, vol. 46, no. 2, pp. 445 – 451, Feb. 2010.

[167] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed., ser. Artif. Intell. Upper Saddle River, NJ: Prentice Hall, 2003.

[168] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, ser. Appl. Math. Sci. New York, NY: Springer, 2002, vol. 42.

[169] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, ser. Die Grundlehren der mathematischen Wissenschaften. Berlin, Germany: Springer, 1970, vol. 163.

[170] M. Tenenhaus, “Canonical analysis of two convex polyhedral cones and applications,” *Psychometrika*, vol. 53, no. 4, pp. 503 – 524, Dec. 1988.

[171] R. L. Dykstra, “An algorithm for restricted least squares regression,” *J. Am. Stat. Assoc.*, vol. 78, no. 384, pp. 837 – 842, Dec. 1983.

[172] N. Gaffke and R. Mathar, “A cyclic projection algorithm via duality,” *Metrika*, vol. 36, no. 1, pp. 29 – 54, Dec. 1989.

[173] T. Hyunh, C. Lassez, and J.-L. Lassez, “Practical issues on the projection of polyhedral sets,” *Ann. Math. Artif. Intell.*, vol. 6, no. 4, pp. 295 – 315, Dec. 1992.

[174] L. M. Bregman, Y. Censor, S. Reich, and Y. Zepkowitz-Malachi, “Finding the projection of a point onto the intersection of convex sets via projections onto half-spaces,” *J. Approx. Theory*, vol. 124, no. 2, pp. 194 – 218, Oct. 2003.

[175] A. B. Németh and S. Z. Németh, “How to project onto an isotone projection cone,” *Linear Alg. Appl.*, vol. 433, no. 1, pp. 41 – 51, Jul. 2010.

[176] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, United Kingdom: Cambridge University Press, 2004.

[177] R. T. Rockafellar, *Convex Analysis*, ser. Princeton Landmarks Math. Princeton, NJ: Princeton University Press, 1970.

[178] F. A. Valentine, *Convex Sets*, ser. McGraw-Hill Ser. High. Math. New York, NY: McGraw-Hill, 1964.

[179] A. Baíllo and A. Cuevas, "On the estimation of a star-shaped set," *Adv. Appl. Probab.*, vol. 33, no. 4, pp. 717 – 726, Dec. 2001.

[180] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, ser. Grundlehren Text Ed. Berlin, Germany: Springer, 2001.

[181] A. N. Michel, K. Wang, and B. Hu, *Qualitative Theory of Dynamical Systems: The Role of Stability Preserving Mappings*, 2nd ed., ser. Pure Appl. Math. New York, NY: Marcel Dekker, 2001, vol. 239.

[182] D. Carlson, "Block diagonal semistability factors and Lyapunov semistability of block triangular matrices," *Linear Alg. Appl.*, vol. 172, pp. 1 – 25, Jul. 1992.

[183] G. Balas, R. Chiang, A. Packard, and M. Safonov, *Robust Control Toolbox^{TM} 3: User's Guide*, MathWorks, Natick, MA, 2010.

[184] F. Morabito, "Sintesi e validazione di tecniche di controllo anti-windup per sistemi Eulero–Lagrange con saturazione sugli ingressi," Master's thesis, Univ. Rome "Tor Vergata", Rome, Italy, 2002.

[185] P. R. Bélanger, *Control Engineering: A Modern Approach.* Orlando, FL: Saunders College Publishing, 1995.

[186] *Optimization Toolbox^{TM} 5: User's Guide*, Mathworks, Natick, MA, 2010.

[187] K. J. Åström and L. Rundqwist, "Integrator windup and how to avoid it," in *Proc. American Control Conf.*, Pittsburgh, PA, Jun. 1989, pp. 1693 – 1698.

[188] K. J. Åström and T. Hägglund, *PID Controllers: Theory, Design, and Tuning*, 2nd ed. Research Triangle Park, NC: Instrument Society of America, 1995.

[189] A. O'Dwyer, *Handbook of PI and PID Controller Tuning Rules*, 2nd ed. London, United Kingdom: Imperial College Press, 2006.

[190] G. J. Silva, A. Datta, and S. P. Bhattacharyya, *PID Controllers for Time-Delay Systems*, ser. Control Eng. Boston, MA: Birkhäuser, 2005.

[191] Q.-G. Wang, Z. Ye, W.-J. Cai, and C.-C. Hang, *PID Control for Multivariable Processes*, ser. Lect. Notes Control Inf. Sci. Berlin, Germany: Springer, 2008, vol. 373.

[192] Y. Choi and W. K. Chung, *PID Trajectory Tracking Control for Mechanical Systems*, ser. Lect. Notes Control Inf. Sci. Berlin, Germany: Springer, 2004, vol. 298.