



University of Pennsylvania
ScholarlyCommons


Publicly Accessible Penn Dissertations

2020

Constrained Learning And Inference

Luiz Fernando De Oliveira Chamon
University of Pennsylvania

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Computer Sciences Commons](#), [Electrical and Electronics Commons](#), and the [Mathematics Commons](#)

Recommended Citation

De Oliveira Chamon, Luiz Fernando, "Constrained Learning And Inference" (2020). *Publicly Accessible Penn Dissertations*. 3818.
<https://repository.upenn.edu/edissertations/3818>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/3818>
For more information, please contact repository@pobox.upenn.edu.

Constrained Learning And Inference

Abstract

Data and learning have become core components of the information processing and autonomous systems upon which we increasingly rely on to select job applicants, analyze medical data, and drive cars. As these systems become ubiquitous, so does the need to curtail their behavior. Left untethered, they are susceptible to tampering (adversarial examples) and prone to prejudiced and unsafe actions. Currently, the response of these systems is tailored by leveraging domain expert knowledge to either construct models that embed the desired properties or tune the training objective so as to promote them. While effective, these solutions are often targeted to specific behaviors, contexts, and sometimes even problem instances and are typically not transferable across models and applications. What is more, the growing scale and complexity of modern information processing and autonomous systems renders this manual behavior tuning infeasible. Already today, explainability, interpretability, and transparency combined with human judgment are no longer enough to design systems that perform according to specifications.

The present thesis addresses these issues by leveraging constrained statistical optimization. More specifically, it develops the theoretical underpinnings of constrained learning and constrained inference to provide tools that enable solving statistical problems under requirements. Starting with the task of learning under requirements, it develops a generalization theory of constrained learning akin to the existing unconstrained one. By formalizing the concept of probability approximately correct constrained (PACC) learning, it shows that constrained learning is as hard as its unconstrained learning and establishes the constrained counterpart of empirical risk minimization (ERM) as a PACC learner. To overcome challenges involved in solving such non-convex constrained optimization problems, it derives a dual learning rule that enables constrained learning tasks to be tackled by through unconstrained learning problems only. It therefore concludes that if we can deal with classical, unconstrained learning tasks, then we can deal with learning tasks with requirements.

The second part of this thesis addresses the issue of constrained inference. In particular, the issue of performing inference using sparse nonlinear function models, combinatorial constrained with quadratic objectives, and risk constraints. Such models arise in nonlinear line spectrum estimation, functional data analysis, sensor selection, actuator scheduling, experimental design, and risk-aware estimation. Although inference problems assume that models and distributions are known, each of these constraints pose serious challenges that hinder their use in practice. Sparse nonlinear functional models lead to infinite dimensional, non-convex optimization programs that cannot be discretized without leading to combinatorial, often NP-hard, problems. Rather than using surrogates and relaxations, this work relies on duality to show that despite their apparent complexity, these models can be fit efficiently, i.e., in polynomial time. While quadratic objectives are typically tractable (often even in closed form), they lead to non-submodular optimization problems when subject to cardinality or matroid constraints. While submodular functions are sometimes used as surrogates, this work instead shows that quadratic functions are close to submodular and can also be optimized near-optimally. The last chapter of this thesis is dedicated to problems involving risk constraints, in particular, bounded predictive mean square error variance estimation. Despite being non-convex, such problems are equivalent to a quadratically constrained quadratic program from which a closed-form estimator can be extracted.

These results are used throughout this thesis to tackle problems in signal processing, machine learning, and control, such as fair learning, robust learning, nonlinear line spectrum estimation, actuator scheduling, experimental design, and risk-aware estimation. Yet, they are applicable much beyond these illustrations to perform safe reinforcement learning, sensor selection, multiresolution kernel estimation, and wireless resource allocation, to name a few.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Electrical & Systems Engineering

First Advisor

Alejandro Ribeiro

Keywords

Combinatorial optimization, Constrained learning, Constrained optimization, Functional optimization, Inference, Machine learning

Subject Categories

Computer Sciences | Electrical and Electronics | Mathematics

CONSTRAINED LEARNING AND INFERENCE

Luiz Fernando de Oliveira Chamon

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania

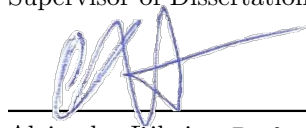
in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation



Alejandro Ribeiro, Professor of Electrical and Systems Engineering

Graduate Group Chairperson



Victor Preciado, Associate Professor of Electrical and Systems Engineering

Dissertation Committee

George J. Pappas, UPS Foundation Professor and Chair of the Department of Electrical and Systems Engineering, University of Pennsylvania

Yonina C. Eldar, Professor of Electrical Engineering, Weizmann Institute of Science

Hamed Hassani, Assistant Professor of Electrical and Systems Engineering, University of Pennsylvania

Nikolai Matni, Assistant Professor of Electrical and Systems Engineering, University of Pennsylvania

CONSTRAINED LEARNING AND INFERENCE

COPYRIGHT

2020

Luiz Fernando de Oliveira Chamon

A certa(s) personagem(s) desvanecida(s)

Acknowledgments

I would like to thank my advisor, Alejandro Ribeiro, without whose guidance, advice, and constant state of disagreement over these past five years this thesis would not exist. His support and understanding in all matters were fundamental in shaping the researcher I am today. Thank you also for shamelessly lying on my recommendation letters by saying “I am an easy person to work with.”

I would like to thank my committee, Yonina C. Eldar, George J. Pappas, Hamed Hassani, and Nikolai Matni, for helping shape the ideas and contributions in this work and beyond. While I have only had the pleasure to formally collaborate with Yonina and George during my Ph.D. (and plan to continue to do so after), I hope to soon find something interesting enough to write with Hamed and Nikolai.

I was fortunate enough to have collaborators outside of my committee, both within and without my lab. Santiago Paternain, Miguel Calvo-Fullana, Mark Eisen, Dionysios Kalogerias, Anastasios Tsiamis, Luana Ruiz, Vinícius Silva, Maria Peifer, Behnaz Arzani, Boon Thau Loo: thank you all for allowing me to contribute to your body of work. Though we have never written papers together, I would like to also thank all the staff of the ESE office, past and present, who have certainly collaborated their fair share to my Ph.D.

My colleagues at Penn who have indulged me in long conversations, both technical and as far from technical as possible: this work would not have been possible without you. While I will forget many, Fernando, Tasos, Markos, Caiman, Papi, Misha, Segarra, Aryan, Weiyu, Harshat, Mohammad, Alp, Alëna, Kelsey, Mariliza, Jacob, Hadi, Shahin, and Ling come to mind right now.

After all the house parties and barbecues that kept me (arguably) sane during this Ph.D., I could not leave out my Philadelphia friends from these acknowledgments. My little sister Melissa and all of my adopted (and adoptive) Greek family in Philadelphia and elsewhere: *mou leipete hdh kai akoma*.

The Philadelphia Open Soccer crew. Those who are here and those who have left. Thank you.

Finally, I would like to thank my always present parents, Edna and Marco, and brother, Paulo.

Louvado seja Deus que o acabei.

ABSTRACT

CONSTRAINED LEARNING AND INFERENCE

Luiz F. O. Chamon
Alejandro Ribeiro

Data and learning have become core components of the information processing and autonomous systems upon which we increasingly rely on to select job applicants, analyze medical data, and drive cars. As these systems become ubiquitous, so does the need to curtail their behavior. Left untethered, they are susceptible to tampering (adversarial examples) and prone to prejudiced and unsafe actions. Currently, the response of these systems is tailored by leveraging domain expert knowledge to either construct models that *embed* the desired properties or *tune* the training objective so as to promote them. While effective, these solutions are often targeted to specific behaviors, contexts, and sometimes even problem instances and are typically not transferable across models and applications. What is more, the growing scale and complexity of modern information processing and autonomous systems renders this manual behavior tuning infeasible. Already today, explainability, interpretability, and transparency combined with human judgment are no longer enough to design systems that perform according to specifications.

The present thesis addresses these issues by leveraging constrained statistical optimization. More specifically, it develops the theoretical underpinnings of constrained learning and constrained inference to provide tools that enable solving statistical problems under requirements. Starting with the task of learning under requirements, it develops a generalization theory of constrained learning akin to the existing unconstrained one. By formalizing the concept of probability approximately correct constrained (PACC) learning, it shows that constrained learning is as hard as its unconstrained learning and establishes the constrained counterpart of empirical risk minimization (ERM) as a PACC learner. To overcome challenges involved in solving such non-convex constrained optimization problems, it derives a dual learning rule that enables constrained learning tasks to be tackled by through unconstrained learning problems only. It therefore concludes that if we can deal with classical, unconstrained learning tasks, then we can deal with learning tasks with requirements.

The second part of this thesis addresses the issue of constrained inference. In particular, the issue of performing inference using sparse nonlinear function models, combinatorial constrained with

quadratic objectives, and risk constraints. Such models arise in nonlinear line spectrum estimation, functional data analysis, sensor selection, actuator scheduling, experimental design, and risk-aware estimation. Although inference problems assume that models and distributions are known, each of these constraints pose serious challenges that hinder their use in practice. Sparse nonlinear functional models lead to infinite dimensional, non-convex optimization programs that cannot be discretized without leading to combinatorial, often NP-hard, problems. Rather than using surrogates and relaxations, this work relies on duality to show that despite their apparent complexity, these models can be fit efficiently, i.e., in polynomial time. While quadratic objectives are typically tractable (often even in closed form), they lead to non-submodular optimization problems when subject to cardinality or matroid constraints. While submodular functions are sometimes used as surrogates, this work instead shows that quadratic functions are close to submodular and can also be optimized near-optimally. The last chapter of this thesis is dedicated to problems involving risk constraints, in particular, bounded predictive mean square error variance estimation. Despite being non-convex, such problems are equivalent to a quadratically constrained quadratic program from which a closed-form estimator can be extracted.

These results are used throughout this thesis to tackle problems in signal processing, machine learning, and control, such as fair learning, robust learning, nonlinear line spectrum estimation, actuator scheduling, experimental design, and risk-aware estimation. Yet, they are applicable much beyond these illustrations to perform safe reinforcement learning, sensor selection, multiresolution kernel estimation, and wireless resource allocation, to name a few.

Keywords. Statistical learning. Inference. Machine learning. Constrained learning. Constrained optimization. Combinatorial optimization. Functional optimization.

Contents

| | |
|--|-------------|
| Acknowledgements | iv |
| Abstract | vi |
| List of Tables | xiii |
| List of Figures | xiv |
| 1 Introduction | 1 |
| 1.1 Motivation | 1 |
| 1.2 Objectives | 2 |
| 1.3 Contributions | 2 |
| 1.4 List of publication | 5 |
| 1.4.1 Preprints | 5 |
| 1.4.2 Journals | 6 |
| 1.4.3 ML conferences | 6 |
| 1.4.4 Control conferences | 6 |
| 1.4.5 Signal processing conferences | 7 |
| 1.5 Notation | 8 |
| 2 Learning and inference under requirements | 10 |
| 2.1 Solving statistical problems with requirements today | 10 |
| 2.2 Constrained learning and inference | 12 |
| 2.2.1 The learning setting | 13 |

| | | |
|-----------|---|-----------|
| 2.2.2 | The inference setting | 14 |
| I | Constrained learning | 19 |
| 3 | Constrained learning theory | 20 |
| 3.1 | Classical learning theory | 20 |
| 3.2 | Probably Approximately Correct Constrained Learning | 22 |
| 3.3 | PACC Learning is as Hard as PAC Learning | 23 |
| 4 | Dual constrained learning | 25 |
| 4.1 | Empirical dual learning | 27 |
| 4.1.1 | The duality gap | 32 |
| 4.1.2 | The parameterization gap | 33 |
| 4.1.3 | The empirical gap | 34 |
| 4.2 | A Constrained Learning Algorithm | 35 |
| 5 | Applications | 38 |
| 5.1 | Fairness | 38 |
| 5.1.1 | Invariance and fair learning | 39 |
| 5.1.2 | Rate-constrained learning | 43 |
| 5.2 | Robustness | 46 |
| II | Constrained inference | 50 |
| 6 | Sparse nonlinear functional models | 51 |
| 6.1 | Sparse functional programs and duality | 53 |
| 6.1.1 | Sparse functional programs | 53 |
| 6.1.2 | The Lagrangian dual of sparse functional programs | 55 |
| 6.2 | Strong duality and its implications | 58 |
| 6.2.1 | Strong duality of sparse functional programs | 58 |
| 6.2.2 | SFPs and L_1 -norm optimization problems | 61 |

| | | |
|----------|---|------------|
| 6.3 | Solving sparse functional programs | 65 |
| 6.4 | Applications | 68 |
| 6.4.1 | Nonlinear line spectrum estimation | 68 |
| 6.4.2 | Robust functional data analysis | 74 |
| 7 | Combinatorial constraints and nonsubmodular optimization | 79 |
| 7.1 | (Approximate) supermodularity | 80 |
| 7.1.1 | Approximately supermodular scalarizations | 84 |
| 7.2 | Approximately supermodular minimization | 88 |
| 7.2.1 | Cardinality constraints | 89 |
| 7.2.2 | Intersection of matroids constraints | 91 |
| 7.3 | The multiset case | 95 |
| 7.4 | Applications | 97 |
| 7.4.1 | Graph Signal Sampling | 97 |
| 7.4.2 | Control input scheduling | 113 |
| 7.4.3 | Experimental design | 123 |
| 8 | Risk-aware minimum mean square error estimation | 131 |
| 8.1 | The risk-constrained MMSE problem | 133 |
| 8.2 | Convex Variational QCQP Reformulation | 135 |
| 8.3 | Risk-Aware MMSE Estimators | 136 |
| 8.4 | Applications | 139 |
| 9 | Concluding remarks | 143 |
| A | Proofs of Part I | 145 |
| A.1 | Proof of Theorem 1 | 145 |
| A.2 | Proof of Theorem 2 | 146 |
| A.2.1 | The approximation gap | 147 |
| A.2.2 | The estimation gap | 148 |
| A.2.3 | The PACC solution | 149 |

| | | |
|----------|--|------------|
| A.2.4 | Proof of Proposition 19: The Approximation Gap | 149 |
| A.2.5 | Proof of Proposition 20: The Estimation Gap | 153 |
| A.3 | Proof of Proposition 1 | 155 |
| A.4 | Duality Gap for Regression | 160 |
| A.5 | Proof of Proposition 2: The Approximation Gap | 163 |
| A.6 | Proof of Proposition 3: The Estimation Gap | 166 |
| A.7 | Proof of Theorem 4 | 168 |
| B | Proofs of Part II | 172 |
| B.1 | Chapter 6: Nonconvex functional models | 172 |
| B.1.1 | Proof of Lemma 2 | 172 |
| B.1.2 | A step-by-step guide to solving SFPs | 175 |
| B.1.3 | Proof of Proposition 7 | 178 |
| B.2 | Chapter 7: Combinatorial constraints | 181 |
| B.2.1 | Proof of Lemma 5 | 181 |
| B.2.2 | Proof of Lemma 7 | 182 |
| B.2.3 | Proof of Theorem 8 | 184 |
| B.3 | Proof of Theorem 9 | 185 |
| B.3.1 | Proof of Theorem 10 | 186 |
| B.3.2 | Proof of Theorem 11 | 187 |
| B.3.3 | Proof of Theorem 12 | 188 |
| B.3.4 | Proof of Proposition 14 | 189 |
| B.3.5 | Proof of Proposition 15 | 190 |
| B.3.6 | Proof of Proposition 16 | 192 |
| B.3.7 | Proof of Proposition 17 | 194 |
| B.3.8 | Proof of Theorem 15 | 196 |
| B.3.9 | Proof of Theorem 16 | 199 |
| B.4 | Chapter 8: Risk constraints | 200 |
| B.4.1 | Proof of Lemma 6 | 200 |
| B.4.2 | Proof of Theorem 18 | 203 |

List of Tables

| | | |
|-----|--|-----|
| 5.1 | Preprocessing of the Adult dataset | 40 |
| 5.2 | Preprocessing of the COMPAS dataset | 42 |
| 5.3 | Classifier insensitivity on the COMPAS dataset | 43 |
| 8.1 | Classification of statistical uncertainty. | 133 |

List of Figures

| | | |
|-----|---|----|
| 5.1 | Classifier sensitivity on the Adult test set. | 41 |
| 5.2 | Dual variable analysis for Adult dataset: (a) distribution of the dual variables values and (b) prevalence of different groups among the 20% training set examples with largest dual variables. | 42 |
| 5.3 | Dual variables of different counterfactual constraints for the COMPAS dataset. . . . | 44 |
| 5.4 | Dual variable relative to the fairness constraint. | 46 |
| 5.5 | Dual variable relative to the fairness constraint. | 46 |
| 5.6 | Robust constrained learning (FMNIST): (a) Accuracy of classifiers under the PGD attack for different perturbation magnitudes and (b) distribution of ε used during training. | 49 |
| 5.7 | Robust constrained learning (CIFAR-10): (a) Accuracy of classifiers under the PGD attack for different perturbation magnitudes and (b) distribution of ε used during training. | 49 |
| 6.1 | Illustration of Example 1. | 64 |
| 6.2 | Reconstruction MSE for line spectral estimation of linear sources. | 70 |
| 6.3 | Support size estimation for line spectral estimation of linear sources. | 71 |
| 6.4 | Solutions obtained for line spectral estimation of saturated sources. | 72 |
| 6.5 | Reconstruction MSE for line spectral estimation of saturated sources. | 74 |
| 6.6 | Solution of functional logistic regression for ECG classification. | 75 |
| 6.7 | Receiver operating characteristic (ROC) curve for logistic classifiers in the presence of impulsive noise. | 77 |

| | | |
|------|--|-----|
| 6.8 | ECG and sparse functional coefficients (positive coefficients: blue, negative coefficients: red): (a) healthy heart and (b) heart with myocardial infarction. | 78 |
| 7.1 | Illustration of the near-optimal guarantee from Theorem 8. | 90 |
| 7.2 | Comparison between the bound in (7.48) and α | 104 |
| 7.3 | Relative suboptimality of sampling schemes for low SNR (SNR = -20 dB) | 106 |
| 7.4 | Relative suboptimality of sampling schemes for high SNR (SNR = 20 dB) | 107 |
| 7.5 | Relative suboptimality of MSE and log det (SNR = 20 dB) | 108 |
| 7.6 | Sampling set size for 90% reduction of MSE | 109 |
| 7.7 | Classification performance: (a) PCA and kPCA and (b) greedy subsampled kPCA. . | 109 |
| 7.8 | Classification performance of subsampled kPCA for different sampling schemes: (a) $k = 12$ components and (b) $k = 15$ components. Mean (solid line), median (\times), and error bars (one standard deviation) based on 100 sampling realizations. | 110 |
| 7.9 | Illustration of time-specific and overall input sets and schedules. | 113 |
| 7.10 | Bound on α from (7.59) for a schedule of length $N = 4$: (a) deterministic dynamical system ($\mathbf{W}_k = \mathbf{0}$, LQR) and (b) dynamical system with disturbance (LQG). Shaded regions span two standard deviations from the mean. | 118 |
| 7.11 | Relative suboptimality of greedy scheduling for different constraints (100 system realizations). (a) \mathcal{I}_1 (less than 2 actuators per time step); (b) \mathcal{I}_1 and \mathcal{I}_2 (less than 5 control actions over the horizon); (c) \mathcal{I}_1 , \mathcal{I}_2 , and \mathcal{I}_3 (inputs cannot be used on consecutive time slots). | 119 |
| 7.12 | Amazon basin, amazon river (dark blue trace), system states (light grey circles), and chemical spill origins (red circles). | 120 |
| 7.13 | Greedy schedule and actuation energy of the spill control agents. | 122 |
| 7.14 | Agents actions and chemical concentrations in the ocean (light blue mass in Figure 7.12) for the autonomous system (no agent), greedy schedule, and full actuation. | 123 |
| 7.15 | A-optimal design: (a) Thm. 15; (b) A-optimality (low SNR); (c) A-optimality (high SNR). The plots show the unnormalized A-optimality value for clarity. | 128 |
| 7.16 | E-optimal design: (a) Thm. 16; (b) E-optimality (low SNR); (c) E-optimality (high SNR). The plots show the unnormalized E-optimality value for clarity. | 128 |

| | | |
|-----|---|-----|
| 8.1 | Comparison between risk-neutral and risk-aware estimates. | 132 |
| 8.2 | Mean squared error and risk for different values of μ in the state-dependent noise scenario. | 140 |
| 8.3 | Mean squared error and risk for different values of μ in the communication scenario. | 141 |
| 8.4 | Risk-aware, MMSE, and robust estimates of z and h in the communication scenario for different values of Y | 142 |

Chapter 1

Introduction

1.1 Motivation

The information processing and autonomous systems underlying modern technologies such as self-driving cars and “smart” applications (e.g., grid, city, home, car...) have reached scales and complexity levels that can no longer be handled by traditional, hand-designed algorithms. Simultaneously, the access to data has become pervasive due to the lower cost of sensors and the expansion of the internet-of-things (IoT) and other interconnected devices. These trends have made statistical learning and inference a core component of the modern systems upon which we increasingly rely to select job candidates, analyze medical data, and manage assets. As its societal impact grows, the shortcomings of this approach become clear from the growing number of catastrophic failure reports involving biased, prejudiced models, systems prone to tampering and unsafe behaviors [1–6]. The need to control the behavior of data-based solutions has become critical to address these issues.

However, the models and structures found in state-of-the-art solutions are often too complex and opaque for their behaviors to be analyzed or predicted. Take the case of (convolutional, graph) neural networks (NNs). While they are now the *de facto* models in a wide variety of applications, we lack a fundamental understanding of their inner workings, something that remains an active topic of research [7–10]. Hence, explainability, interpretability, and transparency combined with human judgment is no longer enough to produce systems that perform according to specifications. We need

statistical algorithms and models whose behaviors can be controlled, curtailed, and dictated in a systematic manner.

Typically, learning requirements are imposed by using domain expert knowledge to *tune* the training objective (see, e.g., [11, 12, 12, 13, 13–17]). This approach, known as *regularization*, is ubiquitous in practice even though it need not yield feasible solutions [18]. In fact, existing results from classical learning theory guarantee generalization only with respect to the value of the regularized objective, which says nothing about satisfying the requirements it describes [19, 20]. Hence, even if a solution is feasible for the empirical problem, its statistical performance remains unclear. This issue is sometimes addressed by constructing models that explicitly *embed* the required properties (e.g., [21–27]), although the complexity of modern machine learning (ML) parametrizations renders this approach impractical.

1.2 Objectives

The goal of this work is to provide systematic, guaranteed way to impose requirements on the solution of statistical problems. In particular, it sets out to both develop the *theoretical underpinnings of constrained learning* and provide tools to *(approximately) solve inference problems subject to intricate requirements*, such as sparsity, risk, and combinatorial constraints. The objective is to establish constrained statistical optimization as a formal solution to impose requirements in both learning and inference settings. More specifically, this work

- formalizes the problems of learning and inference under requirements using the language of constrained statistical optimization;
- determines if, when, and to what extent these problems can be solved; and
- provides practical methods and algorithms to obtain these solutions.

1.3 Contributions

The main contributions of this thesis are listed below.

- 1) Formal theory of constrained learning (recipient of the *best student paper award* at ICASSP 2020)
 - (a) formalize the concept of a probably approximately correct constrained (PACC) learnability, the constrained learning counterpart of PAC learning;
 - (b) show that, under mild conditions, constrained learning is as hard as unconstrained learning in the sense that a hypothesis class is PACC learnable if and only if it is PAC learnable;
 - (c) prove that, under mild conditions, if unconstrained learning tasks can be performed, then so can constrained learning tasks. In other words, it is possible to learn under requirements by solving only unconstrained learning problems.
- 2) Definition and solution of sparse functional programs (SFPs)
 - (a) define a new class of nonlinear, non-convex, infinite dimensional optimization programs involving sparse objectives: the SFP;
 - (b) prove that despite their non-convexity, SFPs are strongly dual under mild conditions
 - (c) show that the value of SFPs (" L_0 -norm minimization") and their continuous compressive sensing counterparts (L_1 -norm minimization) are essentially the same, although not all solutions of the latter are sparse;
 - (d) put forward a practical, efficient algorithm to solve SFPs based on primal-dual dynamics.
- 3) Near-optimal guarantees for the greedy optimization of non-submodular functions
 - (a) define different forms of approximate submodularity, namely α and ϵ -submodularity;
 - (b) derive near-optimal guarantees for the greedy maximization of monotonically increasing, α and ϵ -submodular functions;
 - (c) derive the first explicit (P-computable) bounds on α and ϵ for quadratic costs such as the mean-square error (MSE), the worst-case error (maximum eigenvalue of the estimation error covariance), the average control cost, and the LQR/LQG costs.
- 4) Closed-form solution for risk-aware minimum MSE estimation problem (recipient of the *best paper award* at ICASSP 2020)

- (a) define a new risk constrained estimation problem using the predictive variance as a measure of risk;
- (b) show that, under mild conditions, this problem is equivalent to a convex program for which strong duality holds;
- (c) derive a closed-form expression for the risk-aware minimum MSE estimator, showing that it has the form of a biased estimator.

Due to time and space constraints, some of the work developed during these five years of Ph.D. could not be included in this thesis. Nevertheless, since they are closely related to topics it covers, I present them below along with their relevant references.

- *Model predictive selection* puts forward an online, receding horizon solution to the problem of actuator scheduling with stability guarantees [28];
- *Bayesian posterior learning* proposes a novel way to perform Bayesian inference in which Bayes rule is replaced by a statistical learning problem. An example in the context of Gaussian Processes (GPs) can be found in [29].
- *Learning for wireless resource allocation* uses constrained learning in order to obtain optimal resource (e.g., power) allocation policies without knowledge of channel distributions or capacity models [30];
- My work on *constrained reinforcement learning* essentially shows that a lot of the results that hold in the supervised learning scenario, also hold for reinforcement learning. It provides tools and algorithms to impose requirements on the policy learned through reinforcement learning, which has applications, e.g., in safety [31–33].
- *Resilient optimization* seeks to develop a theory of resilience akin to that of robustness using the language of constrained optimization. Here, resilience refers to the ability of a system to adapt to disturbances so extreme that it cannot continue to function normally. Overcoming such externalities requires a system to modify its behavior and possibly even change its nominal equilibrium. Applications in control can be found in [34, 35].

These theoretical contributions have impact in different application spanning different domains, a non-exhaustive list of which is presented below.

- **signal processing:** nonlinear line spectrum estimation, risk-aware estimation, graph signal sampling, wireless resource allocation;
- **control:** sensor selection, actuator/agent scheduling, safe policy learning, resilient control;
- **statistics/machine learning:** sparse functional data analysis, experimental design, constrained reinforcement learning, Bayesian inference, fairness, robustness, prediction credibility.

1.4 List of publication

The list of manuscripts below includes the work related to this thesis published during the course of my Ph.D. While most of these publications are featured in this thesis, some could not be included either because they have appeared (or will appear) as part of other theses or because they are part of research thrusts that were not mature enough at the time of this writing.

1.4.1 Preprints

1. L. Chamon and A. Ribeiro, “Probably approximately correct constrained learning,” *Advances in Neural Information Processing Systems (under review)*, 2020, <https://arxiv.org/abs/2006.05487>
2. L. Chamon, A. Amice, and A. Ribeiro, “Approximately supermodular scheduling subject to matroid constraints,” *IEEE Trans. on Autom. Control (under review)*, 2020, <https://arxiv.org/abs/2003.08841>
3. S. Paternain, M. Calvo-Fullana, L. Chamon, and A. Ribeiro, “Safe policies for reinforcement learning via primal-dual methods,” *IEEE Trans. on Autom. Control (under review)*, 2019, <https://arxiv.org/abs/1911.09101>

1.4.2 Journals

1. L. Chamon, Y. Eldar, and A. Ribeiro, “Functional nonlinear sparse models,” *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2449–2463, 2020
2. M. Peifer, L. Chamon, S. Paternain, and A. Ribeiro, “Sparse multiresolution representations with adaptive kernels,” *IEEE Trans. on Signal Process.*, vol. 68[1], pp. 2031–2044, 2020
3. L. Chamon, G. J. Pappas, and A. Ribeiro, “Approximate supermodularity of Kalman filter sensor selection,” *IEEE Trans. on Autom. Control.*, 2020
4. M. Eisen, C. Zhang, L. F. Chamon, D. D. Lee, and A. Ribeiro, “Learning optimal resource allocations in wireless systems,” *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2775–2790, 2019 [**Top 50 most accessed articles in IEEE TSP: May, July, Sept, Oct 2019**]
5. L. F. O. Chamon and A. Ribeiro, “Greedy sampling of graph signals,” *IEEE Trans. Signal Process.*, vol. 66[1], pp. 34–47, 2018

1.4.3 ML conferences

1. S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, “Constrained reinforcement learning has zero duality gap,” in *Advances in Neural Information Processing Systems*, 2019, pp. 7553–7563
2. L. Chamon and A. Ribeiro, “Approximate supermodularity bounds for experimental design,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5403–5412

1.4.4 Control conferences

1. L. Chamon, A. Amice, S. Paternain, and A. Ribeiro, “Resilient control: Compromising to adapt,” in *IEEE Control and Decision Conference (CDC)*, 2020, <https://arxiv.org/abs/2004.03726>

2. A. Tsiamis, D. Kalogierias, L. Chamon, A. Ribeiro, and G. Pappas, “Risk-constrained linear-quadratic regulators,” in *IEEE Control and Decision Conference (CDC)*, 2020, <https://arxiv.org/abs/2004.04685>
3. L. Chamon, S. Paternain, and A. Ribeiro, “Counterfactual programming for optimal control,” in *Learning for Dynamics & Control (L4DC)*, 2020
4. L. Chamon, A. Amice, and A. Ribeiro, “Matroid-constrained approximately supermodular optimization for near-optimal actuator scheduling,” in *IEEE Control and Decision Conference (CDC)*, 2019, pp. 3391–3398
5. S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, “Learning safe policies via primal–dual methods,” in *IEEE Conference on Decision and Control*, 2019
6. V. Silva, L. Chamon, and A. Ribeiro, “Model predictive selection: A receding horizon scheme for actuator selection,” in *American Control Conference*, 2019, pp. 347–353
7. L. Chamon, G. Pappas, and A. Ribeiro, “The mean square error in Kalman filtering sensor selection is approximately supermodular,” in *Conf. on Decision and Contr.*, 2017, pp. 343–350

1.4.5 Signal processing conferences

1. L. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro, “The empirical duality gap of constrained statistical learning problems,” in *International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2613–2616 [**Best student paper award**]
2. D. Kalogierias, L. Chamon, G. J. Pappas, and A. Ribeiro, “Better safe than sorry: Risk-aware nonlinear Bayesian estimation,” in *IEEE International Conference in Acoustic, Speech, and Signal Processing (ICASSP)*, 2020 [**Best paper award**]
3. L. Chamon, S. Paternain, and A. Ribeiro, “Learning gaussian processes with bayesian posterior optimization,” in *Asilomar*, 2019
4. L. Chamon, Y. C. Eldar, and A. Ribeiro, “Sparse recovery over nonlinear dictionaries,” in *ICASSP*, 2019

5. M. Peifer, L. Chamon, S. Paternain, and A. Ribeiro, “Sparse learning of parsimonious reproducing kernel Hilbert space models,” in *IEEE International Conference in Acoustic, Speech, and Signal Processing (ICASSP)*, 2019, pp. 3292–3296
6. M. Eisen, C. Zhang, L. Chamon, D. D. Lee, and A. Ribeiro, “Dual domain learning of optimal resource allocations in wireless systems,” in *IEEE International Conference in Acoustic, Speech, and Signal Processing (ICASSP)*, 2019, pp. 4729–4733
7. M. Peifer, L. Chamon, S. Paternain, and A. Ribeiro, “Locally adaptive kernel estimation using sparse functional programming,” in *Asilomar Conference on Signals, Systems and Computers*, 2018, pp. 2022–2026
8. M. Eisen, C. Zhang, L. Chamon, D. D. Lee, and A. Ribeiro, “Online deep learning in wireless communication systems,” in *Asilomar Conference on Signals, Systems and Computers*, 2018, pp. 1289–1293
9. L. Chamon, Y. C. Eldar, and A. Ribeiro, “Strong duality of sparse functional optimization,” in *ICASSP*, 2018, pp. 4739–4743
10. L. Chamon and A. Ribeiro, “Universal bounds for the sampling of graph signals,” in *Int. Conf. on Acoust., Speech and Signal Process.*, 2016
11. —, “Near-optimality of greedy set selection in the sampling of graph signals,” in *Global Conf. on Signal and Inform. Process.*, 2016, pp. 1265–1269

1.5 Notation

Throughout this manuscript, we use lowercase boldface letters for vectors (\mathbf{x}), uppercase boldface letters for matrices (\mathbf{X}), calligraphic letters for sets (\mathcal{A}), and *fraktur* font for measures (\mathfrak{h}). In particular, we denote the Lebesgue measure by \mathfrak{m} . We denote the set of complex numbers by \mathbb{C} , real numbers by \mathbb{R} , non-negative real numbers by \mathbb{R}_+ , and positive semi-definite (PSD) matrices by \mathbb{S}_+ . We also use $\mathbf{X} \succeq 0$ to mean that the matrix \mathbf{X} is PSD, so that for $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{n \times n}$, $\mathbf{X} \preceq \mathbf{Y} \Leftrightarrow \mathbf{b}^H \mathbf{X} \mathbf{b} \leq \mathbf{b}^H \mathbf{Y} \mathbf{b}$, for all $\mathbf{b} \in \mathbb{C}^n$. Similarly, we write $\mathbf{X} \succ 0$ to say that \mathbf{X} is positive definite.

We indicate the i -th entry of a vector by $[\mathbf{x}]_i$, i.e., if $\mathbf{x} \in \mathbb{C}^n$ and $\mathbf{x} = [x_1 \ \cdots \ x_n]^T$ then $[\mathbf{x}]_i = x_i$. Set subscripts refer either to the vector obtained by keeping only the elements with indices in the set $([\mathbf{x}]_{\mathcal{A}})$ or to the submatrix whose columns have indices in the set $([\mathbf{X}]_{\mathcal{A}})$. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we write $\mathbf{x} \succeq \mathbf{y}$ to denote that $[\mathbf{x}]_i \geq [\mathbf{y}]_i$ for all $i = 1, \dots, n$ and $\mathbf{x} \succ \mathbf{y}$ when $[\mathbf{x}]_i > [\mathbf{y}]_i$ for all i . We use $|\mathcal{A}|$ for the cardinality of \mathcal{A} , denote the empty set by \emptyset , and write $2^{\mathcal{A}}$ for the power set of \mathcal{A} (i.e., the collection of all finite subsets) and $\mathcal{P}(\mathcal{A})$ for the free monoid of \mathcal{A} (i.e., the collection of all finite multisets with elements in \mathcal{A}). Recall that a multiset is a set with repetition.

For a complex number $z = a + jb$, $j = \sqrt{-1}$, we write $\operatorname{Re}[z] = a$ for its real part and $\operatorname{Im}[z] = b$ for its imaginary part. We use \mathbf{z}^H for the conjugate transpose of the complex vector \mathbf{z} . We denote the support of a function $f : \Omega \rightarrow \mathbb{C}$ as $\operatorname{supp}(f) = \{\beta \in \Omega \mid f(\beta) \neq 0\}$ and define the indicator function $\mathbb{I} : \Omega \rightarrow \{0, 1\}$ as $\mathbb{I}(\beta \in \mathcal{E}) = 1$, if β belongs to the event \mathcal{E} , and zero otherwise.

We use L_p to denote the space of measurable functions whose p -th power is integrable. Of particular interest is the L_1 space of absolutely integrable functions and the L_2 space of square integrable functions. The measurable space over which these Hilbert spaces are defined can be inferred from the context. We additionally define the space $L_{1,+} = \{f \in L_1 \mid f \geq 0 \text{ a.e.}\}$, i.e., the space of a.e.-positive absolutely integrable functions. We write “a.e.” to mean “almost everywhere.” Unless explicitly stated, e.g., “p-a.e.,” we mean the Lebesgue measure.

Finally, we will often find it convenient to write the gradient as an explicit derivative (as, e.g., in [56]). In this case, we consider the derivative of a function f with respect to an $n \times 1$ vector \mathbf{x} to be its $1 \times n$ gradient vector, i.e.,

$$\frac{\partial f}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}.$$

Chapter 2

Learning and inference under requirements

Data-driven solutions have become a core component of the modern information systems we increasingly rely upon to select job candidates, analyze medical data, and control “smart” applications (home, grid, city). As these systems become ubiquitous, the number of reports involving biased, prejudiced models or systems prone to tampering (e.g., adversarial examples) and unsafe behaviors grows [1–6], evoking the need to impose requirements on their behavior. By requirements, we mean properties, either statistical or structural, that these solution must satisfy. These properties include fairness [57–62], robustness [15, 63, 64], safety [33, 65, 66], smoothness [67, 68], and sparsity [69].

2.1 Solving statistical problems with requirements today

Currently, the problem of learning and inference under requirements is addressed in two fashions. The first involves leveraging domain expert knowledge to build models that embed the desired behavior. This approach is often deployed to deal with geometric properties, such as invariance [21–27], or functional shape constraints, e.g., monotonicity. The effectiveness of this approach comes from the fact that the solution satisfies the desired properties by design, so no additional modification of the learning procedure is required. For instance, much of the success of convolutional neural networks (CNNs) in image processing is now attributed to their equivariance to translations and robustness to small image deformations [70, 71]. Other geometric properties, however, are harder to encode in the model structure, such as rotation invariance or invariances in non-Euclidian domains.

More challenging yet is to find ways of structurally embedding statistical properties such as fairness, safety, or risk-awareness. Even if possible, there may be multiple ways to encode a behavior and no definite path to choosing among them. What is more, these models are often designed for a specific property of a specific data type and are therefore difficult to transfer and often impossible to combine.

The second approach modifies the learning objective instead of the model. By leveraging the fact that learning is typically expressed as an optimization problem, it modifies the cost function of that problem to promote the desired behaviors. This often takes the form of a weighted combination between the original cost function, often some measure of statistical performance or fit, and regularizers. The role of these regularizers is to penalize violations of the requirements and/or reward desired behaviors [11–15, 17]. This approach does not require major modifications of the learning procedure, since these multiobjective problems can often be tackled using the same algorithms as their unregularized counterparts. What is more, the objective can include as many regularizers as necessary to simultaneously cope with a plurality of behaviors, overcoming a serious drawback of the previous approach.

Selecting these regularizers and their weight, however, can be challenging, especially as the number of constraints grows. In fact, their values often depend on the problem instance, the objective value, and can interact in non-trivial ways [72–76]. In the case of convex optimization problems, a straightforward relation between constraints and regularization costs can be obtained due to strong duality. A myriad of primal-dual methods can then be used to obtain optimal, feasible solutions [77]. However, most modern parametrizations (e.g., CNNs) lead to non-convex programs for which a regularized formulation need not yield feasible solutions, all the more so good ones [18]. While primal-dual algorithms have been used in practice, no guarantees can be given for their outcome in general [63, 64, 78, 79]. Regularization parameter tuning must therefore be performed by experts that understand not only the application, but also the underlying optimization algorithm. Though this trial-and-error process can be partially automated, it remains a brute force approach whose complexity scales exponentially with the number of regularizers.

2.2 Constrained learning and inference

Notice that two components of optimization problems have been targeted to address the issue of imposing requirements on data-driven solutions: the domain, i.e., the model structure, and the objective function. In contrast, this work posits that the proper way to impose requirements is actually to use their remaining component, namely, constraints. This approach has several advantages from a formulation perspective as requirements are often expressed as constraints in the first place. What is more, it comes with the same feasibility guarantee of the model design approach without limiting the number of requirements that can be imposed or having to deal with the choice or tuning of regularizers.

Formally, let \mathfrak{D}_i , $i = 0, \dots, m$ denote probability distributions over data pairs (\mathbf{x}, y) , with $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$, and let \mathcal{H} a hypothesis class containing functions $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$. For convenience, we can interpret \mathbf{x} as a feature vector or a system input, y as a label or a measurement, and ϕ as a classifier or estimator. For classification problems, \mathcal{Y} is a finite set, typically a subset of \mathbb{N} , whereas it is uncountable in the case of regression. We define the constrained statistical optimization (CSO) problem as

$$\begin{aligned} P^\star &= \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y)] \\ \text{subject to} \quad &\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] \leq c_i, \quad i = 1, \dots, m, \\ &\ell_j(\phi(\mathbf{x}), y) \leq c_j \quad \mathfrak{D}_j\text{-a.e.}, \quad j = m+1, \dots, m+q, \end{aligned} \tag{P-CSO}$$

where $\ell_i : \mathbb{R}^k \times \mathcal{Y} \rightarrow [0, B]$, $i = 0, \dots, m+q$, are bounded functions that, together with the c_i , encode performance metrics and the desired statistical properties of the solution. In general, we think of \mathfrak{D}_0 as a nominal joint distribution and the additional \mathfrak{D}_i as different conditional distributions over which requirements are imposed either on average, through the losses ℓ_i , $i \leq m$, or pointwise, through the losses ℓ_j , $j > m$. Note that we do not assume that any of the ℓ_i are convex. Observe that the unconstrained version of (P-CSO), namely

$$P_U^\star = \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y)], \tag{PI}$$

is at the core of celebrated Bayesian estimators, such as Kalman filters, and virtually every modern learning algorithm [20, 56, 80].

Throughout this work, we will distinguish two settings under which we wish to solve (P-CSO), learning and inference, depending on whether the distributions \mathfrak{D} are known. In the sequel, we describe these settings and detail the specific problems we investigate.

2.2.1 The learning setting

Learning refers to statistical problems in which the distributions \mathfrak{D}_i , $i = 0, \dots, m + q$, are not known *a priori* and are accessible only through independently drawn samples $(\mathbf{x}_{n_i}, y_{n_i}) \sim \mathfrak{D}_i$, $n_i = 1, \dots, N_i$. In classical statistics, this is known as the *non-parametric* setting. In the case of the unconstrained problem, classical learning theory shows that there exist conditions under which (PI) can be (approximately) solved based only on samples and that when this is the case, a (approximate) solution can be obtained by replacing expectations by sample averages, i.e., by solving the empirical risk minimization (ERM) problem [19, 20, 81, 82]

$$\hat{P}_U^* = \min_{\phi \in \mathcal{H}} \frac{1}{N_0} \sum_{n=1}^{N_0} \ell_0(\phi(\mathbf{x}_n), y_n). \quad (\text{P-ERM})$$

The phenomenon by which a solution obtained from samples, i.e., (P-ERM), approximates the solution over the population statistics, i.e., (P-CSO), is known as *generalization*.

For the constrained problem (P-CSO), similar results exist only in specific cases, e.g., for coherence constraints or rate-constrained learning [57, 58, 62, 83], and often hold for randomized solutions, e.g., [57, 58, 60, 62]. In general, there is however no reason to expect that the empirical approximation used to obtain (P-ERM) is also valid in the context constrained learning. Explicitly, it remains an open question whether the empirical constrained risk minimization (ECRM) problem

$$\begin{aligned} \hat{P}^* &= \min_{\phi \in \mathcal{H}} \frac{1}{N_0} \sum_{n_0=1}^{N_0} \ell_0(\phi(\mathbf{x}_{n_0}), y_{n_0}) \\ \text{subject to } &\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(\phi(\mathbf{x}_{n_i}), y_{n_i}) \leq c_i, \quad i = 1, \dots, m \\ &\ell_j(\phi(\mathbf{x}_{n_j}), y_{n_j}) \leq c_j, \text{ for all } n_j, \quad j = m + 1, \dots, m + q \end{aligned} \quad (\text{P-ECRM})$$

yields a (approximate) solution of (P-CSO). In fact, it is not clear if it is even possible to obtain a solution of (P-CSO) in the learning setting. This issue is aggravated when the infinite dimensional hypothesis class \mathcal{H} is parametrized (e.g., using neural networks), leading to non-convex constrained optimization problems, which from an algorithmic standpoint are considerably harder than their unconstrained counterparts. Indeed, while gradient descent can sometimes be used to approximately minimize a loss function even if it is non-convex [7, 9, 84, 85], it does not guarantee feasibility. In fact, even obtaining a feasible solution of (P-CSO) can be challenging, all the more so a good one. These questions are addressed in Part I.

2.2.2 The inference setting

Inference problems, on the other hand, assume direct access to the distributions \mathfrak{D}_i in some analytical form. Though this setting may seem considerably simpler than the previous one, it is nevertheless useful and several classical problems in statistics, such as hypothesis testing and detection [REF], have been cast and solved using constrained inference since Neyman-Pearson [86]. What is more, the additional statistical information given in this scenario allows us to explore more challenging requirements involving sparsity, combinatorial constraints, and risk. These will be the topic of Part II.

Before we briefly introduce those constraints we explore in Part II, we make a short remark on notation. In an attempt to not alienate readers familiar with the Bayesian inference literature, we reverse the roles of \mathbf{x} and \mathbf{y} in this part and consider vector measurements, i.e., $\mathbf{y} \in \mathbb{C}^k$. Explicitly, we now consider the measurements \mathbf{y} as observations based on which we wish to estimate the hidden variables \mathbf{x} . In other words, instead of observing both \mathbf{x} and \mathbf{y} but not \mathfrak{D}_i (learning), we observe \mathbf{y} and \mathfrak{D} , but not \mathbf{x} . Since the two parts of this manuscript stand mostly on their own, this is the only place where both notations will appear next to each other. Hence, this should not generate too much confusion.

2.2.2.1 Sparsity constraints

The first type of constrained inference problems we study involves infinite dimensional, nonlinear parametrizations with sparsity requirements. These are often found in signal processing due to the

analog and nonlinear nature of the physical world from where the signals are collected. Indeed, there are many examples of inherently continuous (as opposed to “discrete”) applications, such as spectral estimation, image recovery, and source localization [87–90], as well as nonlinear ones, e.g., magnetic resonance fingerprinting, spectrum cartography, and manifold data sparse coding [91–93].

These challenges are often tackled by imposing structure on the signals, such as bandlimitedness, finite rate of innovation, bounded total variation, or lying in a reproducing kernel Hilbert space (RKHS), which then allows them to be processed using an appropriate finite set of samples [94–97]. Hence, the challenges of infinite dimensionality and nonlinearity can be overcome by means of sampling theorems and “linear-in-the-parameters” methods. Due to the limited number of measurements, however, these discretizations often lead to underdetermined problems. Sparsity priors then play an important role in achieving state-of-the-art results by leveraging the assumption that there exists a signal representation in terms of only a few atoms from an overparametrized dictionary [69, 97, 98].

Since fitting these models leads to non-convex (and possibly NP-hard [99]) problems, sparsity is typically replaced by a tractable relaxation based on an atomic norm (e.g., the l_1 -norm). For linear, incoherent dictionaries, these relaxed problems have been shown to retrieve the desired sparse solution [69, 98]. Discretized continuous problems, however, rarely meet these conditions, which are NP-hard to check (e.g., restricted isometry/eigenvalue properties [100–102]). What is more, due to grid mismatch issues, infinite dimensional sparse signals need not be sparse when discretized [103–105]. Only in specific instances, such as line spectrum estimation, there exist guarantees for relaxations that forgo discretization [96, 106–111].

Instead, we propose to forgo both discretization and relaxation and directly tackle the continuous problem

$$\begin{aligned}
& \underset{\Theta \in \mathcal{H}}{\text{minimize}} && \mathbb{E}_{\beta \in \mathfrak{D}_0} \left[\mathbb{I}[\Theta(\beta) \neq 0] \right] \\
& \text{subject to} && \frac{1}{N} \sum_{i=1}^N g(z_i) \leq c \\
& && z_i = \mathbb{E}_{\beta \in \mathfrak{D}_i} \left[F_i[\Theta(\beta), \beta] \right], \quad i = 1, \dots, N,
\end{aligned} \tag{PII}$$

where \mathcal{H} is once again the space of measurable functions, $g : \mathbb{C} \rightarrow \mathbb{R}$ is a convex performance measure, and $F_i : \mathbb{C} \times \Omega \rightarrow \mathbb{C}$ is a (possibly nonlinear) function representing, e.g., the dictionary.

Note that if \mathfrak{D}_0 is the Lebesgue measure \mathfrak{m} , the objective function of (PII) is the measure of the support of Θ . Then, the empirical average of g together with c determines the goodness-of-fit of the parameters Θ for the nonlinear functional model defined for the z_i . Observe that while the statistical performance constraint on g is written in empirical form, the theory developed in Part I shows that (PII) can also be used in a learning setting. Although SFPs combine the infinite dimensionality of functional programming with the non-convexity of sparsity and nonlinear models, we show that they are tractable under mild conditions.

2.2.2.2 Combinatorial constraints

The second type of requirement considered in Part II are combinatorial. They arise in applications where access to the full measurement \mathbf{y} is impractical or impossible. The issue is then to select the subset of measurements based on which to solve the statistical inference problem and the constraints limit which subsets can be selected. For instance, power or sensing constraints may limit the number of observations allowed (cardinality constraint) or disallow certain combinations of observations due to, e.g., duty cycle limitations (matroid constraints). These problems arise in applications such as sensor selection, experimental design, and scheduling.

Formally, let $\mathcal{V} = \{1, 2, \dots, k\}$ be the ground set of possible measurements, i.e., the set of indices of the elements of $\mathbf{y} \in \mathbb{C}^k$. Our goal is to solve the combinatorial problem

$$\begin{aligned} & \underset{\mathcal{S} \subseteq \mathcal{V}}{\text{minimize}} && J(\mathcal{S}) \\ & \text{subject to} && \mathcal{S} \in \mathcal{I}_j, \quad j = 1, \dots, p, \end{aligned} \tag{PIII}$$

where $\mathcal{I}_j \subseteq 2^{\mathcal{V}}$ are families of subsets of \mathcal{V} that enumerates admissible sets of measurements and $J(\mathcal{S})$ is the value of a statistical problem solve with $[\mathbf{y}]_{\mathcal{S}}$, e.g., the minimum MSE $J(\mathcal{S}) = \min_{\phi \in \mathcal{H}} \mathbb{E} [\|\mathbf{x} - \phi([\mathbf{y}]_{\mathcal{S}})\|^2]$. Notice that $J(\mathcal{S})$ is itself the value of a statistical optimization problem.

Notice that the unconstrained version of (PIII) is trivial: inference typically improves as the number of measurements increases, so the solution is simply $\mathcal{S} = \mathcal{V}$. The constraints in (PIII), however, make that problem NP-hard in general. Hence, we cannot expect to find its solution even for problems of moderate size. Nevertheless, when the objective function J displays certain

diminishing return structures (e.g., supermodularity), it is well-known that greedy algorithms can be used to obtain approximate, near-optimal solutions of (PIII). Yet, many figures of merit of interest, chief among them the MSE, do not have this property.

2.2.2.3 Risk constraints

Finally, we consider requirements related to risk. Mitigating risk is crucial in critical applications in which decisions must be made not only on the basis of minimizing average losses, but also safeguarding against less probable, though possibly catastrophic, events. Examples can be found in areas such as wireless industrial control, energy [112, 113], finance [114–116], robotics [117, 118], LIDAR [119], and networking [120]. A typical approach is to use regularized objectives that involve not only the mean performance, but also some measure of statistical volatility of the losses, such as mean-variance functionals [114, 121], mean-semideviations [122], and Conditional Value-at-Risk (CVaR) [123]. In this work, however, we obtain risk-averse estimators by solving a constrained inference problem. Of particular interest is the risk-averse minimum mean square error (MSE) problem, in which a squared loss is minimized on average subject to a bound on its expected conditional variance. Explicitly,

$$\begin{aligned} & \underset{\phi \in \mathcal{H}}{\text{minimize}} && \mathbb{E}_{\mathfrak{D}} \left[\|\mathbf{x} - \phi(\mathbf{y})\|^2 \right] \\ & \text{subject to} && \mathbb{E}_{\mathfrak{D}} \left[\text{var} \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \right) \right] \leq \varepsilon \end{aligned} \tag{P-RISK}$$

where \mathcal{H} denotes the space of \mathbf{y} measurable functions,

$$\text{var} \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \right) \triangleq \mathbb{E} \left[\left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 - \mathbb{E} \left[\|\mathbf{x} - \phi(\mathbf{y})\|^2 \right] \right)^2 \mid \mathbf{y} \right] \tag{2.1}$$

is the predictive variance of the squared loss $\|\mathbf{x} - \phi(\mathbf{y})\|^2$ with respect to \mathbf{y} , and $\varepsilon > 0$ is a fixed risk tolerance. While (P-RISK) has many desirable properties (see Chapter 8 for details), it is a non-convex, functional optimization problem. Hence, even with knowledge of the statistics \mathfrak{D} , solving (P-RISK) directly is quite challenging.

In summary, whereas learning and inference problems involving requirements are straightforward to pose as constrained statistical optimization problems, solving them is certainly not. Among the main challenges are unknown statistics, infinite dimensionality, non-convexity, and combinatorial

nature. In the remainder of this manuscript, we overcome these challenges using a diverse set of tools spanning learning theory, constrained optimization, duality, vector measure theory, discrete algebra, and spectral theory.

Part I

Constrained learning

Chapter 3

Constrained learning theory

3.1 Classical learning theory

Classical learning theory is concerned with studying unconstrained learning problems of the form (PI) to identify what it means to solve a statistical learning problem, when it can be done, and how [20]. While there exist today a myriad of learning models, the first and still quite popular one is known as (agnostic) *probably approximately correct* (PAC) learning [19, 20, 81, 82]. Agnostic is used to differentiate the statistical data model introduced in Chapter 2, i.e., the joint distribution \mathfrak{D}_0 , with the original PAC learning model in which a deterministic labeling function h is used to generate y . We do not dwell on this distinction, but bear in mind that whenever we say “learning theory” in this work, we actually mean agnostic PAC learning theory.

The PAC framework describes the best we can expect to achieve in view of the fact that we only have access to the distributions \mathfrak{D}_i through random samples. In particular, we cannot expect to obtain an exact solution of (PI) since we can only obtain coarse estimates of its objective function. This fact is encapsulated in the *Approximately* part of PAC. What is more, since our estimate is based on random sample, we must account for the possibility of sometimes obtaining a completely uninformative sample set, in which case we will completely fail to solve the original problem. Hence, we can only expect to be Approximately correct in probability, thus the *Probably* in PAC. When such an approximate solution of (PI) can be obtained in \mathcal{H} with high probability, the hypothesis

class is said to be *PAC learnable* as formalized in the definition below [19, 20, 81, 82]. Notice that PAC learnability is a property of the hypothesis class, i.e., the domain \mathcal{H} of (PI), since it must hold for all distributions \mathfrak{D}_0 .

Definition 1 (PAC learnability). A hypothesis class \mathcal{H} is (*agnostic*) *probably approximately correct* (PAC) learnable if for every $\epsilon, \delta \in (0, 1)$ and every distribution \mathfrak{D}_0 , a $\phi^\dagger \in \mathcal{H}$ can be obtained from $N \geq N_{\mathcal{H}}(\epsilon, \delta)$ samples of \mathfrak{D}_0 such that

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi^\dagger(\mathbf{x}), y)] \leq P_U^* + \epsilon, \quad \text{with probability } 1 - \delta. \quad (3.1)$$

A classical result from learning theory states that \mathcal{H} is PAC learnable if and only if it has finite VC (Vapnik–Chervonenkis) dimension and that the ϕ^\dagger from Def. 1 can be obtained by solving the ERM problem P-ERM (possibly with the addition of a regularizer on ϕ) [19, 20]. The VC dimension is a measure of complexity or richness of a hypothesis class:

Definition 2 (VC dimension). Let $\mathcal{C} = \{c_1, \dots, c_m\} \subset \mathcal{X}$ be a finite set of points and define the set of all vectors in $\{-1, 1\}$ achievable by the hypothesis class \mathcal{H}

$$\mathcal{H}_{\mathcal{C}} = \{(\text{sign}(h(c_1)), \dots, \text{sign}(h(c_m))) \mid h \in \mathcal{H}\}.$$

The class \mathcal{H} is said to *shatter* \mathcal{C} if $|\mathcal{H}_{\mathcal{C}}| = 2^{|\mathcal{C}|}$. The VC dimension of \mathcal{H} is the cardinality of the largest set \mathcal{C} it can shatter. If \mathcal{H} can shatter sets of arbitrary size, we say \mathcal{H} has infinite VC dimension.

Notice, however, that the objects studied in (PAC) learning theory are unconstrained statistical learning problem. Definition 1 is therefore not enough to enable constrained learning since the a PAC ϕ^\dagger may not be feasible for (P-CSO). In fact, feasibility often takes priority over performance in constrained learning problems. For instance, regardless of how good a fair classifier is, it serves no “fair” purpose in practice unless it meets some fairness requirements. We therefore need a new framework to deal with constrained learning problems, which we introduce in the sequel.

3.2 Probably Approximately Correct Constrained Learning

As with unconstrained learning problems, we cannot expect to solve (P-CSO) exactly without access to the \mathfrak{D}_i against which to evaluate its expectations. Additionally, (P-CSO) is a variational problem in general, which can be challenging to solve unless \mathcal{H} is finite. Yet, obtaining a PAC solution (Definition 1) of (P-CSO) is not enough since it does not account for constraints or in a broader sense, for the problem requirements. The following definition formalizes what it means to learn under constraints by defining what is a *good enough* solution of (P-CSO).

Definition 3 (PACC learnability). A hypothesis class \mathcal{H} is *probably approximately correct constrained* (PACC, pronounced “PAC-cee”) learnable if for every $\epsilon, \delta \in (0, 1)$ and every distribution \mathfrak{D}_i , $i = 0, \dots, m + q$, a $\phi^\dagger \in \mathcal{H}$ can be obtained based $N \geq N_{\mathcal{H}}(\epsilon, \delta)$ samples from each \mathfrak{D}_i such that, with probability $1 - \delta$, it is

- 1) probably approximately optimal, i.e.,

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi^\dagger(\mathbf{x}), y)] \leq P^* + \epsilon, \quad \text{and} \quad (3.2)$$

- 2) probably approximately feasible, i.e.,

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi^\dagger(\mathbf{x}), y)] \leq c_i + \epsilon \quad \text{and} \quad (3.3a)$$

$$\ell_j(\phi^\dagger(\mathbf{x}), y) \leq c_j, \quad \text{for all } (\mathbf{x}, y) \in \mathcal{K}_j, \quad (3.3b)$$

where the $\mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}$ are sets of \mathfrak{D}_j measure at least $1 - \epsilon$.

Definition 3 is an extension of PAC learnability (Definition 1) to the problem of learning under requirements. Indeed, observe that (3.2) is equivalent to (3.1). Hence, we immediately obtain that every PACC learnable class is also PAC learnable. However, a PACC learnable class must also meet the probably approximate feasibility conditions in (3.3). In the case of average constraints, namely (3.3a), the PACC solution ϕ^\dagger can violate the specification c_i by at most ϵ . In other words, it cannot be too far from the original requirement. In the case of pointwise constraints, namely (3.3b), ϕ^\dagger must satisfy the constraint with \mathfrak{D}_j -probability at least $1 - \epsilon$. However, there is no requirement

on the magnitude of the violation. The additional “C” in PACC serves to remind ourselves of this distinction, although as we show next, PAC and PACC learnability are in fact equivalent properties.

3.3 PACC Learning is as Hard as PAC Learning

Having formalized what we mean by constrained learning (Sec. 3.2), we turn to the issue of when it can be done. To do so, we follow the unconstrained learning lead and consider the ECRM rule in (P-ECRM) based on N_i samples $(\mathbf{x}_{n_i}, y_{n_i}) \sim \mathfrak{D}_i$. The following theorem shows that, under mild assumptions, if \mathcal{H} is PAC learnable, then it is also PACC learnable using (P-ECRM).

Theorem 1. *Let the ℓ_i , $i = 0, \dots, m + q$, be bounded on \mathcal{X} . The hypothesis class \mathcal{H} is PACC learnable if and only if it is PAC learnable and (P-ECRM) is a PACC learner of \mathcal{H} . Explicitly, let $d_{\mathcal{H}} < \infty$ be the VC dimension of \mathcal{H} . If $N_i \geq C\zeta^{-1}(\epsilon, \delta, d_{\mathcal{H}})$, $i = 0, \dots, m + q$, for an absolute constant C and*

$$\zeta^{-1}(\epsilon, \delta, d) = \frac{d + \log(1/\delta)}{\epsilon^2}, \quad (3.4)$$

then any solution $\hat{\phi}^$ of (P-ECRM) is a PACC solution of (P-CSO).*

Proof. See Appendix A.1. ■

Theorem 1 shows that, from a learning theoretic standpoint, constrained learning is as hard as unconstrained learning. Not only that, but notice the sample complexity of constrained described by (3.4) matches that of PAC learning [19, 20]. It is therefore not surprising that a constrained version of ERM is a PACC learner. A similar result appeared in [59] for a particular rate constraint and not in the context of PACC learning. Still, solving (P-ECRM) remains challenging. Indeed, while it addresses the statistical issue of (P-CSO), it remains, in most practical cases, an infinite dimensional (functional) problem. This issue is often addressed by leveraging a finite dimensional parametrization of (a subset of) the \mathcal{H} , such as a kernel model or a (C)NN. Explicitly, we associate

to each parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ a function $f_{\boldsymbol{\theta}} \in \mathcal{H}$, replacing (P-ECRM) by

$$\begin{aligned} \hat{P}_{\boldsymbol{\theta}}^* &= \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \quad \frac{1}{N_0} \sum_{n_0=1}^{N_0} \ell_0(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_0}), y_{n_0}) \\ \text{subject to} \quad &\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_i}), y_{n_i}) \leq c_i, \quad i = 1, \dots, m \\ &\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_j}), y_{n_j}) \leq c_j, \text{ for all } n_j, \quad j = m+1, \dots, m+q. \end{aligned} \tag{PIV}$$

Even if (P-ECRM) is a convex program in ϕ , (PIV) typically is not a convex program in $\boldsymbol{\theta}$, except in particular cases. While this issue also arises in unconstrained learning problems, it is exacerbated by the presence of constraints. Though it is sometimes possible to find good approximate minimizers of ℓ_0 using, e.g., gradient descent rules [7,9,84,85,124], even obtaining a feasible $\boldsymbol{\theta}$ may be challenging. Regularized formulations and relaxations are often used to sidestep this issue by incorporating a linear combination of the constraints into the objective and solving the resulting unconstrained problem [11–15,17]. Yet, whereas the generalization guarantees of classical learning theory apply to this modified objective, they say nothing of the requirements it describes. Since strong duality need not hold for the non-convex (PIV), this procedure need not be PACC (Definition 3) and may lead to solutions that are either infeasible or whose performance is unacceptably poor [18].

While no formal connection can be drawn between (PIV) and its regularized formulation, its dual problem turns out to be related to (P-CSO). In fact, this dual learning approach provides a practical algorithm that yields a (near-)PACC solutions for (P-CSO) based only on solving unconstrained learning problems. This result concludes definitely closes the gap between unconstrained and constrained learning.

Chapter 4

Dual constrained learning

In order to overcome the shortcomings of (P-ECRM) [and its parametrized version (PIV)], we put forward a different learning rule based on the (parametrized) empirical dual of (P-CSO). Our goal is to quantify the loss of optimality incurred by replacing the constrained, variational, statistical (P-CSO) by an unconstrained, finite dimensional, empirical problem. In doing so, we prove that it is a PACC learner except for an approximation error determined by the quality of the parametrization. We formalize this concept by introducing the parametrized hypothesis class $\mathcal{P} = \{f_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathbb{R}^p\}$ and defining the concept of near-PACC solution:

Definition 4 (Near-PACC learnability). A hypothesis class \mathcal{H} is *nearly probably approximately correct constrained* (near-PACC) learnable using a parametrized class \mathcal{P} if there exists $\epsilon_0 \geq 0$ such that for every $\epsilon, \delta \in (0, 1)$ and every distribution \mathfrak{D}_i , $i = 0, \dots, m + q$, a $\phi^\dagger \in \mathcal{H}$ can be obtained from $N \geq N_{\mathcal{P}}(\epsilon, \delta)$ samples from each \mathfrak{D}_i that is, with probability at least $1 - \delta$,

- 1) probably near-approximately optimal, i.e.,

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi_N(\mathbf{x}), y)] \leq P^* + \epsilon_0 + \epsilon, \quad \text{and} \quad (4.1)$$

2) probably approximately feasible, i.e.,

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi^\dagger(\mathbf{x}), y)] \leq c_i + \epsilon \quad \text{and} \quad (4.2a)$$

$$\ell_j(\phi^\dagger(\mathbf{x}), y) \leq c_j, \quad \text{for all } (\mathbf{x}, y) \in \mathcal{K}_j, \quad (4.2b)$$

where the $\mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}$ are sets of \mathfrak{D}_j measure at least $1 - \epsilon$.

The main difference between near-PACC (Definition 4) and PACC learnability (Definition 3) is the presence of a minimum gap (ϵ_0) between the optimal value P^* of (P-CSO) and the value of ϕ^\dagger . Indeed, for $\epsilon_0 = 0$, near-PACC reduces to PACC learnability. Note that ϵ_0 does not depend on the number of samples N or the distributions \mathfrak{D}_i . It is a property of the learning task and the parametrized class \mathcal{P} . In that sense, it is akin to what is known in classical learning known as the *approximation error*, as opposed to ϵ , the *estimation error* [20]. In contrast to unconstrained learning, however, the approximation error can not be treated separately due to the constraints. Yet, it is *fixed*, i.e., independent of the sample set, and affects neither the sample complexity nor the constraint satisfaction. Hence, the parametrized constrained learner sacrifices optimality, but not feasibility, which remains dependent only on the number of samples N [(4.2)]. Finally, note that the sample complexity $N_{\mathcal{P}}$ does not depend on the original hypothesis class \mathcal{H} , but on the parametrized \mathcal{P} . Near-PACC is therefore also related to representation-independent learning [20].

In the sequel, we define the (parametrized) empirical dual problem of (P-CSO) and proceed to show that it is a near-PACC learner (Theorems 2 and 3), by first analyzing the *duality gap* of (P-CSO) (Section 4.1.1), then bounding the error due to the finite dimensional parametrization (the *parametrization gap*, Section 4.1.2), and finally studying the effect of replacing the expectations by sample averages (the *empirical gap*, Section 4.1.3).

4.1 Empirical dual learning

Before defining the (parametrized) empirical dual of (P-CSO), let us derive its classical Lagrangian dual problem. Explicitly, we begin by defining the Lagrangian of (P-CSO) as

$$\begin{aligned} L(\phi, \boldsymbol{\mu}, \boldsymbol{\lambda}) = & \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y)] + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] - c_i \right] \\ & + \sum_{j=m+1}^{m+n} \int \lambda_j(\mathbf{x}, y) [\ell_j(\phi(\mathbf{x}), y) - c_j] p_{\mathfrak{D}_j}(\mathbf{x}, y) d\mathbf{x} dy, \end{aligned} \quad (4.3)$$

where $p_{\mathfrak{D}_j}$ is the density of the distribution \mathfrak{D}_j , $\boldsymbol{\mu} \in \mathbb{R}_+^m$ collects the dual variables μ_i relative to the average constraints, and $\boldsymbol{\lambda}$ is an $n \times 1$ vector that collects the functional dual variables $\lambda_j \in L_{1,+}$ relative to the pointwise constraints. Recall that by $f \in L_{1,+}$ we mean that $f \in L_1$ (absolutely integrable) and $f \geq 0$ a.e. For conciseness, we leave the measure (\mathfrak{D}_j) implicit. Observe that, since the losses ℓ_j are bounded (Assumption 4), the integral in (4.3) exists and is well-defined. This is a direct consequence of Hölder's inequality [125, Thm. 1.5.2]. Additionally, note that while \mathfrak{D}_j need not have a density, we assume that it is absolutely continuous with respect to the Lebesgue measure to simplify the exposition. The dual problem of (P-CSO) can then be written as

$$D^* = \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m, \lambda_j \in L_{1,+}} \min_{\phi \in \mathcal{H}} L(\phi, \boldsymbol{\mu}, \boldsymbol{\lambda}). \quad (\text{D-CSL})$$

The dual problem (D-CSL) is a composite functional optimization programs. It involves both a minimization with respect to ϕ and a maximization with respect to the dual variables λ_j , which are infinite dimensional due to the pointwise constraints. Hence, solving it (even approximately) remains quite challenging. What is more, we only have in general that D^* is a lower bound on the value of P^* (weak duality [18]), which means that even if we could solve (D-CSL), it may not get us closer to solving the constrained learning problem we are actually interested in, namely (P-CSO). Nevertheless, (D-CSL) is an unconstrained problem, which suggests that it could be tackled using the same techniques as unconstrained learning. The goal of this chapter is to show that doing so actually provide a near-PACC solution of (P-CSO).

Start by defining the (*parametrized*) *empirical Lagrangian* of (P-CSO) as

$$\begin{aligned} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j) = & \frac{1}{N_0} \sum_{n_0=1}^{N_0} \ell_0(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_0}), y_{n_0}) + \sum_{i=1}^m \mu_i \left[\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right] \\ & + \sum_{j=m+1}^{m+q} \left[\frac{1}{N_j} \sum_{n_j=1}^{N_j} [\lambda_j]_{n_j} (\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_j}), y_{n_j}) - c_j) \right], \end{aligned} \quad (4.4)$$

based on N_i samples $(\mathbf{x}_{n_i}, y_{n_i}) \sim \mathfrak{D}_i$, $i = 0, \dots, m+q$, where $\boldsymbol{\mu} \in \mathbb{R}_+^m$ collects the dual variables μ_i relative to the average constraints and $\boldsymbol{\lambda}_j \in \mathbb{R}_+^{N_j}$ collects the dual variables $[\lambda_j]_{n_j}$ relative to the j -th pointwise constraint. The *empirical dual problem* of (P-CSO) is then written as

$$\hat{D}^* = \max_{\substack{\boldsymbol{\mu} \in \mathbb{R}_+^m, \\ \boldsymbol{\lambda}_j \in \mathbb{R}_+^{N_j}}} \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j). \quad (\hat{\text{D-CSL}})$$

Note that the empirical Lagrangian (4.4) has the same form as the regularized problems often used to tackle learning under requirements. However, whereas $\boldsymbol{\mu}$ is kept fixed (or estimated based on trial and error and cross-validation) in the latter, it is an optimization variable in $(\hat{\text{D-CSL}})$. Also observe that $(\hat{\text{D-CSL}})$ is the dual problem of the parametrized ECRM (PIV). Yet, due to its non-convexity, only weak duality holds, so that in general a saddle-point of $(\hat{\text{D-CSL}})$ is in no way related to a solution of (PIV) [18]. In contrast, $(\hat{\text{D-CSL}})$ is a parametrized, empirical version of the dual problem (D-CSL) that can be related directly to the original learning problem (P-CSO).

The assumptions required to establish these results depend on whether the original learning task includes pointwise constraints, i.e., on q in (P-CSO). We introduce both results below for reference, but proceed to showcase only the proof for $q = 0$, i.e., for the case without pointwise constraints. We do so because this result considers non-convex losses ℓ_i , so its proof is more involved than the $q > 0$ case, in which the losses are required to be convex. Since the techniques used to establish both results are similar, we defer the proof of the latter to the appendices.

Let us start with the case with pointwise constraints, i.e., $q > 0$ in (P-CSO). We establish that $(\hat{\text{D-CSL}})$ is a near-PACC learner under the following assumptions:

Assumption 1. The losses $\ell_i(\cdot, y)$, $i = 0, \dots, m+q$, are M -Lipschitz, convex functions for all $y \in \mathcal{Y}$.

Assumption 2. The hypothesis class \mathcal{H} is a convex, closed functional space, the parametrized

class $\mathcal{P} \subseteq \mathcal{H}$ is PAC learnable, and there is $\nu > 0$ such that for each $\phi \in \mathcal{H}$ there exists $f_{\theta} \in \mathcal{P}$ for which

$$\sup_{\mathbf{x} \in \mathcal{X}} |f_{\theta}(\mathbf{x}) - \phi(\mathbf{x})| \leq \nu. \quad (4.5)$$

Assumption 3. There exists $\theta' \in \mathbb{R}^p$ such that $f_{\theta'}$ is strictly feasible for (P-CSO) with constraints $c_i - M\nu$ and $c_j - M\nu$ and for each datasets $\mathcal{S} = \{(\mathbf{x}_{n_i}, y_{n_i})\}_{i=0, \dots, m+q}$ there exists a θ'' that is strictly feasible for (PIV).

In this setting, we can obtain the following result:

Theorem 2. Define V_N as

$$V_N = B \sqrt{\frac{1}{N} \left[1 + \log \left(\frac{4(m+q+2)(2N)^{d_{\mathcal{P}}}}{\delta} \right) \right]}, \quad (4.6)$$

where $d_{\mathcal{P}}$ is the VC dimension of \mathcal{P} . Under Assumptions 1–3, it holds with probability at least $1 - \delta$ that

$$P^* \leq \hat{D}^* \leq P^* + \left(1 + \|\tilde{\mu}^*\|_1 + \sum_{j=m+1}^{m+q} \|\tilde{\lambda}_j^*\|_{L_1} \right) M\nu + V_{N_0} \quad \text{and} \quad (4.7a)$$

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\hat{\theta}^*}(\mathbf{x}), y)] \leq c_i + V_{N_i}, \quad \text{for all } i \quad (4.7b)$$

$$\ell_i(f_{\hat{\theta}^*}(\mathbf{x}), y) \leq c_i, \quad \text{for all } (\mathbf{x}, y) \in \mathcal{K}_j, \quad (4.7c)$$

where P^* is the value of (P-CSO), $(\tilde{\mu}^*, \tilde{\lambda}^*_j)$ are the optimal dual variables of (P-CSO) with the constraints tightened to $c_i - M\nu$, $i = 1, \dots, m+q$, and $\mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}$ is a set of \mathfrak{D}_j -measure at least $1 - V_{N_j}$ for all $j = m+1, \dots, m+q$.

Proof. See Appendix A.2. ■

From Theorem 2, we immediately obtain that $(\hat{D}\text{-CSL})$ is a near-PACC learner (Definition 4).

Corollary 1. Under Assumptions 1–3, $(\hat{D}\text{-CSL})$ is a near-PACC learner of \mathcal{H} through \mathcal{P} for

$$\epsilon_0 = \left(1 + \|\tilde{\mu}^*\|_1 + \sum_{j=m+1}^{m+q} \|\tilde{\lambda}_j^*\|_{L_1} \right) M\nu,$$

where $(\tilde{\boldsymbol{\mu}}^*, \tilde{\boldsymbol{\lambda}}^*_j)$ are the optimal dual variables of (P-CSO) with the constraints tightened to $c_i - M\nu$, $i = 1, \dots, m + q$.

Since the approximation guarantees are very similar, we defer commenting on the implications of these results until after we have stated the result for learning tasks without pointwise constraints, i.e., $q = 0$ in (P-CSO). In this case, we rely on the following milder assumptions:

Assumption 4. The losses $\ell_i(\cdot, y)$, $i = 0, \dots, m$, are M -Lipschitz continuous (possibly non-convex) functions for each $y \in \mathcal{Y}$.

Assumption 5. The hypothesis class \mathcal{H} is a convex, closed functional space, the parametrized class $\mathcal{P} \subseteq \mathcal{H}$ is PAC learnable and there is $\nu > 0$ such that for each $\phi \in \mathcal{H}$ there exists $f_{\boldsymbol{\theta}} \in \mathcal{P}$ for which

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} [|f_{\boldsymbol{\theta}}(\mathbf{x}) - \phi(\mathbf{x})|] \leq \nu, \quad \text{for } i = 0, \dots, m. \quad (4.8)$$

For clarity, we introduce and prove this result in the classification setting, i.e., the case in which \mathcal{Y} is finite. The regression problem is considered using an additional smoothness assumption on the losses to transform it into a classification problem as is done for regression trees [80]. These results are presented in Appendix A.4.

Theorem 3. Consider the learning task (P-CSO) without pointwise constraints, i.e., for $q = 0$. Let \mathcal{Y} be finite, the conditional random variables $\mathbf{x}|y$ induced by the \mathcal{D}_i be non-atomic, and \mathcal{H} be a decomposable function space (see Section 4.1.1). Under assumptions 3–5, it holds that

$$P^* \leq \hat{D}^* \leq P^* + (1 + \|\tilde{\boldsymbol{\mu}}^*\|_1) M\nu + V_{N_0} \quad \text{and} \quad (4.9a)$$

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} [\ell_i(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y)] \leq c_i + V_{N_i}, \quad \text{for all } i \quad (4.9b)$$

with probability $1 - \delta$, where P^* is the value of (P-CSO) with $q = 0$, V_N is as in (4.6), and $\tilde{\boldsymbol{\mu}}^*$ are the dual variables of (P-CSO) with the constraints tightened to $c_i - M\nu$, $i = 1, \dots, m$.

Theorem 3 immediately implies that $(\hat{D}\text{-CSL})$ is a near-PACC learner (Definition 3).

Corollary 2. Under Assumptions 3–5, $(\hat{D}\text{-CSL})$ is a near-PACC learner of \mathcal{H} through \mathcal{P} in (P-CSO) with $q = 0$ for

$$\epsilon_0 = (1 + \|\tilde{\boldsymbol{\mu}}^*\|_1) M\nu, \quad (4.10)$$

where $\tilde{\mu}^*$ are the dual variables of (P-CSO) with its constraints tightened to $c_i - M\nu$, $i = 1, \dots, m$.

Theorems 2 and 3 establish conditions under which $(\widehat{D}\text{-CSL})$ is a near-PACC learner of the constrained learning problem (P-CSO). When the learning task has pointwise requirements, Theorem 2 requires that the losses ℓ_i be convex (Assumption 1). This is in contrast to unconstrained learning or the ECRM result from Theorem 1 which hold regardless of the losses convexity. Nevertheless, this does not imply that either $(\widehat{D}\text{-CSL})$ or (PIV) are convex problems since $\ell_i(f_{\theta}(\mathbf{x}), y)$ need not be a convex function of θ . Theorem 3 shows this restriction can be lifted when there are no pointwise constraints under some mild conditions on the distributions and the hypothesis class (see Section 4.1.2 for more details).

Another distinction between Theorems 2 and 3 is with respect to the assumption on the richness of the parametrized class \mathcal{P} . A stronger, uniform approximation requirement (Assumption 2) is needed to meet the pointwise constraints. Assumptions 2 and 5 also require \mathcal{H} to be a convex class of functions. These conditions are met by a myriad of function space-parametrization pairs, e.g., when \mathcal{H} is the space of continuous functions or an RKHS, f_{θ} can be a neural network [126–128] or a finite linear combinations of kernels [68, 129] respectively. This requirement can be relaxed in the absence of pointwise constraints to a total variation covering (Assumption 5). Notice that \mathcal{H} can be any closed functional space, since only the parametrized class \mathcal{P} needs to be PAC learnable. Assumption 3 is used in both theorems to guarantee that the problem is well-posed, i.e., that there exist feasible solution for (P-CSO) in \mathcal{P} .

In either case, the quality of the approximation is dictated by (i) the sample size, (ii) the difficulty of the learning task, and (iii) the richness of the parametrization. Indeed, as the sample size increases, the estimation error V_N decreases at the classical rate of $O(\log(N)/\sqrt{N})$. This has a direct impact on both the near-optimality and approximate feasibility of the problem. In fact, note that when the \mathfrak{D}_i denote conditional distributions of \mathfrak{D}_0 , N_i can be considerably smaller than N_0 and jeopardize the ability to impose the constraints. This is particularly critical for classes that are minority in the dataset (see, e.g., the fairness examples in Chapter 5). On the other hand, factors (ii) and (iii) affect only the optimal value of the solution, as required by the PACC framework (Definition 4).

By learning task difficulty [(ii)], we mean the sensitivity of the statistical problem to the constraints. This is embodied by the well-known sensitivity interpretation of the dual variables [130,

Sec. 5.6], which can be formalized here due to the lack of duality gap (see Lemma 1 below or Proposition 18 in Appendix A.4). Thus, $(1 + \|\tilde{\boldsymbol{\mu}}^*\|_1)M$ in (4.9a) effectively quantifies how stringent the constraints are for the learning problem in terms of how much performance could be gained by relaxing them. A similar interpretation applies to (4.7a).

Finally, the approximation error is affected by the approximation capability ν of the parametrization [(iii)]. It is worth noting that better parametrizations typically involve more parameters, which in turn affects the VC dimension of \mathcal{P} , leading to a typical compromise between approximation error and complexity. For small sample sets, (4.9a) is dominated by the estimation error $\epsilon = V_N$. This motivates the use of lower complexity classes which, though more restricted, will generalize better. If there is abundance of data or the learning requirements are particularly stringent, the approximation error ϵ_0 in (4.10) dominates and finer, even if more complex, parametrizations should be used.

In the sequel, we describe the steps necessary to prove Theorem 3. The proof is a combination of three results that analyze the errors incurred while transforming (P-CSO) into $(\widehat{\mathbf{D}}\text{-CSL})$. First, we show that (P-CSO) is strongly dual under the conditions of Theorem 3. Hence, it is equivalent to its dual problem (D-CSL) and the *duality gap* is zero (Section 4.1.1). Then, we bound the *parametrization gap*, i.e., the error incurred from solving $(\widehat{\mathbf{D}}\text{-CSL})$ using the parametrized space \mathcal{P} instead of \mathcal{H} (Section 4.1.2). Finally, we study the *empirical gap* from approximating expectations by sample averages (Section 4.1.3). Since some of these results may be of independent interest, we briefly discuss each of them in the sequel. The complete proof of Theorem 2 is deferred to Appendix A.2.

4.1.1 The duality gap

Consider the dual problem of (P-CSO) in (D-CSL) for $q = 0$, in which case the Lagrangian in (4.3) reduces to

$$L(\phi, \boldsymbol{\mu}) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_0} [\ell_i(\phi(\mathbf{x}), y)] + \sum_{i=1}^m \mu_i \left(\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} [\ell_i(\phi(\mathbf{x}), y)] - c_i \right). \quad (4.11)$$

Since (4.11) is a relaxation of (P-CSO), we know from weak duality that $D^* \leq P^*$. If the ℓ_i are convex, it is well-known that this relation holds with equality under some constraint qualification (e.g., Assumption 3) [18]. The next result shows that under mild condition on the distributions \mathfrak{D}_i , equality holds even if the ℓ_i are non-convex. We note that, besides being the first step in the construction of Theorem 3, this result is of independent interest and has been used in other contexts (see, e.g., [30, 38, 39]).

Proposition 1. *Suppose there exists ϕ' such that $\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi'(\mathbf{x}), y)] < c_i$ for $i = 1, \dots, m$. If \mathcal{Y} is finite, the conditional random variable $\mathbf{x}|y$ induced by each of the joint distributions \mathfrak{D}_i is non-atomic, and \mathcal{H} is decomposable, then (P-CSO) is strongly dual, i.e., $P^* = D^*$.*

Proof. See Appendix A.3. ■

Hence, even if (P-CSO) is a non-convex program (e.g., in the rate-constrained learning example of Section 5.1.2), it remains strongly dual under mild conditions. In particular, the conditional $\mathbf{x}|y$ induced by the distributions \mathfrak{D}_i must be non-atomic, i.e., cannot contain Dirac deltas. Additionally, \mathcal{H} must be a *decomposable* function space, i.e., if $\phi, \phi' \in \mathcal{H}$, then for any measurable set \mathcal{Z} it holds that $\bar{\phi} \in \mathcal{H}$ for

$$\bar{\phi}(\mathbf{x}) = \begin{cases} \phi(\mathbf{x}), & \mathbf{x} \in \mathcal{Z} \\ \phi'(\mathbf{x}), & \mathbf{x} \notin \mathcal{Z} \end{cases}.$$

Lebesgue spaces (e.g., L_2 or L_∞) or more generally Orlicz spaces are typical examples of decomposable function spaces [131].

Note that Proposition 1 requires \mathcal{Y} be finite, i.e., this result holds only for classification problems. We address the regression case, i.e., continuous output y , in Appendix A.4 using a slightly stronger assumption on the continuity of the losses. The regression case can then be approximated arbitrarily well by a sequence of ever finer classification problems yielding the required strong duality result (Proposition 22). A similar approach is used in the construction of regression trees [80].

4.1.2 The parameterization gap

Whereas Proposition 1 shows there is no duality gap between (P-CSO) and (D-CSL), the latter remains a variational, statistical problem. The next step is therefore to approximate the functional

space \mathcal{H} using the finite dimensional parametrization $f_{\boldsymbol{\theta}}$, which leads to the finite dimensional problem

$$D_{\boldsymbol{\theta}}^* = \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} \min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\mu}) \triangleq L(f_{\boldsymbol{\theta}}, \boldsymbol{\mu}), \quad (\text{D}_{\boldsymbol{\theta}}\text{-CSL})$$

where L is the Lagrangian in (4.11). Notice that the minimization in (D $_{\boldsymbol{\theta}}$ -CSL) is performed over $\mathcal{P} \subseteq \mathcal{H}$ (Assumption 5). Hence, it is straightforward that $D_{\boldsymbol{\theta}}^* \geq D^* = P^*$. Yet, if the parametrization is rich enough, we would expect the gap $D_{\boldsymbol{\theta}}^* - P^*$ to be small. This intuition is formalized in the following proposition.

Proposition 2. *Let $\boldsymbol{\theta}^*$ achieve the saddle-point in (D $_{\boldsymbol{\theta}}$ -CSL). Under the conditions of Theorem 3, $f_{\boldsymbol{\theta}^*}$ is a feasible, near-optimal solution of (P-CSO). Explicitly,*

$$P^* \leq D_{\boldsymbol{\theta}}^* \leq P^* + (1 + \|\tilde{\boldsymbol{\mu}}^*\|_1) M\nu, \quad (4.12)$$

where P^* and $D_{\boldsymbol{\theta}}^*$ are as in (P-CSO) and (D $_{\boldsymbol{\theta}}$ -CSL) respectively and $\tilde{\boldsymbol{\mu}}^*$ are the dual variables of (P-CSO) with the constraints tightened to $c_i - M\nu$ for $i = 1, \dots, m$.

Proof. See Appendix A.5. ■

Despite being finite dimensional, (D $_{\boldsymbol{\theta}}$ -CSL) remains a statistical problem. Hence, though Proposition 2 establishes that its solutions are not only (P-CSO)-feasible but also near-optimal, the issue remains of how to solve it without explicit knowledge of the distributions \mathfrak{D}_i . Observe, however, that the objective of (D $_{\boldsymbol{\theta}}$ -CSL) involves an unconstrained statistical minimization problem. We have therefore done most of the heavy lifting and can now rely on generalization bounds from classical learning theory.

4.1.3 The empirical gap

The final step to transform (D $_{\boldsymbol{\theta}}$ -CSL) into the empirical dual problem ($\widehat{\text{D}}$ -CSL) is to turn the statistical Lagrangian (4.11) into the empirical (4.4). The estimation error incurred in this step is detailed in the next proposition.

Proposition 3. *Let $\hat{\boldsymbol{\theta}}^*$ achieve the saddle-point in ($\widehat{\text{D}}$ -CSL). Under the conditions of Theorem 3,*

Algorithm 1 Primal-dual constrained learning

- 1: *Initialize:* $\boldsymbol{\theta}^{(0)} = \mathbf{0}$, $\boldsymbol{\mu}^{(0)} = \mathbf{0}$, $\boldsymbol{\lambda}_j^{(0)} = \mathbf{0}$
- 2: **for** $\text{dot} = 1, \dots, T$
- 3: Obtain $\boldsymbol{\theta}^{(t-1)}$ such that $\hat{L}(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}) \leq \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}) + \rho$
- 4: Update dual variables

$$\begin{aligned}\mu_i^{(t)} &= \left[\mu_i^{(t-1)} + \eta \left(\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right) \right]_+ \\ \lambda_{j,n_j}^{(t)} &= \left[\lambda_{j,n_j}^{(t-1)} + \frac{\eta}{N_j} \left(\ell_j(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_j}), y_{n_j}) - c_j \right) \right]_+\end{aligned}$$

5: **end for**

it holds with probability $1 - \delta$ over the samples drawn from the distributions \mathfrak{D}_i that

$$|D_\theta^* - \hat{D}^*| \leq V_{N_0} \quad \text{and} \quad (4.13)$$

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y)] \leq c_i + V_{N_i}. \quad (4.14)$$

where V_N is as in (4.6) and D_θ^* and \hat{D}^* are as in (D $_\theta$ -CSL) and (\hat{D} -CSL) respectively.

Proof. See Appendix A.6. ■

Theorem 3 is obtained directly from Propositions 1–3 using the triangle inequality. Observe that the order in which these transformations is applied to (P-CSO) is crucial. If we begin by replacing the expectations with sample averages, which leads to (P-ECRM), we need \mathcal{H} to be PAC learnable in order to derive any generalization guarantee. To overcome this issue, we could start by parametrizing (P-CSO) using \mathcal{P} . However, as we have argued before, this leads to a non-convex problem for which strong duality does not hold.

4.2 A Constrained Learning Algorithm

Having established that (\hat{D} -CSL) is worth solving, i.e., it is a near-PACC learning of (P-CSO) (Theorems 2 and 3), we now proceed with the issue of *how* it can be solved. We have argued that (\hat{D} -CSL) is preferable to (PIV) for constrained learning because it is unconstrained. That is not to say that (\hat{D} -CSL) is easy to solve, but that it is not harder than classical ERM. We show that this is

the case next by describing a practical algorithm that (approximately) solves ($\widehat{\text{D}}\text{-CSL}$) by (approximately) solving unconstrained learning problems.

To do so, start by noting that the outer maximization in ($\widehat{\text{D}}\text{-CSL}$) is a convex optimization program. Indeed, the dual function $\hat{d}(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) = \min_{\boldsymbol{\theta}} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ is the pointwise minimum of a set of affine functions and is therefore always concave [18]. Additionally, its (sub)gradients can be easily computed by evaluating the constraint slacks at the minimizer of \hat{L} [77, Ch. 3]. Hence, the main challenge in ($\widehat{\text{D}}\text{-CSL}$) is the inner minimization.

Note, however, that this minimization is a classical, unconstrained ERM problem. In fact, it is equivalent to solving an instance of a regularized learning problem. Hence, despite the non-convexity of the Lagrangian (4.4), it is often the case that it is possible to find good minimizers, especially for differentiable losses and parametrizations (i.e., most common ML models). For instance, there is ample empirical and theoretical evidence that gradient descent can learn good parameters for (C)NNs [7, 9, 84, 85, 124]. This is in contrast to (PIV) for which even obtaining a feasible $\boldsymbol{\theta}$ can be intricate. In that vein, we thus assume that we have access to the following oracle:

Assumption 6. There exists an oracle $\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ and $\rho > 0$ such that $\hat{L}(\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j), \boldsymbol{\mu}, \boldsymbol{\lambda}_j) \leq \min_{\boldsymbol{\theta}} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j) + \rho$ for all $\boldsymbol{\mu} \in \mathbb{R}_+^m$ and $\boldsymbol{\lambda}_j \in \mathbb{R}_+^{N_j}$, $j = m+1, \dots, m+q$.

Assumption 6 essentially states that we are able to (approximately) train regularized unconstrained learners using the parametrization $f_{\boldsymbol{\theta}}$. We can then alternate between minimizing the Lagrangian (4.4) with respect to $\boldsymbol{\theta}$ for fixed $(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ and updating the dual variables using the resulting minimizer. This procedure is summarized in Algorithm 1 and analyzed in the following theorem. It is worth noting that this result applies to the deterministic output $(\boldsymbol{\theta}^{(T)}, \boldsymbol{\mu}^{(T)})$ of Algorithm 1 and not to a randomized solution obtained by sampling from $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\mu}^{(t)})$, $t = 0, \dots, T$, as in [57, 58, 62].

Theorem 4. *Under the conditions of Theorem 2 or 3, the primal-dual pair $(\boldsymbol{\theta}^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\lambda}^{(T)})$ obtained after running Algorithm 1 for*

$$T \leq \left\lceil \frac{U_0}{2\eta M\nu} \right\rceil + 1 \text{ steps}$$

with step-size

$$\eta \leq \frac{2\epsilon_0}{\left(m + \frac{q}{N}\right) B^2}$$

satisfies

$$\left| \hat{L}(\boldsymbol{\theta}^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\lambda}^{(T)}) - P^* \right| \leq \rho + \epsilon_0 + V_{N_0} \quad (4.15)$$

with probability $1 - \delta$ over sample sets, where $U_0 = \|\boldsymbol{\mu}^*\|^2 + \sum_{j=m+1}^{m+q} \|\boldsymbol{\lambda}_j^*\|^2$ is the initial distance to the optimal dual variables $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}_j^*)$ of $(\widehat{\text{D}}\text{-CSL})$ and ϵ_0 is as in Corollary 1 or 2.

Proof. See Appendix A.7. ■

Underlying the oracle in Assumption 6 (step 3 in Algorithm 1) is often an iterative procedure, e.g., gradient descent, and the cost of running this procedure until convergence to obtain an approximate minimizer can be prohibitive. A common alternative is to adopt an Arrow-Hurwicz-style approach in which the primal variable $\boldsymbol{\theta}^{(t)}$ and the dual variables $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\lambda}_j^{(t)})$ are updated iteratively [132]. While the convergence guarantee of Theorem 4 no longer holds in this case, good results are observed in practice by performing the primal and dual updates at different timescales, e.g., by performing step 3 once per epoch. This is what we do in the next chapter to illustrate the usefulness of this algorithm in constrained learning applications.

Chapter 5

Applications

In this chapter, we deploy the dual constrained learning results from Chapter 4 in two applications: fairness and robustness. Throughout the experiments, we explore a myriad of models, from logistic regression to fully connected NNs to CNNs, in order to showcase the wide applicability of these results. We also illustrate how the dual variables can be used to obtain information on the sensitivity of the constraints and deliver insights into which requirements are more stringent.

5.1 Fairness

Fair learning has garnered considerable attention recently due to the growing number of reports denouncing biases in learning applications [1–3, 6]. Data-driven models, as it turns out, tend to amplify discrepancies and biases in the data, an issue that has been studied at length in the classical statistics literature. However, the complexity and scale of system and models used in modern ML, combined with the over-reliance on convenience samples, requires that different solutions be put forward. Constrained learning lends itself naturally to address this issue.

Different forms of fairness have been studied in the literature [57–59, 61, 62, 133]. We showcase two of them below: the first based on invariance and the second on rate parity.

5.1.1 Invariance and fair learning

Fairness can be seen as a form of invariance in which a model is trained to be insensitive to certain protected variable (e.g., a gender). All things being equal, the model should have a similar outcome regardless of whether the individual is female or male. Formally, consider a model ϕ whose output is a discrete distribution over k possible classes. Then, (P-CSO) can be used to write

$$\begin{aligned} & \underset{\phi \in \mathcal{H}}{\text{minimize}} && \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell_0(\phi(\mathbf{x}), y) \right] \\ & \text{subject to} && \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[D_{\text{KL}}(\phi(\mathbf{x}) \parallel \phi(\rho(\mathbf{x}))) \right] \leq c, \end{aligned} \tag{PV}$$

where ρ is an input transformation we wish the model to be invariant to (e.g., a specific change in gender) and $c > 0$ determines the level of insensitivity. Formulation (PV) can be extended trivially to multiple transformations. When the average invariance from (PV) is not enough, a stricter, pointwise requirement can be imposed by using

$$D_{\text{KL}}(\phi(\mathbf{x}, z) \parallel \phi(\mathbf{x}, 1 - z)) \leq c \quad \mathcal{D}\text{-a.e.} \tag{5.1}$$

The constraint in (PV) bounds the average causal effect (ACE) and (5.1) relates to *counterfactual fairness* [133].

We begin by showcasing the use of (PV) in the Adult dataset [134], where the goal is to predict whether an individual makes more than US\$ 50,000.00 while being insensitive to gender. The transformations performed on the data are listed in Table 5.1.1. We use a neural network with two outputs and a single hidden-layer with 64 nodes using a sigmoidal activation function. The output is encoded into a probability using a softmax transformation ($f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow [0, 1]^2$). Using this parametrization, we then pose the constrained learning problem

$$\begin{aligned} & \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{minimize}} && \mathbb{E} \left[\ell_0(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \right] \\ & \text{subject to} && D_{\text{KL}}(f_{\boldsymbol{\theta}}(\mathbf{x}, z) \parallel f_{\boldsymbol{\theta}}(\mathbf{x}, 1 - z)) \leq c, \end{aligned} \tag{PVI}$$

where z is the variable gender (encoded 0 for female and 1 for male) and ℓ_0 is the negative logistic log-likelihood, i.e., $-\log \left([f_{\boldsymbol{\theta}}(\mathbf{x})]_y \right)$. To solve (PVI), we use ADAM [135] for step 3 of Algorithm 1,

Table 5.1: Preprocessing of the Adult dataset

| Variable names | Transformation |
|-----------------|--|
| fnlwgt | Dropped |
| educational-num | Dropped |
| relationship | Dropped |
| capital-gain | Dropped |
| capital-loss | Dropped |
| education | Grouped the levels Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th |
| race | Grouped the levels Other and Amer-Indian-Eskimo |
| marital-status | Grouped the levels Married-civ-spouse, Married-AF-spouse, Married-spouse-absent |
| marital-status | Grouped the levels Divorced, Separated |
| race | Grouped the levels Other and Amer-Indian-Eskimo |
| native-country | Grouped the levels Columbia, Cuba, Guatemala, Haiti, Ecuador, El-Salvador, Dominican-Republic, Honduras, Jamaica, Nicaragua, Peru, Trinidad&Tobago |
| native-country | Grouped the levels England, France, Germany, Greece, Holand-Netherlands, Hungary, Italy, Ireland, Portugal, Scotland, Poland, Yugoslavia |
| native-country | Grouped the levels Cambodia, Laos, Philippines, Thailand, Vietnam |
| native-country | Grouped the levels China, Hong, Taiwan |
| native-country | Grouped the levels United-States, Outlying-US(Guam-USVI-etc), Puerto-Rico |
| age | Binned by quantiles (6 bins) |
| hours-per-week | Binned levels into less than 40 and more than 40 |

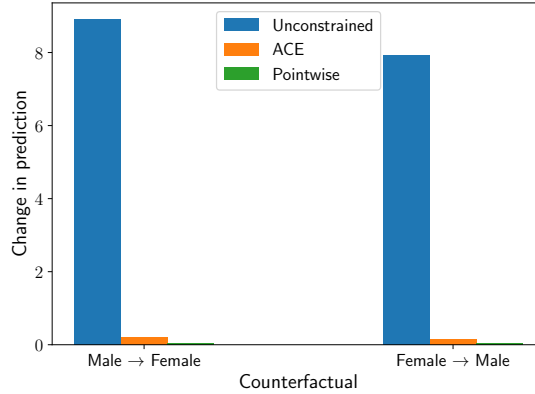


Figure 5.1: Classifier sensitivity on the Adult test set.

with batch size 128 and learning rate 0.1. All other parameters were kept as in the original paper. After each epoch, we update the dual variables (step 4), also using ADAM with a step size of 0.01. We take $c = 10^{-3}$. All classifiers were trained over 300 epochs.

Without the constraint in (PVI), the resulting classifier is quite sensitive to gender: its prediction would change for approximately 8% of the test samples if their gender were reversed (Figure 5.1). With the pointwise constraint, the classifier becomes insensitive to the protected variable in 99.9% of the test set, which is on the order of $1/\sqrt{N} \approx 0.008$. While the less strict ACE can also be imposed, it leads to slightly more sensitive classifiers (for $c = 5 \times 10^{-4}$, the classifier changes prediction in 0.2% of the test set).

Recall that due to the bound on the duality gap, the dual variables of (PVI) obtained in Algorithm 1 have a sensitivity interpretation (Lemma 9): the larger their value, the harder the constraint is to satisfy [18]. Almost 96% of the dual variables are zero after convergence, meaning that the constraint was tight for only 4% of the individuals. In Figure 5.2a, we show the distribution of $\lambda > 0$ over the Adult training set. If we analyze the group with the largest dual variables (the 80% percentile to be exact), we find a significantly higher prevalence of married individuals, non-white, non-US natives, and with a Masters degree (Figure 5.2b). Hence, while attempting to control for gender invariance, the constrained learner also had to overcome other prejudices correlated to sexism in the dataset.

This situation becomes clearer when applying (PV) to the COMPAS dataset [136]. Here, the

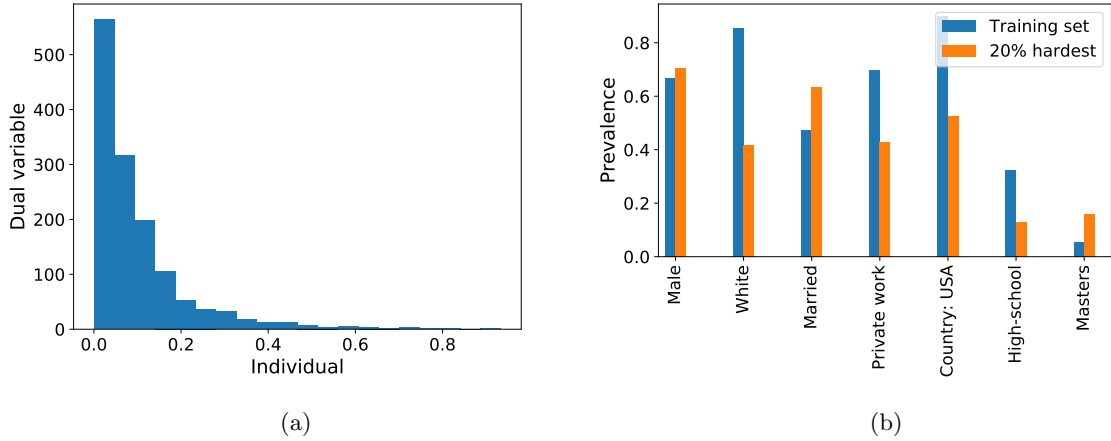


Figure 5.2: Dual variable analysis for Adult dataset: (a) distribution of the dual variables values and (b) prevalence of different groups among the 20% training set examples with largest dual variables.

Table 5.2: Preprocessing of the COMPAS dataset

| Variable names | Transformation |
|------------------|---|
| age_cat | Dropped |
| is_recid | Dropped |
| is_violent_recid | Dropped |
| score_text | Dropped |
| v_score_text | Dropped |
| decile_score | Dropped |
| v_decile_score | Dropped |
| race | Grouped the levels Other, Asian, Native American |
| age | Binned by quantiles (5 bins) |
| priors_count | Binned levels into 0, 1, 2, 3, 4, and more than 4 |
| juv_misd_count | Binned levels into 0, 1, and more than 1 |
| juv_other_count | Binned levels into 0, 1, and more than 1 |

goal is to predict recidivism based on an individual’s past offense data (see Table 5.1.1 for details on the data processing). We use the same neural network as before trained over 400 iterations using a similar procedure, but with batch size 256, primal learning rate 0.1, and dual variables learning rate 2 (halved every 50 iterations). Unconstrained, it reaches an accuracy of almost 70%, but is sensitive to both gender, race, and gender \times race (Table 5.1.1). By including ACE constraints on these counterfactuals, we obtain a classifier that is now invariant to these variables.

Once again, the value of the dual variables bring insights into the different forms of biases in the dataset (Figure 5.3). If we do not include constraints on the cross-term counterfactuals, then the hardest constraint to satisfy is the gender-invariant one. Invariance to the Caucasian:Hispanic

Table 5.3: Classifier insensitivity on the COMPAS dataset

| Counterfactual | Unc. (Acc: 69.4%) | ACE (Acc: 67.9%) |
|--|-------------------|------------------|
| Male \leftrightarrow Female | 21.4% | 0% |
| African-American \leftrightarrow Caucasian | 10.86% | 0% |
| African-American \leftrightarrow Hispanic | 14.32% | 0.02% |
| African-American \leftrightarrow Other | 11.38% | 0% |
| Caucasian \leftrightarrow Hispanic | 9.11% | 0% |
| Caucasian \leftrightarrow Other | 6.54% | 0% |
| Hispanic \leftrightarrow Other | 3.08% | 0% |
| Male \leftrightarrow Female + African-American \leftrightarrow Caucasian | 28.84% | 0.02% |
| Male \leftrightarrow Female + African-American \leftrightarrow Hispanic | 27.47% | 0% |
| Male \leftrightarrow Female + African-American \leftrightarrow Other | 29.17% | 0% |
| Male \leftrightarrow Female + Caucasian \leftrightarrow Hispanic | 22.71% | 0% |
| Male \leftrightarrow Female + Caucasian \leftrightarrow Other | 24.27% | 0% |
| Male \leftrightarrow Female + Hispanic \leftrightarrow Other | 21.15% | 0% |

and Hispanic:Other counterfactuals is effectively “implied” by the other constraints, since their dual variables vanish. If we include all 13 counterfactuals, i.e., add the cross-terms between gender and race, then the cross-terms dominate the satisfaction difficulty, with the Male:Female \times African-American:Caucasian dichotomy dominating over all others. Interestingly, however, the dual variable for the African-American:Caucasian counterfactual does not completely vanish, indicating the existence of a gender-independent race bias in the dataset. This does not occur with other combinations of the race factor. This type of combinatorial (gerrymandering) fairness poses a serious challenge in fair classification [60].

5.1.2 Rate-constrained learning

Many definitions of fairness involve some form of rate constraint parity. Rate constraints, or more precisely probability or chance constraints, have been used in statistics since Neyman-Pearson [86] and have been applied beyond fair learning to control classifier performance, such as its coverage, precision, or accuracy [57, 58, 60, 62]. Explicitly, a rate-constrained learning problems can be written as

$$\begin{aligned}
P^* &= \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_0} [\ell_0(\phi(\mathbf{x}), y)] \\
\text{subject to } & \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} [\mathbb{I}(\mathcal{E})] \leq c_i,
\end{aligned} \tag{P-RCL}$$

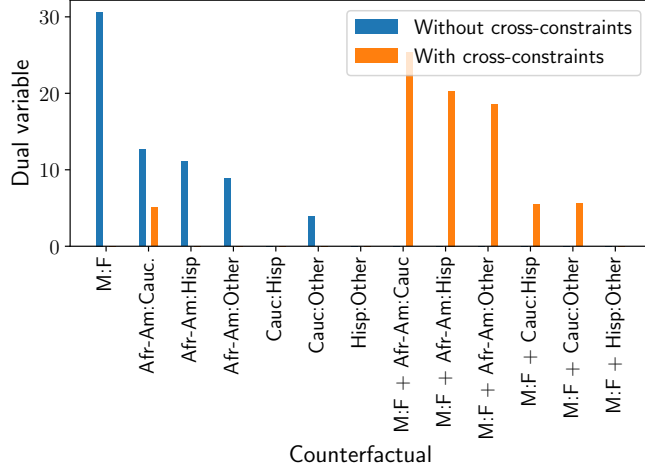


Figure 5.3: Dual variables of different counterfactual constraints for the COMPAS dataset.

where $\mathbb{I}(\mathcal{E})$ denotes the indicator function of an event \mathcal{E} , i.e., $\mathbb{I}(\mathcal{E}) = 1$ when \mathcal{E} occurs and zero otherwise.

Rate constraints are challenging due to their non-convexity and non-differentiability. Hence, most existing generalization guarantees for rate-constrained problems apply to specific algorithms and hold for randomized models [57, 58, 60, 62]. In contrast, Theorems 3 and 4 provide conditions under which (P-RCL) is (near-)PACC learnable, i.e., under which deterministic models can be obtained. Still, the non-differentiability of the indicator function can hinder step 3 of Algorithm 1. Since we only need an approximate solution, we use a smooth surrogate of the indicator in the sequel, namely a sigmoidal function, in order to allow step 3 to be performed using gradient descent. This is a typical approach in the statistics (e.g., logistic models) and rate constraints literature [57, 62].

To illustrate the use of rate constraints in learning, let us revisit the problem of fair classification application using the COMPAS dataset [136]. The goal is to predict recidivism based on the individual’s characteristics and past offenses. Yet, while the overall recidivism rate in the dataset is 45.5%, this rate is 52.3% for African-Americans, which compose 51.4% of the data (Figure 5.4). Considering how this data was collected (based on arrests), we expect this disparity to be due to sampling bias. Training an unconstrained logistic classifier exacerbates this skewness. While its test accuracy is 68.5%, it predicts an overall recidivism rate of 38.4% (the actual rate on the test set

is 45.7%) while maintaining the African-American group rate at 52.2% (Figure 5.4). This classifier was trained over a random sample containing 80% of the dataset using ADAM [135] for 1000 epochs with batch size of 128 samples, learning rate 0.2, and all other parameters as in the original paper. While we use a logistic classifier, the same results are obtained for a single layer feedforward neural network.

We overcome this issue, by imposing an asymmetric form of statistical parity that upper bounds the difference between each protected group’s recidivism rate and the overall one. Explicitly,

$$\begin{aligned} P^* &= \min_{\phi \in \mathcal{H}} \mathbb{E}[\ell_0(\phi(\mathbf{x}), y)] \\ \text{subject to} \quad &\mathbb{E}[\mathbb{I}[\phi(\mathbf{x}) \geq 0.5] \mid \text{Race} = r] \leq \mathbb{E}[\mathbb{I}[\phi(\mathbf{x}) \geq 0.5]] + 0.01, \end{aligned} \tag{PVII}$$

where ℓ_0 is the negative log-likelihood of the logistic distribution and $r = \{\text{African-American, Caucasian, Hispanic, Other}\}$. In other words, the final classifier is required to predict recidivism within each group at most 1% above the rate of the overall population.

We solve (PVII) using a logistic classifier for ϕ trained with Algorithm 1 over $T = 1000$ epochs. For step 3, we used ADAM with the same hyperparameters as above and a sigmoidal approximation for the indicator function. Explicitly, we replaced $\mathbb{I}[\phi(\mathbf{x}) \geq 0.5]$ by $\sigma(a(\phi(\mathbf{x}) - 0.5))$, where σ denotes the sigmoid function. In our simulations, we set $a = 8$. After each epoch, we updated the dual variables (step 5) also using ADAM with step size 0.001. The results are shown in the last row of Figure 5.4.

Notice that compared to the unconstrained model, the predicted recidivism rate over the test set remains almost the same (38.9%), but the rates within each group are now more homogeneous. For instance, the rate for African-Americans is now only 1.5% above the overall one. Interestingly, the model now predicts Caucasians have a considerably higher recidivism, from 24.07% in the unconstrained model to 39.8%, which is much closer to the actual rate in the data set (39.1%). In doing so, the test accuracy of (PVII) (66.3%) remains close to that of the unconstrained classifier.

Since we used a logistic classifier, we can interpret its coefficients as odds ratio to analyze the differences between the constrained and unconstrained models. The coefficients with largest changes are displayed in Figure 5.5. Note that while the original model estimates that being African-American increases your chances of recidivism by almost 30%, the constrained model compensates for the

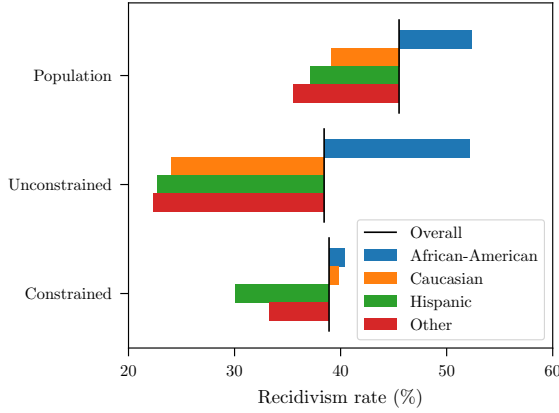


Figure 5.4: Dual variable relative to the fairness constraint.

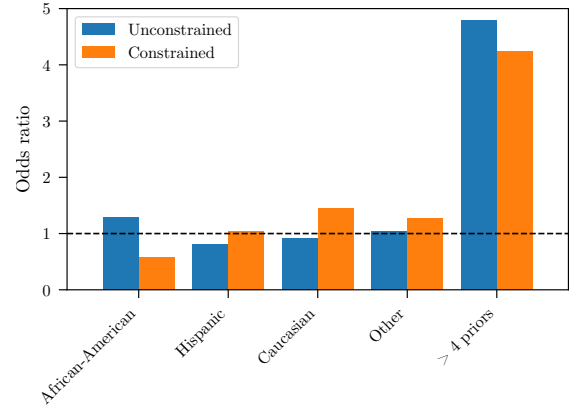


Figure 5.5: Dual variable relative to the fairness constraint.

dataset biases by instead decreasing the probability by 40%. The opposite effect occurs in the Caucasian group. The model also compensates for the individual having a large number of priors, a group composed mostly of African-Americans in the dataset (69%).

5.2 Robustness

Another issue affecting ML models, especially CNNs, is robustness. It is straightforward to construct small input perturbations that lead to misclassification and there are now numerous methods to do so. Although adversarial training has been successfully used to train robust ML models, it often leads to solutions with poor nominal performance, i.e., poor performance on original, clean data [11, 15, 63, 78, 79, 137]. To overcome this issue, [64] poses a constrained learning that explicitly trades-off nominal performance and with the model invariance to a worst-case perturbation. They propose an algorithm that optimizes over an upper bound of this robustness constraint, leading to solutions that are accurate on clean data and have similar classification on noisy inputs. Similarly, we address this nominal performance vs. robustness compromise by using (P-CSO) to constraint the performance of the classifier on an adversarial distribution (rather than making it invariant as in

Section 5.1.1). Explicitly, we pose

$$\begin{aligned} & \underset{\phi \in \mathcal{H}}{\text{minimize}} && \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell_0(\phi(\mathbf{x}), y) \right] \\ & \text{subject to} && \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{A}} \left[\ell_0(\phi(\mathbf{x}), y) \right] \leq c \end{aligned} \tag{PVIII}$$

where \mathfrak{A} is an adversarial data distributions. Hence, we optimize against a given adversarial distribution \mathfrak{A} rather a worst-case one. This distribution can be tailored to provide a smooth performance degradation, as we illustrate next.

Consider the problem of training a ResNet18 [138] to classify images from the FMNIST dataset [139]. We reserve 100 images from each class sampled at random for validation. When trained without constraints over 100 epochs using the ADAM optimizer with the settings from [135] and batches of 128 images, it reaches its best accuracy over the validation set after 67 epochs. The nominal accuracy of this solution (over the test set) is 93.5%. However, when the input is attacked using PGD [63], it fails to classify any of the test images for perturbation magnitudes as low as $\varepsilon = 0.04$ (Figure 5.6a). In what follows, ε indicates the maximum pixel modification allowed (ℓ_∞ -norm of the perturbation) and we run the PGD attack using a step size of $\varepsilon/30$ for 50 iterations displaying the worst result over 10 restarts, unless stated otherwise.

A first attempt is then to use PGD with $\varepsilon = 0.04$ to sample from a hypothetical adversarial distribution and constrain its performance against that distribution as in (PVIII). Though the adversarial distribution is now dependent on the model ϕ , by using a smaller learning rate for the dual variables, ϕ can be considered almost static for the dual update and we have observed no instability issues in practice. To accelerate training, we use a much weaker attack running PGD without restarts for only 5 steps with step size $\varepsilon/3$. Notice from Figure 5.6a that when training against $\varepsilon = 0.04$ ($c = 0.4$), the resulting classifier trades-off nominal performance (now 88%) for adversarial performance (now 85%). However, as the strength of the attack increases, the performance of the classifier deteriorates abruptly: for $\varepsilon = 0.08$, it is down to 9%. Increasing the training adversarial strength to $\varepsilon = 0.1$ ($c = 0.7$) yields a more robust classifier, albeit at the cost of a lower nominal accuracy (84.6%). Still, the performance degradation remains quite abrupt.

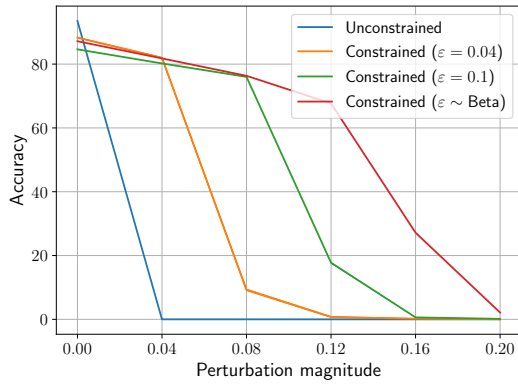
This issue can be fixed by training against using a hierarchical adversarial distribution. Explicitly,

we build the adversarial distribution \mathfrak{A} as

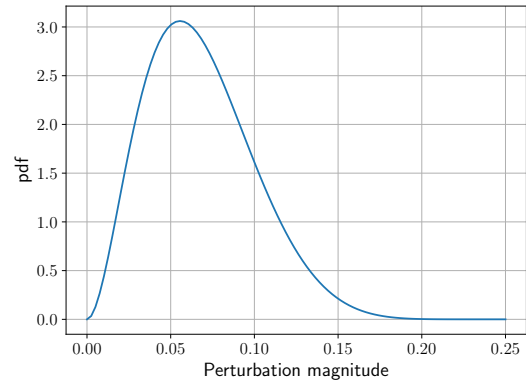
$$\Pr(\mathfrak{A}) = \Pr(\mathfrak{A} \mid \varepsilon) \Pr(\varepsilon), \quad (5.2)$$

where $\Pr(\mathfrak{A} \mid \varepsilon)$ is induced by an adversarial attack of magnitude at most ε (in our case, PGD) and $\Pr(\varepsilon)$ denotes a prior distribution on the magnitude of the attacks. In Figure 5.6a we take $\varepsilon \sim 0.25 \times \text{Beta}(3, 8)$ (Figure 5.6b). Notice that even though the mean value of the perturbation is approximately 0.07, the resulting classifier has a nominal performance close to 87% and retains a 67% accuracy for perturbations of magnitude up to 0.12.

Similar results are obtained when training a ResNet18 [138] to classify images in the CIFAR-10 dataset. The training was performed as above, once again reserving 100 random images from each class sampled for validation. The unconstrained classifier trained over 100 epochs reached its best accuracy over the validation set after 82 epochs, which corresponds to a nominal test accuracy of 85.4%. However, when the input is attacked using PGD [63], the accuracy falls to 5% already for $\varepsilon = 0.01$ (Figure 5.7a). When using the fixed ε training method described above, we once again observe a trade-off between nominal accuracy and robustness. This can, however, be improved using the hierarchical training technique from (5.2). Taking $\varepsilon \sim 0.1 \times \text{Beta}(3, 10)$, such that $\mathbb{E}[\varepsilon] = 0.02$, we obtain the same nominal accuracy as for the fixed- ε , but improve the robustness for higher perturbation values.

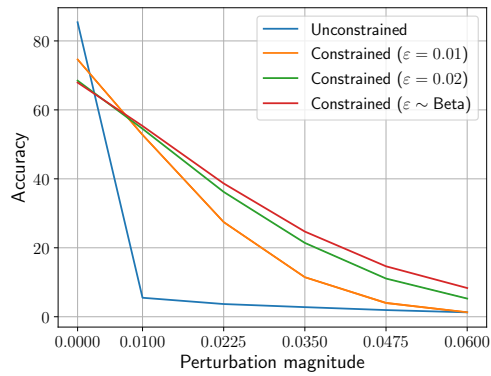


(a)

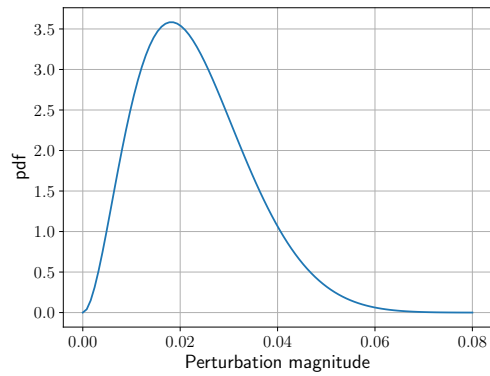


(b)

Figure 5.6: Robust constrained learning (FMNIST): (a) Accuracy of classifiers under the PGD attack for different perturbation magnitudes and (b) distribution of ϵ used during training.



(a)



(b)

Figure 5.7: Robust constrained learning (CIFAR-10): (a) Accuracy of classifiers under the PGD attack for different perturbation magnitudes and (b) distribution of ϵ used during training.

Part II

Constrained inference

Chapter 6

Sparse nonlinear functional models

In this chapter, we address the issue of nonlinear functional inference under sparsity constraints. These constraints are commonly used to obtain parsimonious, interpretable solutions, reduce measurement costs, or enable estimation problems to be solved in challenging, high-dimensional scenarios. Although we now consider the inference setting in which models (distributions) are available to us analytically, the estimation problems tackled next are by no means easy. In fact, they have the appearance of intractability due to their infinite dimensional, nonlinear, non-convex properties. Yet, these functional, nonlinear statistical inference problems appear in numerous applications, such as spectral estimation, image recovery, source localization, magnetic resonance fingerprinting, spectrum cartography, and manifold data sparse coding [87–93].

The infinite dimensionality and nonlinearity challenges of these applications are often tackled by imposing structure on the signals. For instance, bandlimited, finite rate of innovation, or union-of-subspaces signals can be processed using an appropriate discrete set of samples [94–97]. Functions in reproducing kernel Hilbert spaces (RKHSs) also admit finite descriptions through variational results known as representer theorems [140, 141]. The infinite dimensionality of continuous problems is therefore often overcome by means of sampling theorems. Similarly, nonlinear functions with bounded total variation or lying in an RKHS can be written as a finite linear combination of basis functions. Under certain smoothness assumptions, nonlinearity can be addressed using “linear-in-the-parameters” methods.

In these contexts, sparsity priors play an important role in achieving state-of-the-art results since discretization often leads to underdetermined problems. This is achieved by leveraging the assumption that there exists a signal representation in terms of only a few atoms from an overparametrized dictionary [69, 97, 98]. Since fitting these models leads to non-convex (and possibly NP-hard [99]) problems, sparsity is typically replaced by a tractable relaxation based on an atomic norm (e.g., the ℓ_1 -norm). For linear, incoherent dictionaries, these relaxed problems have been shown to retrieve the desired sparse solution [69, 98]. Nevertheless, discretized continuous problems rarely meet these conditions. Only in specific instances, such as line spectrum estimation, there exist guarantees for relaxations that forgo discretization [96, 106–111].

This discretization/relaxation approach, however, is not always effective. Indeed, discretization can lead to grid mismatch issues and even loss of sparsity: infinite dimensional sparse signals need not be sparse when discretized [103–105]. Also, sampling theorems are sensitive to the function class considered and are often asymptotic: results improve as the discretization becomes finer. This leads to high dimensional statistical problems with potentially poor numerical properties (high condition number). In fact, ℓ_1 -norm-based recovery of spikes on fine grids (essentially) finds twice the number of actual spikes and the number of support candidate points increases as the number of measurements decreases [142, 143]. Furthermore, performance guarantees for convex relaxations rely on incoherence assumptions (e.g., restricted isometry/eigenvalue properties) that may be difficult to meet in practice and are NP-hard to check [100–102]. Finally, these guarantees hold for linear measurements models.

Directly accounting for nonlinearities in sparse models makes a difficult problem harder, since the optimization program remains non-convex even after relaxing the sparsity objective. This is evidenced by the weaker guarantees existing for ℓ_1 -norm relaxations in nonlinear compressive sensing problems [144, 145]. Though “linear-in-the-parameters” models, such as splines or kernel methods, may sometimes be used (e.g., spectrum cartography [92]), they are not applicable in general. Indeed, the number of kernels needed to represent a generic nonlinear model may be so large that the solution is no longer sparse. What is more, there is no guarantee that these models meet the incoherence assumptions required for the convex relaxation to be effective [69, 97, 98].

In what follows, we forgo both discretization and relaxation to directly tackle the continuous problem using *sparse functional programming*. SFPs, briefly introduced in Section 2.2.2.1 combine

the infinite dimensionality of functional programming with the non-convexity of sparsity and non-linear atoms. Nevertheless, they turn out to be tractable under mild conditions (Theorem 5). This result also yields a relation between minimizing the support of a function (“ L_0 -norm”) and its L_1 -norm, although the latter may yield non-sparse solutions. We illustrate the expressiveness of SFPs by casting different signal processing problems and providing numerical examples that showcase their effectiveness.

Recall that in order to be consistent with the Bayesian inference literature, we reverse the roles of \mathbf{x} and \mathbf{y} in this part. In other words, we consider the measurements \mathbf{y} as observations based on which we wish to estimate the hidden variables \mathbf{x} . Thus, instead of observing both \mathbf{x} and \mathbf{y} but not \mathfrak{D}_i (learning, as in Part I), we observe \mathbf{y} and \mathfrak{D} , but not \mathbf{x} .

6.1 Sparse functional programs and duality

6.1.1 Sparse functional programs

Sparse functional programs (SFPs) are variational problems that seek sparsest functions, i.e., functions with minimum support measure. Explicitly, let (Ω, \mathcal{B}) be a measurable space in which \mathcal{B} are the Borel sets of Ω , a compact set of \mathbb{R}^n . In a parallel with the discrete case, define the L_0 -norm to be the measure of the support of a function, i.e., for a measurable function $X : \Omega \rightarrow \mathbb{C}$,

$$\|X\|_{L_0} = \mathfrak{m}[\text{supp}(X)] = \int_{\Omega} \mathbb{I}[X(\boldsymbol{\beta}) \neq 0] d\boldsymbol{\beta}. \quad (6.1)$$

Note that the integral in (6.1) is a multivariate integral over vectors $\boldsymbol{\beta} \in \Omega$. Unless otherwise specified, all integrals are taken with respect to the Lebesgue measure \mathfrak{m} . As in the discrete case, the “ L_0 -norm” in (6.1) is not a norm, but we omit the quotation marks so as not to burden the text.

A general SFP is then defined as the optimization problem

$$\begin{aligned}
& \underset{X \in \mathcal{X}, \mathbf{z} \in \mathbb{C}^p}{\text{minimize}} && \int_{\Omega} F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \lambda \|X\|_{L_0} \\
& \text{subject to} && g_i(\mathbf{z}) \leq 0, \quad i = 1, \dots, m \\
& && \mathbf{z} = \int_{\Omega} \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} \\
& && X(\boldsymbol{\beta}) \in \mathcal{P} \text{ a.e.}
\end{aligned} \tag{P-SFP}$$

where $\lambda > 0$ is a regularization parameter that controls the sparsity of the solution; $g_i : \mathbb{C}^p \rightarrow \mathbb{R}$ are convex functions; $F_0 : \mathbb{C} \times \Omega \rightarrow \mathbb{R}$ is an optional, not necessarily convex regularization term (e.g., take $F_0(x, \boldsymbol{\beta}) = |x|^2$ for shrinkage); $\mathbf{F} : \mathbb{C} \times \Omega \rightarrow \mathbb{C}^p$ is a vector-valued (possibly nonlinear) function; \mathcal{P} is a (possibly non-convex) set defining an almost everywhere (a.e.) pointwise constraints on X , i.e., a constraint that holds for all $\boldsymbol{\beta} \in \Omega$ except perhaps over a set of measure zero (e.g., $\mathcal{P} = \{x \in \mathbb{C} \mid |x| \leq \Gamma\}$ for some $\Gamma > 0$); and \mathcal{X} is a *decomposable* function space as in Chapter 4, i.e., if $X, X' \in \mathcal{X}$, then for any $\mathcal{Z} \in \mathcal{B}$ it holds that $\bar{X} \in \mathcal{X}$ for

$$\bar{X}(\boldsymbol{\beta}) = \begin{cases} X(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{Z} \\ X'(\boldsymbol{\beta}), & \boldsymbol{\beta} \notin \mathcal{Z} \end{cases}.$$

Lebesgue spaces (e.g., $\mathcal{X} = L_2$ or $\mathcal{X} = L_{\infty}$) or more generally Orlicz spaces are typical examples of decomposable function spaces. The space of constant functions, for instance, is *not* decomposable [131].

The linear, continuous sparse recovery/denoising problem is a particular case of (P-SFP). Here, we seek to represent a signal $\mathbf{y} \in \mathbb{R}^p$ as a linear combination of a continuum of atoms $\phi(\beta)$ indexed by $\beta \in \Omega \subset \mathbb{R}$, i.e., as $\int_{\Omega} X(\beta)\phi(\beta)d\beta$, using a sparse functional coefficient X . This problem can be posed as

$$\begin{aligned}
& \underset{X \in L_2, \hat{\mathbf{y}} \in \mathbb{R}^p}{\text{minimize}} && \|X\|_{L_2}^2 + \lambda \|X\|_{L_0} \\
& \text{subject to} && \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \leq \epsilon \\
& && \hat{\mathbf{y}} = \int_{\Omega} X(\beta)\phi(\beta)d\beta,
\end{aligned} \tag{PIX}$$

where $\epsilon > 0$ is a goodness-of-fit parameter. Notice that when discretized, this problem yields the

classical (NP-hard [99]) dictionary denoising problem with $\hat{\mathbf{y}} = \boldsymbol{\phi}^T \mathbf{x}$, where $\boldsymbol{\phi} = [\phi(\beta_j)]$ and $\mathbf{x} = [X(\beta_j)]$, for a set of $\beta_j \in \Omega$, $j = 1, \dots, m$.

The expressiveness of SFPs comes from their ability to accommodate nonlinear measurement models (through \mathbf{F}) and non-convex objective functions. For instance, (PIX) can also be posed using the nonlinear model

$$\hat{\mathbf{y}} = \int_{\Omega} \rho [X(\beta)\phi(\beta)] d\beta, \quad (6.2)$$

where ρ represents, for instance, a source saturation (as in, e.g., Section 6.4). Yet, the abstract formulation in (P-SFP) certainly obfuscates the applicability of SFPs. Additionally, severe technical challenges, such as infinite dimensionality and non-convexity, appear to hinder their usefulness. We defer the issue of applicability to Section 6.4, where we illustrate the use of SFPs in the context of nonlinear spectral estimation and nonlinear functional data analysis. Instead, we first focus on whether problems of the form (P-SFP) can even be solved. Indeed, note that the discrete versions of certain SFPs are known to be NP-hard [99]. Hence, discretizing the functional problem in this case makes it intractable.

We propose to solve SFPs using duality. It is worth noting that duality is often used to solve semi-infinite convex programs [106–108, 146]. In these cases, strong duality holds under mild conditions and solving the dual problem leads to a solution of the original optimization problem of interest. However, SFPs are not convex. To address this issue, we first derive the dual problem of (P-SFP) in the next section, noting that it is both finite dimensional and convex. Then, we show that we can obtain a solution of (P-SFP) from a solution of its dual by proving that SFPs have zero duality gap under quite general conditions (Section 6.2). Finally, we suggest different algorithms to solve the dual problem of (P-SFP) (Section 6.3).

6.1.2 The Lagrangian dual of sparse functional programs

To formulate the dual problem of (P-SFP), we first introduce the Lagrange multipliers $\nu_i \in \mathbb{R}_+$, corresponding to the inequalities $g_i(\mathbf{z}) \leq 0$, and $\boldsymbol{\mu}_R, \boldsymbol{\mu}_I \in \mathbb{R}^p$, corresponding to the real and imaginary parts respectively of the complex-valued equality $\mathbf{z} = \int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta$. To simplify the derivations, we collect the former into the vector $\boldsymbol{\nu} \in \mathbb{R}_+^m$ and combine the latter two multipliers into a single complex-valued dual variable by noticing that for any vector $\mathbf{x} \in \mathbb{C}^m$, it holds

that $\boldsymbol{\mu}_R^T \Re[\mathbf{x}] + \boldsymbol{\mu}_I^T \Im[\mathbf{x}] = \Re[\boldsymbol{\mu}^H \mathbf{x}]$, where $\boldsymbol{\mu} = \boldsymbol{\mu}_R + j\boldsymbol{\mu}_I$.

The Lagrangian dual of (P-SFP) is then defined as

$$\begin{aligned} \mathcal{L}(X, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\nu}) = & \int_{\Omega} F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \lambda \|X\|_{L_0} \\ & + \Re \left[\boldsymbol{\mu}^H \left(\int_{\Omega} \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} - \mathbf{z} \right) \right] \\ & + \sum_{i=1}^m \nu_i g_i(\mathbf{z}) \end{aligned} \quad (6.3)$$

and its dual function is given by

$$d(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\substack{X \in \mathcal{X}, \mathbf{z} \in \mathbb{C}^p, \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \mathcal{L}(X, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\nu}). \quad (6.4)$$

The fact that the pointwise constraint holds almost everywhere in Ω is omitted for conciseness.

Thus, the dual problem of (P-SFP) is given by

$$\underset{\boldsymbol{\mu}, \boldsymbol{\nu} \succeq 0}{\text{maximize}} \quad d(\boldsymbol{\mu}, \boldsymbol{\nu}). \quad (\text{D-SFP})$$

By definition, (D-SFP) is a convex program whose dimensionality is equal to the number of constraints [130]—in this case, on the order of p . It is therefore tractable as long as we can evaluate the dual function d . Indeed, solving (D-SFP) is at least as hard as solving the minimization in (6.4). We next show that the dual function of SFPs is often efficiently computable.

The joint minimization in (6.4) separates as

$$d(\boldsymbol{\mu}, \boldsymbol{\nu}) = d_X(\boldsymbol{\mu}) + d_{\mathbf{z}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \quad (6.5)$$

with

$$\begin{aligned} d_X(\boldsymbol{\mu}) = & \min_{\substack{X \in \mathcal{X}, \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \int_{\Omega} \left\{ F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] + \lambda \mathbb{I}[X(\boldsymbol{\beta}) \neq 0] \right. \\ & \left. + \Re[\boldsymbol{\mu}^H \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}]] \right\} d\boldsymbol{\beta} \end{aligned} \quad (6.6)$$

and $d_{\mathbf{z}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{z}} \sum_{i=1}^m \nu_i g_i(\mathbf{z}) - \Re[\boldsymbol{\mu}^H \mathbf{z}]$. The minimum in $d_{\mathbf{z}}$ is tractable since the objective is convex, given that $\nu_i \geq 0$ and the g_i are convex functions. In certain cases, e.g., when g_i is a

quadratic loss, $d_{\mathbf{z}}$ may even have a closed-form expression. On the other hand, d_X is in general a non-convex problem. When F_0 and \mathbf{F} are normal integrands [131, Def. 14.27], this issue is addressed by exploiting the separability of the objective across $\boldsymbol{\beta}$ as shown in Proposition 4. Examples of normal integrands include functions $f(x, \boldsymbol{\beta})$ that are continuous in x for all fixed $\boldsymbol{\beta}$ and measurable in $\boldsymbol{\beta}$ for all fixed x (also known as Carathéodory) or when Ω is Borel and $f(\cdot, \boldsymbol{\beta})$ is lower semicontinuous for all fixed $\boldsymbol{\beta}$ [131]. Note that these functions can be nonlinear and need not be convex.

Proposition 4. *Consider the functional optimization problem in (6.6) and assume that F_0 and the elements of \mathbf{F} are normal integrands. Let $\gamma^{(0)}(\boldsymbol{\mu}, \boldsymbol{\beta}) = F_0(0, \boldsymbol{\beta}) + \mathbb{R}e [\boldsymbol{\mu}^H \mathbf{F}(0, \boldsymbol{\beta})]$ and define*

$$\gamma^o(\boldsymbol{\mu}, \boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathbb{R}e [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})]. \quad (6.7)$$

Then, for $\mathcal{S}(\boldsymbol{\mu}) = \{\boldsymbol{\beta} \in \Omega : \gamma^o(\boldsymbol{\mu}, \boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}, \boldsymbol{\beta}) - \lambda\}$,

$$d_X(\boldsymbol{\mu}) = \int_{\mathcal{S}(\boldsymbol{\mu})} [\lambda + \gamma^o(\boldsymbol{\mu}, \boldsymbol{\beta})] d\boldsymbol{\beta} + \int_{\Omega \setminus \mathcal{S}(\boldsymbol{\mu})} \gamma^{(0)}(\boldsymbol{\mu}, \boldsymbol{\beta}) d\boldsymbol{\beta}. \quad (6.8)$$

Proof. We start by separating the objective of (6.6) using the following lemma:

Lemma 1 (Separability principle [131, Thm. 14.60]). *Let $G(x, \boldsymbol{\beta})$ be a normal integrand and \mathcal{X} be a decomposable space. Then,*

$$\inf_{\substack{X \in \mathcal{X} \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \int_{\Omega} G[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} = \int_{\Omega} \inf_{x \in \mathcal{P}} G(x, \boldsymbol{\beta}) d\boldsymbol{\beta}. \quad (6.9)$$

Since Ω is a compact subset of \mathbb{R}^n and the indicator function is lower semicontinuous [131, Ex. 14.31], the integrand in (6.6) is normal and we can restrict ourselves to solving the optimization individually for each $\boldsymbol{\beta}$, i.e., if X_d is a solution of (6.6), then

$$X_d(\boldsymbol{\beta}) \in \operatorname{argmin}_{x \in \mathcal{P}} F_0[x, \boldsymbol{\beta}] + \lambda \mathbb{I}[x \neq 0] + \mathbb{R}e [\boldsymbol{\mu}^H \mathbf{F}[x, \boldsymbol{\beta}]]. \quad (\text{PX})$$

Despite the non-convexity of the indicator function, (PX) is a scalar problem, whose solution involves a simple thresholding scheme. Indeed, only two conditions need to be checked: (i) if $X_d(\boldsymbol{\beta}) = 0$, then the indicator function vanishes and the objective of (PX) evaluates to $\gamma^{(0)}(\boldsymbol{\beta})$; (ii) if $X_d(\boldsymbol{\beta}) \neq 0$, then

the indicator function is one and the objective of (PX) evaluates to $\lambda + \gamma^o(\beta)$. The value of (PX) is the minimum of these two cases, which from Lemma 1 yields the desired result in (6.8). ■

Proposition 4 provides a practical way to evaluate (6.5), i.e., to evaluate the objective of the dual problem (D-SFP). Still, it relies on the ability to efficiently solve (6.7), which may be an issue if F_0 , \mathbf{F} , or \mathcal{P} are non-convex. Nevertheless, (6.7) remains a scalar problem that can typically be solved efficiently using global optimization techniques [147] or through efficient local search procedures (see Section 6.4).

The tractability of the dual problem (D-SFP) does not imply that it provides a solution to the original problem (P-SFP). In fact, since SFPs are not convex programs it is not immediate that (D-SFP) is worth solving at all: there is no reason to expect that the optimal value of (D-SFP) is anything more than a lower bound on the optimal value of (P-SFP) [130]. In the sequel, we proceed to show that this is not the case and that we can actually obtain a solution of (P-SFP) by solving (D-SFP).

6.2 Strong duality and its implications

Though we have argued that the dual problem of (P-SFP) is potentially tractable, we are ultimately interested in solving (P-SFP) itself. This section tackles this limitation by showing that (P-SFP) and (D-SFP) have the same values (Theorem 5). In Section 6.3, we show how, under mild conditions, this result allows us to efficiently find a solution for (P-SFP). Before that, however, we use strong duality to derive a relation between SFPs and L_1 -norm optimization problems when their solution saturates (Section 6.2.2).

6.2.1 Strong duality of sparse functional programs

The main result of this section is presented in the following theorem:

Theorem 5. *Suppose that F_0 and \mathbf{F} have no point masses (Dirac deltas) and that there exists a (P-SFP)-feasible pair (X', \mathbf{z}') , i.e., $X' \in \mathcal{X}$ with $X'(\beta) \in \mathcal{P}$ a.e. and $\mathbf{z}' \in \mathbb{C}^p$, such that $\mathbf{z}' = \int_{\Omega} \mathbf{F}[X'(\beta), \beta] d\beta$ and $g_i(\mathbf{z}') < 0$ for all $i = 1, \dots, m$. Then, strong duality holds for (P-SFP), i.e., if P^* is the optimal value of (P-SFP) and D^* is the optimal value of (D-SFP), then $P^* = D^*$.*

Theorem 5 states that although (P-SFP) is a non-convex functional program, it has zero duality gap, suggesting that it can be solved through its tractable dual (D-SFP). A noteworthy feature of this approach is that it precludes discretization by tackling (P-SFP) directly. Discretizing (P-SFP) may not only result in NP-hard problems, but leads to high dimensional, potentially ill-conditioned problems. It is also worth noting that Theorem 5 is *non-parametric* in the sense that it makes no assumption on the existence or validity of the measurement model in (P-SFP). In particular, it does not require that the data arise from a specific model in which the parameters are sparse. This implies, for instance, that the sparsest functional linear model that fits a set of measurements can be determined regardless of whether these measurements arise from a truly sparse, linear model. This is useful in practice when sparse solutions are sought, not for epistemological reasons, but for reducing computational or measurement costs.

Proof of Thm. 5. Recall from weak duality that the dual problem is a lower bound on the value of the primal, so that $D^* \leq P^*$ [130]. Hence, it suffices to prove that $D^* \geq P^*$. To do so, denote the cost function of (P-SFP) by $f_0(X) \triangleq \int_{\Omega} F_0[X(\beta), \beta] d\beta + \lambda \|X\|_{L_0}$ and define the cost-constraints set

$$\begin{aligned} \mathcal{C} = \left\{ (c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I) \mid \exists (X, \mathbf{z}) \in \mathcal{X} \times \mathbb{C}^p \text{ such that} \right. \\ \left. X(\beta) \in \mathcal{P} \text{ a.e., } f_0(X) \leq c, g_i(\mathbf{z}) \leq [\mathbf{u}]_i, \right. \\ \left. \text{and } \int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta - \mathbf{z} = \mathbf{k}_R + j\mathbf{k}_I \right\}. \end{aligned} \quad (6.10)$$

In words, \mathcal{C} describes the range of values taken by the objective and constraints of (P-SFP). Observe that (6.10) separates the real- and complex-valued parts of the equality constraint in (P-SFP). Hence, $\mathcal{C} \subset \mathbb{R}^{2p+m+1}$, allowing us to directly leverage classical convex geometry results. The crux of this proof is to show that \mathcal{C} is a convex set even though (P-SFP) is not a convex program. We summarize this result in the following technical lemma whose proof relies on Lyapunov's convexity theorem [148]:

Lemma 2. *Under the assumptions of Theorem 5, the cost-constraints set \mathcal{C} in (6.10) is a non-empty convex set.*

Proof. See Appendix B.1.1. ■

We may then leverage the following result from convex geometry:

Proposition 5 (Supporting hyperplane theorem [18, Prop. 1.5.1]). *Let $\mathcal{A} \subset \mathbb{R}^n$ be a nonempty convex set. If $\tilde{\mathbf{x}} \in \mathbb{R}^n$ is not in the interior of \mathcal{A} , then there exists a hyperplane passing through $\tilde{\mathbf{x}}$ such that \mathcal{A} is in one of its closed halfspaces, i.e., there exists $\mathbf{p} \neq \mathbf{0}$ such that $\mathbf{p}^T \tilde{\mathbf{x}} \leq \mathbf{p}^T \mathbf{x}$ for all $\mathbf{x} \in \mathcal{A}$.*

Formally, start by observing that the point $(P^*, \mathbf{0}, \mathbf{0}, \mathbf{0})$ cannot be in the interior of \mathcal{C} . Indeed, there would otherwise exist $\delta > 0$ such that $(P^* - \delta, \mathbf{0}, \mathbf{0}, \mathbf{0}) \in \mathcal{C}$, violating the optimality of P . Proposition 5 therefore implies that there exists a non-zero vector $(\lambda_0, \boldsymbol{\nu}, \boldsymbol{\mu}_R, \boldsymbol{\mu}_I) \in \mathbb{R}^{2p+m+1}$ such that for all $(c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I) \in \mathcal{C}$,

$$\lambda_0 c + \boldsymbol{\nu}^T \mathbf{u} + \boldsymbol{\mu}_R^T \mathbf{k}_R + \boldsymbol{\mu}_I^T \mathbf{k}_I \geq \lambda_0 P^*. \quad (6.11)$$

Observe that the vector defining the hyperplane uses the same notation as for the dual variables of (P-SFP) foreshadowing the fact that these hyperplanes span the values of the Lagrangian (6.3). From (6.11), we immediately obtain that $\lambda_0 \geq 0$ and $\boldsymbol{\nu} \succeq \mathbf{0}$. Indeed, note that \mathcal{C} is unbounded above in its first $m+1$ components, i.e., if $(c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I) \in \mathcal{C}$ then $(c', \mathbf{u}', \mathbf{k}_R, \mathbf{k}_I) \in \mathcal{C}$ for any $(c', \mathbf{u}') \succeq (c, \mathbf{u})$. Hence, if any component of λ_0 or $\boldsymbol{\nu}$ were negative, there would exist a vector in \mathcal{C} that makes the left-hand side of (6.11) arbitrarily small, eventually violating the inequality. Let us now show that $\lambda_0 \neq 0$.

To do so, suppose $\lambda_0 = 0$. Then, (6.11) reduces to

$$\boldsymbol{\nu}^T \mathbf{u} + \boldsymbol{\mu}_R^T \mathbf{k}_R + \boldsymbol{\mu}_I^T \mathbf{k}_I \geq 0, \quad (6.12)$$

for all $(c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I) \in \mathcal{C}$. However, (6.12) leads to a contradiction because its left-hand side can always be made negative. Indeed, if $[\boldsymbol{\nu}]_i > 0$ for any i , then the hypothesis on the existence of a strictly feasible point for (P-SFP) violates (6.12). Explicitly, since there exists (X', \mathbf{z}') such that $\int_{\Omega} \mathbf{F}[X'(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} = \mathbf{z}'$ and $g_i(\mathbf{z}') < 0$ for all $i = 1, \dots, m$, then $(c_0, -\delta \mathbf{1}, \mathbf{0}, \mathbf{0}) \in \mathcal{C}$ for some c_0 and $\delta > 0$, where $\mathbf{1}$ is a vector of ones. Thus, if $[\boldsymbol{\nu}]_i > 0$ for any i , we obtain $-\delta(\boldsymbol{\nu}^T \mathbf{1}) < 0$, which violates (6.12).

On the other hand, if $\boldsymbol{\nu} = \mathbf{0}$, then (6.12) reduces to $\boldsymbol{\mu}_R^T \mathbf{k}_R + \boldsymbol{\mu}_I^T \mathbf{k}_I \geq 0$ which cannot hold because for $(\bar{c}, \bar{\mathbf{u}}, -\boldsymbol{\mu}_R, -\boldsymbol{\mu}_I) \in \mathcal{C}$, (6.12) evaluates to $-\|\boldsymbol{\mu}_R\|^2 - \|\boldsymbol{\mu}_I\|^2 < 0$. To see that this vector is indeed an element of \mathcal{C} , simply choose any $\bar{X} \in \mathcal{C}$ with $\bar{X}(\beta) \in \mathcal{P}$ a.e. and let $\bar{\mathbf{z}} = -\boldsymbol{\mu}_R - j\boldsymbol{\mu}_I - \int_{\Omega} \mathbf{F}[\bar{X}(\beta), \beta] d\beta$, $[\bar{\mathbf{u}}]_i = g_i(\bar{\mathbf{z}})$, and $\bar{c} = f_0(\bar{X})$. Hence, it must be that $\lambda_0 \neq 0$.

However, for $\lambda_0 \neq 0$, (6.11)

$$c + \tilde{\boldsymbol{\nu}}^T \mathbf{u} + \tilde{\boldsymbol{\mu}}_R^T \mathbf{k}_R + \tilde{\boldsymbol{\mu}}_I^T \mathbf{k}_I \geq P,$$

where $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu}/\lambda_0$, $\tilde{\boldsymbol{\mu}}_R = \boldsymbol{\mu}_R/\lambda_0$, and $\tilde{\boldsymbol{\mu}}_I = \boldsymbol{\mu}_I/\lambda_0$, which from the definition of \mathcal{C} implies that

$$\begin{aligned} f_0(X) + \sum_{i=1}^m \tilde{\nu}_i g_i(\mathbf{z}) + \tilde{\boldsymbol{\mu}}_R^T \Re \left[\int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta - \mathbf{z} \right] \\ + \tilde{\boldsymbol{\mu}}_I^T \Im \left[\int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta - \mathbf{z} \right] \geq P^*, \end{aligned} \quad (6.13)$$

for any (P-SFP)-feasible pair (X, \mathbf{z}) . Letting $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}_R + j\tilde{\boldsymbol{\mu}}_I$, we recognize that (6.13) in fact bounds the value of the Lagrangian in (6.3) for any (P-SFP)-feasible pair (X, \mathbf{z}) , i.e., $\mathcal{L}(X, \mathbf{z}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}) \geq P^*$. Taking the minimum of the left-hand side of (6.13) hence implies $D^* \geq P^*$, thus concluding the proof. \blacksquare

6.2.2 SFPs and L_1 -norm optimization problems

Similar to the discrete case, there is a close relation between L_0 - and L_1 -norm minimization. Formally, consider

$$\begin{aligned} & \underset{X \in L_{\infty}, \mathbf{z} \in \mathcal{C}^p}{\text{minimize}} && \|X\|_{L_q} \\ & \text{subject to} && g_i(\mathbf{z}) \leq 0 \\ & && \mathbf{z} = \int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta \\ & && |X| \leq \Gamma \text{ a.e.} \end{aligned} \quad (\text{P}_q)$$

Problem (P₀) [i.e., (P_q) with $q = 0$] is an instance of (P-SFP) without regularization ($F_0 \equiv 0$) in which \mathcal{P} is the set of measurable functions bounded by $\Gamma > 0$. On the other hand, (P₁) [(P_q) for $q = 1$] is a functional version of the classical ℓ_1 -norm minimization problem. The following

proposition shows that for a wide class of dictionaries, the optimal values of (P_0) and (P_1) are the same (up to a constant).

Proposition 6. *Let $x^o(\boldsymbol{\mu}, \boldsymbol{\beta}) = \operatorname{argmin}_{|x| \leq \Gamma} |x| + \operatorname{Re} [\boldsymbol{\mu}^T \mathbf{F}(x, \boldsymbol{\beta})]$ saturate, i.e., $x^o(\boldsymbol{\mu}, \boldsymbol{\beta}) \neq 0 \Rightarrow |x^o(\boldsymbol{\mu}, \boldsymbol{\beta})| = \Gamma$ for all $\boldsymbol{\mu} \in \mathbb{C}^p$ and $\boldsymbol{\beta} \in \Omega$. If P_0^* (P_1^*) is the optimal value of (P_q) for $q = 0$ ($q = 1$) and Slater's condition holds, then*

$$P_0^* = \frac{P_1^*}{\Gamma}.$$

Proof. The proof follows by relating the dual values of (P_q) for $q = \{0, 1\}$ and then using strong duality. Start by defining the Lagrangian of (P_q) as

$$\begin{aligned} \mathcal{L}(X, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \|X\|_{L_q} + \sum_i \nu_i g_i(\mathbf{z}) \\ &\quad + \operatorname{Re} \left[\boldsymbol{\mu}^H \left(\int_{\Omega} \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} - \mathbf{z} \right) \right]. \end{aligned} \quad (6.14)$$

For $q = 0$, Proposition 1 yields

$$\begin{aligned} d_0(\boldsymbol{\mu}, \boldsymbol{\nu}) &= \int_{S_0(\boldsymbol{\mu})} \left\{ 1 + \min_{|x| \leq \Gamma} \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] \right\} d\boldsymbol{\beta} \\ &\quad + w(\boldsymbol{\mu}, \boldsymbol{\nu}). \end{aligned} \quad (6.15)$$

where

$$S_0(\boldsymbol{\mu}) = \{\boldsymbol{\beta} \in \Omega \mid \min_{|x| \leq \Gamma} \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] < -1\} \quad (6.16)$$

and $w(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{z}} \sum_i \nu_i g_i(\mathbf{z}) - \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{z}]$. Notice that w is homogeneous, i.e., $w(\alpha \boldsymbol{\mu}, \alpha \boldsymbol{\nu}) = \alpha w(\boldsymbol{\mu}, \boldsymbol{\nu})$ for $\alpha > 0$. Proceeding similarly from (6.14), the dual function of (P_1) is

$$\begin{aligned} d_1(\boldsymbol{\mu}, \boldsymbol{\nu}) &= \int_{\Omega} \left\{ \min_{|x| \leq \Gamma} |x| + \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] \right\} d\boldsymbol{\beta} \\ &\quad + w(\boldsymbol{\mu}, \boldsymbol{\nu}). \end{aligned} \quad (6.17)$$

Using the the saturation hypothesis, the integrand in (6.17) is non-trivial only over the set

$$S_1(\boldsymbol{\mu}) = \{\boldsymbol{\beta} \in \Omega \mid \min_{|x| \leq \Gamma} \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] < -\Gamma\}. \quad (6.18)$$

Hence,

$$d_1(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_{\mathcal{S}_1(\boldsymbol{\mu})} \left\{ \Gamma + \min_{|x| \leq \Gamma} \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] \right\} d\boldsymbol{\beta} + w(\boldsymbol{\mu}, \boldsymbol{\nu}). \quad (6.19)$$

To proceed, note that the dual functions in (6.15) and (6.19) are related by

$$d_0(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{\Gamma} d_1(\Gamma \boldsymbol{\mu}, \Gamma \boldsymbol{\nu}). \quad (6.20)$$

Indeed, observe from (6.16) and (6.18) that $\mathcal{S}_1(\Gamma \boldsymbol{\mu}) = \mathcal{S}_0(\boldsymbol{\mu})$. Thus,

$$\begin{aligned} \frac{1}{\Gamma} \int_{\mathcal{S}_1(\Gamma \boldsymbol{\mu})} \left\{ \Gamma + \min_{|x| \leq \Gamma} \operatorname{Re} [\Gamma \boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] \right\} d\boldsymbol{\beta} \\ = \int_{\mathcal{S}_0(\boldsymbol{\mu})} \left\{ 1 + \min_{|x| \leq \Gamma} \operatorname{Re} [\boldsymbol{\mu}^H \mathbf{F}(x, \boldsymbol{\beta})] \right\} d\boldsymbol{\beta}. \end{aligned}$$

The homogeneity of w then yields (6.20). Immediately, it holds that if $(\boldsymbol{\mu}^o, \boldsymbol{\nu}^o)$ is a maximum of d_0 , then $(\Gamma \boldsymbol{\mu}^o, \Gamma \boldsymbol{\nu}^o)$ is a maximum of d_1 . To see this is the case, note from (6.20) that

$$\nabla d_0(\boldsymbol{\mu}^o, \boldsymbol{\nu}^o) = \mathbf{0} \Leftrightarrow \nabla d_1(\Gamma \boldsymbol{\mu}^o, \Gamma \boldsymbol{\nu}^o) = \mathbf{0},$$

so that $(\Gamma \boldsymbol{\mu}^o, \Gamma \boldsymbol{\nu}^o)$ is a critical point of d_1 . Since d_1 is a concave function, $(\Gamma \boldsymbol{\mu}^o, \Gamma \boldsymbol{\nu}^o)$ must be a global maximum.

To conclude, observe that (P_q) has zero duality gap for both $q = 0$, due to Theorem 5, and $q = 1$, because it is a convex program. From (6.20) we then obtain

$$\begin{aligned} P_0^* &= \max_{\boldsymbol{\mu}, \boldsymbol{\nu} \geq 0} d_0(\boldsymbol{\mu}, \boldsymbol{\nu}) = d_0(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) \\ &= \frac{1}{\Gamma} d_1(\Gamma \boldsymbol{\mu}^*, \Gamma \boldsymbol{\nu}^*) = \frac{1}{\Gamma} \max_{\boldsymbol{\mu}, \boldsymbol{\nu} \geq 0} d_1(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{P_1^*}{\Gamma}. \end{aligned} \quad \blacksquare$$

Proposition 6 shows that a large class of L_0 - and L_1 -norm minimization problems found in functional nonlinear sparse recovery are equivalent in the sense that their optimal values are (essentially) the same. It is worth noting that establishing this relation requires virtually no assumptions: the saturation hypothesis is met by a wide class of dictionaries, most notably linear ones. This is in con-

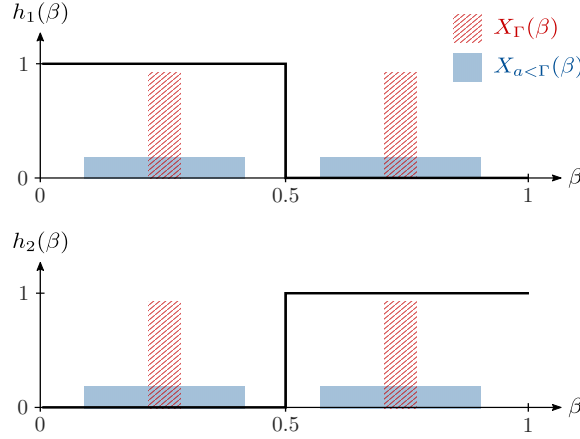


Figure 6.1: Illustration of Example 1.

trast to the discrete case, where such relations exist for incoherent, linear dictionaries [69, 98]. Still, Proposition 6 does not imply that the solution of the L_0 - and L_1 -norm problems are the same, as is the case for discrete results. In fact, though they have the same optimal value, (P_1) admits solutions with larger support (see Example 1). Although conditions exist for which the L_1 -norm minimization problem with linear dictionaries yields minimum support solutions [104, 105, 111], Theorem 5 precludes the use of this relaxation for both linear and nonlinear dictionaries.

Example 1. Proposition 6 gives an equivalence between L_0 - and L_1 -norm minimization problems in terms of their optimal values, but not in terms of their solutions. We illustrate this point using the example depicted in Figure 6.1. Let $\Omega = [0, 1]$, $g(\mathbf{z}) = \|\mathbf{y} - \mathbf{z}\|_2^2$, $\mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}^T$ with $|y_1|, |y_2| < \Gamma/2$, and $\mathbf{F}(x, \beta) = \mathbf{h}(\beta)x$, where $\mathbf{h}(\beta) = \begin{bmatrix} h'(\beta) & 1 - h'(\beta) \end{bmatrix}^T$ with $h'(\beta) = \mathbb{I}(\beta \in [0, 1/2])$. Due to the form of h' , it is readily seen that the optimal value of (P_1) is $P_1^* = |y_1| + |y_2|$.

Now consider the family of functions indexed by $0 < a \leq \Gamma$

$$X_a(\beta) = a \operatorname{sign}(y_1) \mathbb{I}(\beta \in \mathcal{A}_1) + a \operatorname{sign}(y_2) \mathbb{I}(\beta \in \mathcal{A}_2), \quad (6.21)$$

where $\mathcal{A}_1 \subseteq [0, 1/2]$ with $\|\mathcal{A}_1\|_{L_0} = |y_1|/a$ and $\mathcal{A}_2 \subseteq [1/2, 1]$ with $\|\mathcal{A}_2\|_{L_0} = |y_2|/a$ (e.g., X_Γ and $X_{a<\Gamma}$ in Figure 6.1). For all a , X_a is a solution of (P_1) [it is (P_q) -feasible with value P_1^*]. However, its support is given by $\|X_a\|_{L_0} = (|y_1| + |y_2|)/a$. Thus, (P_1) admits solutions that do not have minimum support ($X_{a<\Gamma}$ in Figure 6.1), whereas only X_Γ is a solution of (P_0) .

6.3 Solving sparse functional programs

Theorem 5 from the previous section establishes duality as a fruitful approach for solving the sparse functional program (P-SFP). Indeed, the strong duality of (P-SFP) implies that

$$(X^*, \mathbf{z}^*) \in \underset{\substack{X \in \mathcal{X}, \mathbf{z} \in \mathbb{C}^p, \\ X(\beta) \in \mathcal{P}}}{\operatorname{argmin}} \mathcal{L}(X, \mathbf{z}, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*), \quad (6.22)$$

for the Lagrangian \mathcal{L} in (6.3), where $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ are the solutions of (D-SFP) [130]. When this set is a singleton, the inclusion becomes equality and we recover the unique primal solution X^* . This occurs when the Lagrangian (6.3) has a single minimizer, i.e., when (PX) is a singleton. This is the case, for instance, when $F_0(x, \beta) = |x|^2$, in which case \mathcal{L} is strongly convex in X [130]. Since Proposition 4 allows us to solve (6.22), all that remains is to address the issue of solving (D-SFP) to obtain $(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$.

Note that (D-SFP) is a convex program and can therefore be solved using any (stochastic) convex optimization algorithm [77, 130]. For illustration, this section introduces an algorithm based on supergradient ascent. For ease of reference, a step-by-step guide to solving SFPs is presented in Appendix B.1.2.

Recall that a supergradient of a concave function $f : \Omega \rightarrow \mathbb{R}$ at $\mathbf{x} \in \Omega$ is any vector \mathbf{p} that satisfies the inequality $f(\mathbf{y}) \leq f(\mathbf{x}) + \mathbf{p}^T(\mathbf{y} - \mathbf{x})$ for all $\mathbf{y} \in \Omega$. Although a supergradient may not be an ascent direction at \mathbf{x} , taking small steps in its direction decreases the distance to any maximizer of f [130].

It is straightforward to show that the constraint slacks in (6.3) are supergradients of the dual function d with respect to their corresponding dual variables [130]. Explicitly,

$$\mathbf{p}_{\boldsymbol{\mu}}(\boldsymbol{\mu}', \boldsymbol{\nu}') = \int_{\Omega} \mathbf{F}[X_d(\boldsymbol{\mu}', \beta), \beta] d\beta - \mathbf{z}_d(\boldsymbol{\mu}', \boldsymbol{\nu}') \quad (6.23a)$$

$$\mathbf{p}_{\boldsymbol{\nu}_i}(\boldsymbol{\mu}', \boldsymbol{\nu}') = g_i[\mathbf{z}_d(\boldsymbol{\mu}', \boldsymbol{\nu}')] \quad (6.23b)$$

Algorithm 2 Dual ascent for SFPs

```

 $\boldsymbol{\mu}^{(0)} = \mathbf{0}, \nu_i^{(0)} = 1$ 
for  $t = 1, \dots, T$  do
   $X_{t-1}(\boldsymbol{\beta}) = X_d(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\beta})$ 
   $\mathbf{z}_{t-1} = \mathbf{z}_d(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$ 
   $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \eta_t \left[ \int_{\Omega} \mathbf{F}[X_{t-1}(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} - \mathbf{z}_{t-1} \right]$ 
   $\nu_i^{(t)} = \left[ \nu_i^{(t-1)} + \eta_t g_i(\mathbf{z}_{t-1}) \right]_+$ 
   $P_t = d(\boldsymbol{\mu}^{(t)}, \boldsymbol{\nu}^{(t)})$ 
end for
 $X^*(\boldsymbol{\beta}) = X_d(\boldsymbol{\mu}^{(t^*)}, \boldsymbol{\beta})$  for  $t^* \in \operatorname{argmax}_{1 \leq t \leq T} P_t$ 

```

are supergradients of d for the dual minimizers

$$\begin{aligned}
X_d(\boldsymbol{\mu}, \cdot) \in \operatorname{argmin}_{\substack{X \in \mathcal{X} \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \int_{\Omega} \left\{ F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] + \lambda \mathbb{I}[X(\boldsymbol{\beta}) \neq 0] \right. \\
\left. + \operatorname{Re}[\boldsymbol{\mu}^H \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}]] \right\} d\boldsymbol{\beta},
\end{aligned} \tag{6.24a}$$

$$\mathbf{z}_d(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \operatorname{argmin}_{\mathbf{z}} \sum_i \nu_i g_i(\mathbf{z}) - \operatorname{Re}[\boldsymbol{\mu}^H \mathbf{z}]. \tag{6.24b}$$

Algorithm 2, with step size $\eta_t > 0$, then yields the optimal dual variables $(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$ and a solution X^* of (P-SFP).

Given that the optimization problem in (6.24b) is convex, there are two hurdles in evaluating (6.23): (i) obtaining X_d involves solving the non-convex, infinite dimensional problem in (6.24a) and (ii) the integral in (6.23a) may not have an explicit form or this form is too cumbersome to be useful in practice. We have already argued that despite its non-convexity, the minimization in (6.24a) is tractable by exploiting separability (see Proposition 4). The resulting scalar problem often has a closed-form solution (see Section 6.4 for examples) or can be tackled using global optimization techniques [147]. Note that though this approach does not explicitly yield the function X_d , it allows $X_d(\boldsymbol{\mu}, \boldsymbol{\beta})$ to be evaluated for any $\boldsymbol{\beta} \in \Omega$ using (PX). This is enough to numerically compute the integral in (6.23a). This integral [(ii)] may either be approximated numerically or done without by leveraging stochastic optimization techniques. Step-by-step descriptions of both methods are

presented in Appendix B.1.2.

In the first case, we effectively solve a perturbed version of (P-SFP) and the difference between the optimal value of the original problem and that obtained numerically depends linearly on the precision of the integral computation under mild technical conditions:

Proposition 7. *Suppose that*

- (i) *the perturbation function of (P-SFP) is differentiable around the origin;*
- (ii) $\int_{\Omega} F_0(0, \beta) d\beta = 0$ *and* $\int_{\Omega} \mathbf{F}(0, \beta) d\beta = \mathbf{0}$;
- (iii) *there exists* $\alpha > 0$ *such that* $g_i(\alpha \mathbf{1}), g_i(-\alpha \mathbf{1}) < \infty$; *and*
- (iv) *there exists a strictly feasible pair* $(X^\dagger, \mathbf{z}^\dagger)$ *(Slater's condition) for (P-SFP) such that* $g_i(\mathbf{z}^\dagger) < -\epsilon$, *for* $\epsilon > 0$, *and* $\bar{F}_0 = \int_{\Omega} F_0(X^\dagger(\beta), \beta) d\beta < \infty$.

If P^\star *is the optimal value of (P-SFP) and* P_δ^\star *is the value of the solution obtained by Algorithm 2 when evaluating the integral in (6.23a) with approximation error* $0 < \delta \ll 1$, *then* $|P^\star - P_\delta^\star| \leq \mathcal{O}(\delta)$.

Proof. See Appendix B.1.3. ■

In the second case, the integral in (6.23a) is approximated using Monte Carlo integration, i.e., by drawing a set of β_j independently and uniformly at random from Ω and taking

$$\hat{\mathbf{p}}_\mu = \frac{1}{N} \sum_{j=1}^N \mathbf{F}[X_d(\mu', \beta_j), \beta_j] - \mathbf{z}_d(\mu', \nu'). \quad (6.25)$$

Since Monte Carlo integration is an unbiased estimators, $\hat{\mathbf{p}}_\mu$ is an unbiased estimate of \mathbf{p}_μ . Taking $N = 1$ in (6.25) is akin to performing stochastic (super)gradient ascent on the dual function d . For $N > 1$, we obtain a mini-batch type algorithm. Typical convergence guarantees hold in both cases [130, 149, 150].

Algorithm 2, though effective, may converge slowly depending on the numerical properties of the problem. Faster, problem independent convergence rates can be obtained using, for instance, second-order methods or by exploiting specific structures of SFP instances. Investigating the use and fit of these approaches to solving (D-SFP) is, however, beyond the scope of this thesis.

6.4 Applications

So far, we have focused on whether SFPs are tractable. In this section, we illustrate their expressiveness by using (P-SFP) to cast the problems of nonlinear spectral estimation and robust functional data classification.

6.4.1 Nonlinear line spectrum estimation

The first example application of SFPs is in the context of continuous, possibly nonlinear, sparse dictionary recovery/denoising problems. Formally, let $\mathbf{y} \in \mathbb{C}^p$ collect samples y_i , $i = 1, \dots, p$, of a signal. Our goal is to represent \mathbf{y} using as few atoms as possible from the nonlinear dictionary

$$\mathcal{D} = \{\mathbf{F}(\cdot, \boldsymbol{\beta}) : \mathbb{C} \rightarrow \mathbb{C}^p \mid \boldsymbol{\beta} \in \Omega\}. \quad (6.26)$$

Explicitly, we wish to find $\{(x_k, \boldsymbol{\beta}_k)\}$ such that

$$\hat{\mathbf{y}} = \sum_{k=1}^K \mathbf{F}(x_k, \boldsymbol{\beta}_k) \quad (6.27)$$

is close to \mathbf{y} for some small K . Notice that, in contrast to classical dictionary recovery, the relation between the coefficients x_k and the signal $\hat{\mathbf{y}}$ is not necessarily linear. Moreover, Ω is an uncountable set, so that we select from a continuum of atoms as opposed to the discrete, finite case.

To make the discussion concrete, consider the problem of estimating the parameters of a small number of saturated sinusoids from samples of their superposition. This problem is found in several signal processing applications, such as telecommunication and direction of arrival (DOA) estimation, where nonlinear behaviors are common due to hardware limitations of the sources. Formally, we wish to estimate the frequencies, amplitudes, and phases of K sinusoids from the set of noisy samples

$$y_i = \sum_{k=1}^K \rho [a_k \cos(2\pi f_k t_i)] + n_i, \quad \text{for } i = 1, \dots, p, \quad (6.28)$$

where $f_k \in [0, 1/2]$ is the frequency and $a_k \in \mathbb{R}$ is the amplitude/phase of the k -th component; t_i is the fixed, known sampling time of the i -th sample; $\{n_i\}$ are independent and identically dis-

tributed (i.i.d.) zero-mean random variables with variance $\mathbb{E} n_i^2 = \sigma_n^2$ representing the measurement noise; and ρ is a function that models the source nonlinearity with $\rho(0) = 0$.

To pose this estimation problem as an SFP, we need an approximate continuous representation of the signal model in (6.28). We say approximate because the nonlinearity ρ may prevent us from finding a measurable function X such that $\int \rho[X(\varphi) \cos(2\pi\varphi t_i)] d\varphi = \rho[x \cos(2\pi f t_i)]$ for a fixed amplitude-frequency pair (x, f) . Even if ρ allows it, an exact representation may involve Dirac deltas, which violates a hypothesis of Theorem 5 and prevents us from efficiently finding a solution of (P-SFP). The following proposition introduces a functional signal model that approximates (6.28) arbitrarily well using parameters in L_2 .

Proposition 8. *For fixed $a, t \in \mathbb{R}$, $f \in [0, 1/2]$, define the hyperparameter $B \in \mathbb{R}_+$ and let*

$$r(B) = B \int_0^{\frac{1}{2}} \rho[X'(\varphi) \cos(2\pi\varphi t)] d\varphi. \quad (6.29)$$

If $X'(\varphi) = a$ for $\varphi \in [f - B^{-1}, f + B^{-1}]$ and zero everywhere else, then $r(B) \rightarrow \rho[a \cos(2\pi f t)]$ as $B \rightarrow \infty$.

Proof. Note that (6.29) is equivalent to

$$r(B) = \int_0^{\frac{1}{2}} B \cdot \Pi_{f, B^{-1}}(\varphi) \rho[a \cos(2\pi\varphi t)] d\varphi$$

with $\Pi_{f, b}(\varphi) = \mathbb{I}(\varphi \in [f - b, f + b])$. The result then follows from the fact that $B \cdot \Pi_{f, B^{-1}}(\varphi)$ converges weakly to $\delta(\varphi - f)$ as $B \rightarrow \infty$, where δ is the Dirac's delta [151]. ■

Proposition 8 allows us to cast nonlinear line spectrum estimation as the SFP

$$\begin{aligned} & \underset{X \in L_2}{\text{minimize}} && \|X\|_{L_2} + \lambda \|X\|_{L_0} \\ & \text{subject to} && \sum_{i=1}^p (y_i - \hat{y}_i)^2 \leq \epsilon \\ & && \hat{y}_i = B \int_0^{\frac{1}{2}} \rho[X(\varphi) \cos(2\pi\varphi t_i)] d\varphi, \\ & && \text{for } i = 1, \dots, p, \end{aligned} \quad (\text{PXI})$$

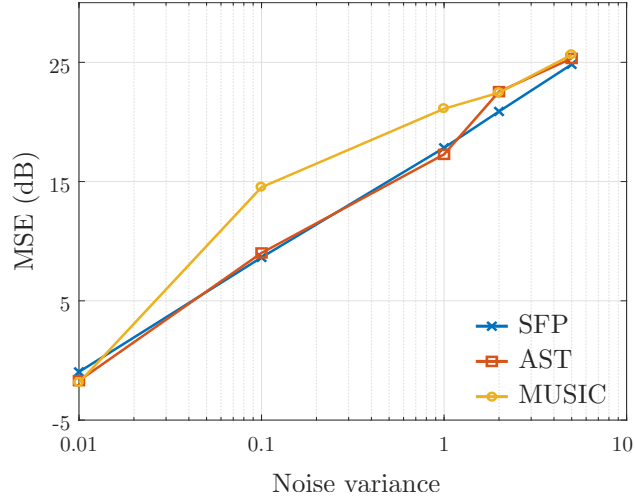


Figure 6.2: Reconstruction MSE for line spectral estimation of linear sources.

where $B > 0$ is an approximation parameter and $\epsilon > 0$ determines the solution fit. Problem (PXI) explicitly seeks the sparsest function X that fits the observations given the model in (6.28). The L_2 -norm regularization improves robustness to noise as well as the numerical properties of the dual by adding shrinkage. Note that since $X \in L_2$, the solution X^* of (PXI) does not contain atoms and is instead a superposition of bump functions around the component frequencies f_k (see, e.g., Figure 6.4). As Proposition 8 suggests, the height and width of each bump depends on the amplitude of the sinusoidal component and the choice of B . Thus, the parameter a_k from (6.28) can be estimated using

$$\hat{a}_k = B \int_{\mathcal{B}_k} X^*(\varphi) d\varphi, \quad (6.30)$$

where X^* is a solution of (PXI) and $\mathcal{B}_k \subset [0, 1/2]$ contains a single bump. The parameter f_k can then be estimated using the center frequency of the bump. Naturally, B should be as large as possible so that (6.29) is a good approximation of (6.28), improving the parameter estimates. Choosing B too large, however, degrades the numerical properties of the dual problem, making it harder to solve in practice. Similar trade-offs are found several methods when tuning regularization parameters, for instance, elastic net [152, 153].

Since ρ is an arbitrary function, a particular case of (PXI) performs spectral estimation with linear sources, i.e., when $\rho(z) = z$ in (6.28) and there is no saturation. The dual function is

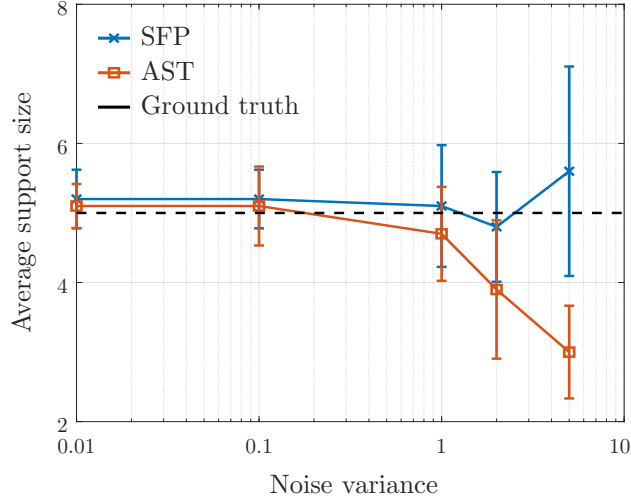


Figure 6.3: Support size estimation for line spectral estimation of linear sources.

straightforward to evaluate in this case since the optimization problem (6.7) from Proposition 4 becomes a quadratic program that admits a closed-form solution. However, a myriad of classical methods such as MUSIC or atomic soft thresholding (AST) have been proposed for the linear case. MUSIC performs line spectrum estimation using the eigendecomposition of the empirical autocorrelation matrix of the measurements y_i [87]. Nevertheless, it can only be used in single snapshot applications when the signal is sampled regularly—see [87] for details—and requires that the number K of components be known *a priori*. The AST approach, on the other hand, is based on an atomic norm relaxation of the sparse estimation problem and leverages duality and spectral properties of Toeplitz matrices to preclude discretization [106, 107]. Both methods first obtain the component frequencies and then determine amplitudes and phases using least squares. These different approaches are compared in Figures 6.2 and 6.3.

These plots display the average performance over 10 realizations that used $p = 61$ samples ($t_i = -30, \dots, 30$) of the superposition of $K = 5$ components whose the frequencies f_k were drawn uniformly at random with a minimum spacing of $4/p$ and whose amplitudes a_k were taken randomly and independently from $[0.5, 3]$. Problem (PXI) was solved using the approximate supergradient method described in Appendix B.1.2 with $B = 1$, $\lambda = 5000$ for all noise levels except $\sigma_n^2 = 5$ which used $\lambda = 6000$, and $\epsilon = p\sigma_n^2$. For the AST method, we used the optimal regularization from [107]

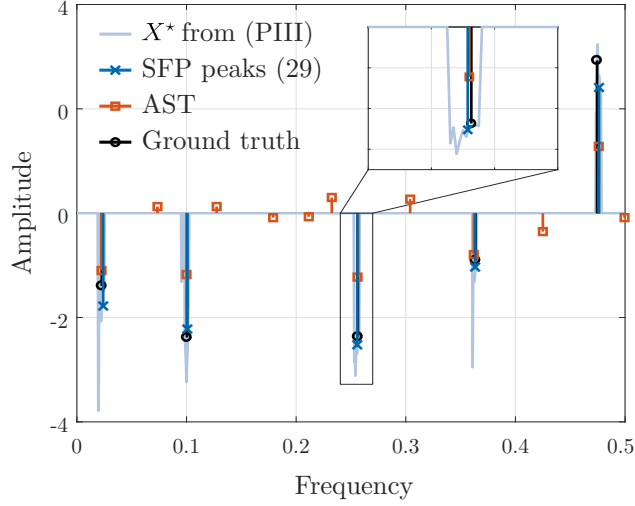


Figure 6.4: Solutions obtained for line spectral estimation of saturated sources.

which depends on σ_n^2 . In all cases, the reconstruction MSE is evaluated as

$$\text{MSE} = \sum_{i=1}^p (y_i - \hat{y}_i)^2,$$

where \hat{y}_i denotes the samples reconstructed based on the K components with largest magnitudes obtained by each algorithm. For AST, the support is obtained from the peaks of the trigonometric polynomial defined by the dual as in [106, 107] and for SFP, from the center of the bumps in the solution X^* of (PXI) (as illustrated in Figure 6.4).

In high SNR scenarios, all methods display similar performance. As the level of noise increases, however, the advantages of explicitly minimizing the L_0 -norm instead of its convex surrogate become clearer, especially with respect to support identification. Observe in Figure 6.3 that as σ_n^2 increases the number of components obtained from AST decreases considerably, despite using the optimal regularization parameter. Finally, it is worth noting that although the performances are similar, AST involves solving a semidefinite program (SDP), which becomes infeasible in practice as the number of samples p grows and has motivated the study of dimensionality reduction techniques and sampling patterns [154]. On the other hand, efficient solvers based on coordinate ascent can be leveraged to solve large-scale SFPs [77, 130].

Still, the signal reconstruction performance is similar across methods in the linear case (Fig-

ure 6.2). This is not surprising given the close relation between L_0 - and L_1 -norm minimization (Theorem 6). In contrast, when the signals are distorted by a nonlinearity, the linear measurement model of AST and MUSIC tends to underestimate the amplitude of the components (Figure 6.4). Though greedy approaches to atomic norm minimization are able to deal with nonlinear dictionaries, optimally selecting single atoms from these infinite dimensional dictionary is challenging. Exhaustive, grid-based heuristics have been proposed for specific problems without guarantees [155].

To illustrate this effect, consider the hard saturation

$$\rho(x) = \begin{cases} x, & |x| \leq r \\ r \cdot \text{sign}(x), & \text{otherwise} \end{cases}, \quad (6.31)$$

where $r > 0$ defines the saturation level. Though computing the dual function may seem challenging in this case due to the nonlinearity, it turns out to be tractable due to the scalar nature of the problem. Indeed, we obtain the dual minimizer from Proposition 4 by evaluating

$$\gamma^o(\boldsymbol{\mu}, \varphi) = \min_{x \in \mathbb{R}} x^2 + \boldsymbol{\mu}^T \rho[x\mathbf{h}(\varphi)],$$

where $[\mathbf{h}]_i = \cos(2\pi\varphi t_i)$ for $i = 1, \dots, p$ and the function ρ applies element-wise. Since we can determine *a priori* which of the elements will saturate, solving this non-convex problem actually reduces to finding the minimum value of p quadratic problems. Namely, assume that \mathbf{h} is sorted such that $h_1 \leq \dots \leq h_p$ and let $\mathbf{w}_i(x) = [h_1 x \ \dots \ h_i x \ r \ \dots \ r]^T$, where r is the saturation level from (6.31). For conciseness, we omit the dependence on φ . Then, $\gamma^o(\boldsymbol{\mu}) = \min_{1 \leq i \leq p} \gamma_i^o(\boldsymbol{\mu})$ for

$$\begin{aligned} \gamma_i^o(\boldsymbol{\mu}) &= \min_{1/|h_{i+1}| \leq |x| \leq 1/|h_i|} x^2 + \boldsymbol{\mu}^T \mathbf{w}_i(x), \quad i = 1, \dots, p-1, \\ \gamma_p^o(\boldsymbol{\mu}) &= \min_{|x| \leq 1/|h_p|} x^2 + \boldsymbol{\mu}^T \mathbf{h}x. \end{aligned}$$

Figure 6.4 shows the solutions obtained using (PXI) and AST for ρ as in (6.31) with $r = 1$. We omit the results for MUSIC in this plot as its performance is similar to AST. Notice that since (PXI) takes the the nonlinear nature of the signal into account it provides more precise parameter estimates. This is evident in Figure 6.5, which shows that (PXI) leads to lower reconstruction errors, especially

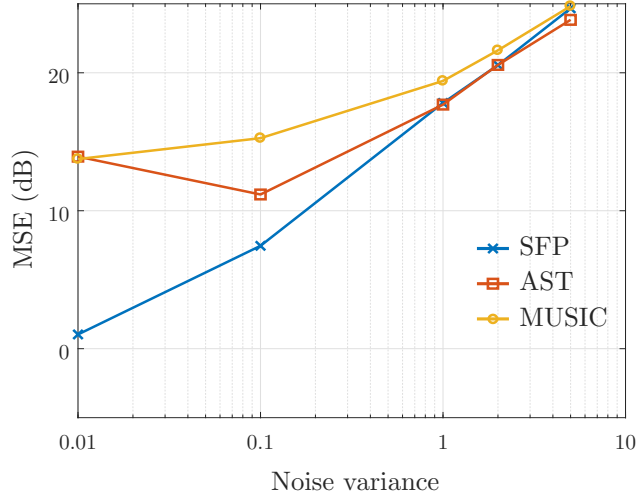


Figure 6.5: Reconstruction MSE for line spectral estimation of saturated sources.

in higher SNRs. This is expected since neither AST nor MUSIC take the nonlinear effects into account. Yet, as the noise increases and begins to dominate over mismodeling, the performance of all methods becomes similar. This effect is more pronounced here than in the linear case because the saturation limits the energy of the signal leading to even lower effective SNRs. For instance, the average SNR for $\sigma_n^2 = 2$ in Figure 6.2 is 6.6 dB, whereas in Figure 6.5, it is 2.05 dB.

In these experiments, the signal samples were constructed as in the linear case, but we used for (PXI) $B = 200$, $\epsilon = p\sigma_n^2$, and $\lambda = 100$ for all noise levels except $\sigma_n^2 \in \{2, 5\}$ which used $\lambda = 80$. For the AST method, we again used the optimal regularization parameter from [107]. Better results could not be obtained by hand-tuning the regularization.

6.4.2 Robust functional data analysis

Functional data analysis extends classical statistical methods to data supported on continuous domains. Since it copes with non-uniformly sampled data and precludes registration, this tool set is especially appropriated for analyzing time series without assuming generative models, such as AR or ARMAX [156]. For concreteness, consider the functional extension of logistic regression: given a data pair (y_i, Z_i) with label $y_i \in \{0, 1\}$ and independent variable $Z_i : [0, 1] \rightarrow \mathbb{R}$, the probability

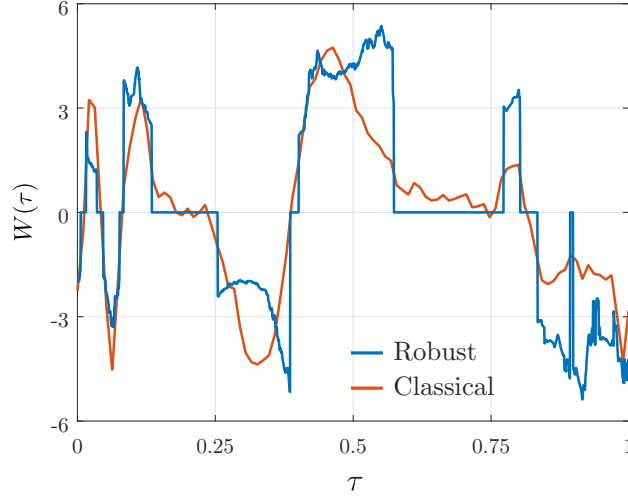


Figure 6.6: Solution of functional logistic regression for ECG classification.

that y_i is positive is modeled as

$$\Pr[y_i = 1] = \frac{1}{1 + \exp\left(-\int_0^1 Z_i(\tau)W(\tau)d\tau + b\right)}, \quad (6.32)$$

where $W : [0, 1] \rightarrow \mathbb{R}$ is the functional classifier parameter and b is the intercept. Although the domain of Z_i and W can be an arbitrary compact set, we use the normalized $[0, 1]$ for simplicity. Typically, some smoothness prior is assumed for W so that the statistical problem is well-posed, e.g., by using splines or imposing that W has small RKHS norm [156]. Observe that if we replace $\int_0^1 Z_i(\tau)W(\tau)d\tau$ by $\mathbf{w}^T \mathbf{z}_i$, for $\mathbf{w}, \mathbf{z}_i \in \mathbb{R}^m$, we recover the classical, finite dimensional logistic model.

As is the case with traditional (discrete) logistic regression, the classifier in (6.32) is sensitive to outliers. In fact, it has been shown recently that any classifier trained by minimizing a convex loss, as is the case of logistic regression or support vector machines (SVM), suffers from this issue [157, 158]. Although sparsity has been used to mitigate this drawback using convex surrogates such as the ℓ_1 -norm [159, 160], these methods remain susceptible to extreme data points caused by impulsive noise or other measurement errors [158].

One approach to addressing this weakness is replacing the inner product in (6.32) by a robust version that reduces the influence of these extreme samples. In [157, 158], this is done by computing

inner products over a subset of the data. Here, however, since (P-SFP) allows us to consider arbitrary nonlinearities in the data model, we can explicitly limit the influence of any sample by saturating the inner product in (6.32). Explicitly,

$$\Pr[y_i = 1] = \frac{1}{1 + \exp\left(-\int_0^1 \rho[Z_i(\tau)W(\tau)] d\tau + b\right)}, \quad (6.33)$$

where ρ is the saturation from (6.31). Notice that (6.33) controls the influence of any data point by using the threshold r from the saturation (6.31). In fact, notice that due to the saturation, the value of the inner product in (6.33) lies in the range $[-r, r]$. Using the negative log likelihood expression for logistic regression [152], we then formulate the following SFP for learning the robust classifier

$$\begin{aligned} & \underset{W \in L_2}{\text{minimize}} && \|W\|_{L_2} + b^2 + \lambda \|W\|_{L_0} \\ & \text{subject to} && -\sum_{i=1}^p \log[1 + \exp((1 - 2y_i)\hat{y}_i)] \leq \epsilon \\ & && \hat{y}_i = \int_{\mathcal{T}} \rho[Z_i(\beta)W(\beta)] d\beta + b, \\ & && \text{for } i = 1, \dots, p, \end{aligned} \quad (\text{PXII})$$

for some fit parameter $\epsilon > 0$. Notice that (PXII) also allows us to fit sparse functional coefficient W by setting $\lambda > 0$. Moreover, although it is written in terms of the logistic likelihood, other convex criteria such as the hinge loss could be used to obtain robust SVMs.

To illustrate the performance of the robust classifier (6.33), we consider the problem of identifying whether an electrocardiogram (ECG) signal comes from a healthy heart or one that suffered a myocardial infarction, i.e., a heart attack. The continuous time series Z_i are obtained by linearly interpolating a single heartbeat (see examples in Figure 6.8). Other techniques, such as sinc or spline interpolation are also commonly used in functional data analysis [156]. The labels y_i indicate whether a heart is healthy (1) or not (0). The samples used in the following experiments were taken from the ECG200 dataset [161], which draws from the MIT-BIH Supraventricular Arrhythmia Database [162]. To train the classical functional logistic classifier in (6.32), we solved (PXII) with $\lambda = 0$ and $r \rightarrow \infty$, i.e., no sparsity regularization and no saturation of the inner product. For the robust version, we used $\lambda = 10$ and $r = 4$. In both cases, the classifier was fitted with $\epsilon = -46$ using the approximate

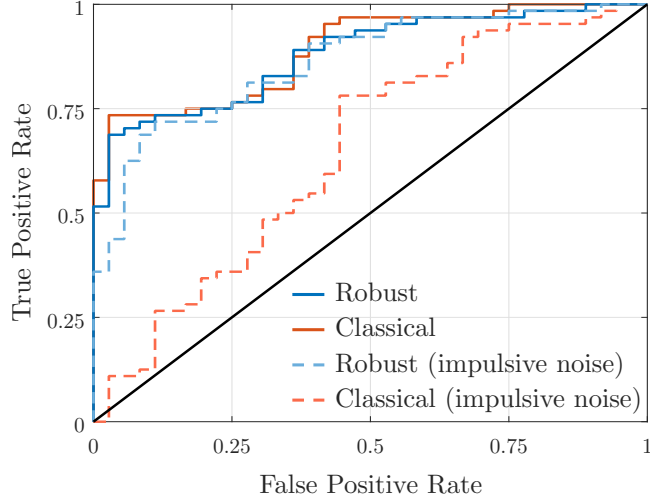


Figure 6.7: Receiver operating characteristic (ROC) curve for logistic classifiers in the presence of impulsive noise.

supergradient method described in Appendix B.1.2.

Notice in Figure 6.6 that the value of the coefficients of the classical and robust classifiers are similar, leading to comparable performance on both training and test sets (approximately 80% accuracy). The receiver operating characteristic (ROC) curve of both classifiers on the test set is displayed in solid lines in Figure 6.7. The robustness of these classifiers to outliers, on the other hand, is considerably different. To illustrate this behavior, corruption by impulsive noise was simulated by randomly adding ± 20 to a random subset of 10% of the samples from each heartbeat in the test set. The resulting ROC curves are shown in dashed lines. Although the performance of the linear logistic classifier has now degraded (the test accuracy dropped to 66%), the ROC of the robust version remains unaltered due to the nonlinearity ρ in (6.33) limiting the effect of the corruption (test accuracy of 76%).

Additionally, the sparsity of the robust classifier parameters improves interpretability by focusing on the portions of the signal that differentiate between normal and abnormal heartbeats (Figure 6.8). For instance, healthy heart signals tend to have negative values for $\tau \in [0.25, 0.4]$ and positive values for $\tau \in [0.4, 0.6]$, whereas hearts that suffered myocardial infarctions do not. On the other hand, there is no discriminant information for $\tau \in [0.6, 0.75]$ and, perhaps less intuitively, between 0.15 and 0.25.

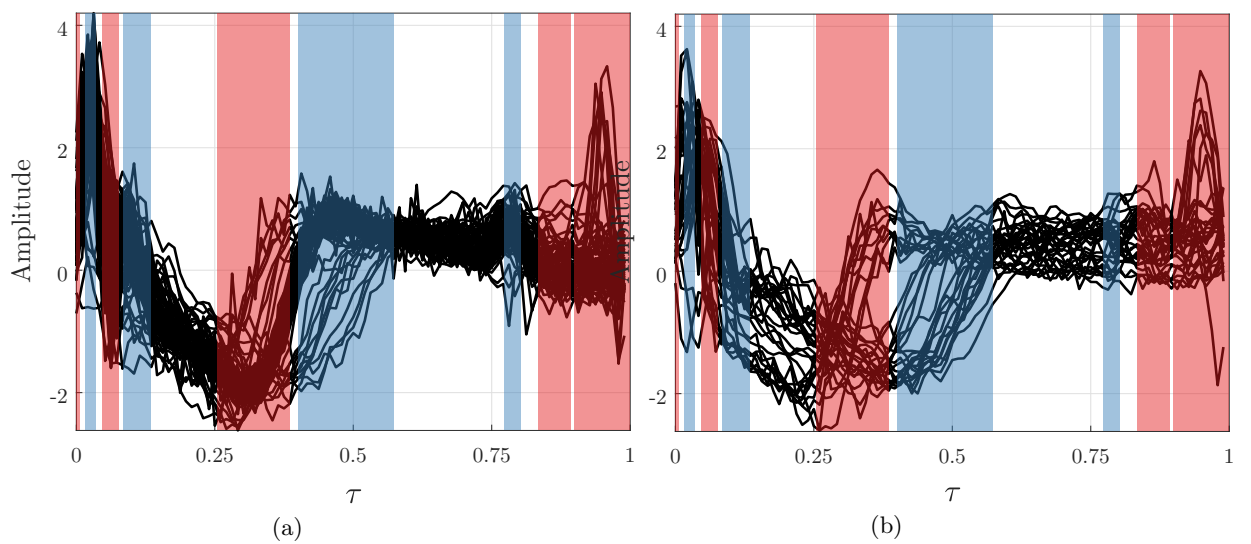


Figure 6.8: ECG and sparse functional coefficients (positive coefficients: blue, negative coefficients: red): (a) healthy heart and (b) heart with myocardial infarction.

Chapter 7

Combinatorial constraints and nonsubmodular optimization

Whereas the SFPs from Chapter 6 turned out to be “easy” to solve (in the sense that they are in P), statistical problems with combinatorial constraints are not. In fact, most of the problems we study in this chapter are well-known NP-hard problems [99, 163, 164]. Our goal here is therefore to efficiently obtain near-optimal approximate solutions, i.e., to derive algorithms that output solutions whose values are within a factor of the optimal value. Typically, the term “near-optimal” is reserved for the case in which that factor is a constant independent of the problem instance (universal). Unfortunately, the lack of structure of the problems we tackle make this hard and in certain cases impossible. It is now known that it is NP-hard to approximate within a constant factor certain quadratic problems with combinatorial constraints (e.g., sensor selection [163]). The best we can do in this case is to provide guarantees that depend only on problem parameters known *a priori*.

To be more specific, we are interested in solving quadratic statistical optimization problems subject to combinatorial constraints. Such problems arise in applications where access to the full measurement \mathbf{y} is impractical or impossible. The issue then becomes that of selecting a subset of observations based on which to perform inference within the limitations of the problem. For instance, power or sensing constraints may limit the number of observations allowed (cardinality constraint) or disallow certain combinations of observations due to, e.g., duty cycle limitations (matroid con-

straints). These problems arise in applications such as sensor selection, experimental design, and scheduling.

A common way to select measurements is based on greedy heuristics in which sensors or experiments are chosen one at a time by selecting the one that most reduces the cost at each step. A classical result from discrete optimization shows that if the cost function displays a certain diminishing returns property known as supermodularity, then greedy procedures are near-optimal: they achieve a value that is at within 63% of the optimal one. The cost functions in many of these problems, however, are quadratic and are well-known to not be supermodular (e.g., the estimation MSE or a quadratic control cost as in the LQR or LQG problems [164–167]). Nevertheless, rather than relying on surrogates and relaxations, we show that greedy methods can still obtain approximate solutions to these problems by building a theory of approximate supermodularity.

7.1 (Approximate) supermodularity

Supermodularity (*submodularity*) encodes a “diminishing returns” property of certain set functions that implies near-optimality bounds on their greedy minimization (maximization). Well-known representatives of this class include the rank or logdet of a sum of PSD matrices, the Shannon entropy, and the mutual information [166, 168]. Still, supermodularity is a stringent condition. In particular, it does not hold for typical the quadratic cost functions such as the MSE [164–167].

The purpose of *approximate supermodularity* (*submodularity*) is to relax the original “diminishing returns” property while controlling the magnitude of the violations. The rationale is that if a function is “almost” supermodular, then it should behave similar to a supermodular function. In what follows, we formalize and quantify these statements.

Consider a set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ whose value for an arbitrary set $\mathcal{X} \subseteq \mathcal{V}$ is denoted by $f(\mathcal{X})$. We say f is *normalized* if $f(\emptyset) = 0$ and f is *monotone decreasing* if for all sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ it holds that $f(\mathcal{A}) \geq f(\mathcal{B})$. Observe that if a function is normalized and monotone decreasing it must be that $f(\mathcal{X}) \leq 0$ for all $\mathcal{X} \subseteq \mathcal{V}$. Define

$$\Delta_u f(\mathcal{X}) = f(\mathcal{X}) - f(\mathcal{X} \cup \{u\}) \tag{7.1}$$

to be the variation in the value of f incurred by adding the element $u \in \mathcal{V} \setminus \mathcal{X}$ to the set \mathcal{X} . Then, a set function f is *supermodular* if for all sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and elements $u \in \mathcal{V} \setminus \mathcal{B}$ it holds that

$$\Delta_u f(\mathcal{A}) \geq \Delta_u f(\mathcal{B}). \quad (7.2)$$

A function f is *submodular* if $-f$ is supermodular.

The relevance of supermodular functions in this work is due to the celebrated bound on the suboptimality of their greedy minimization [169]. Specifically, consider the generic cardinality constrained optimization problem

$$\mathcal{X}^* \in \operatorname{argmin}_{|\mathcal{X}| \leq s} f(\mathcal{X}),$$

and construct its greedy solution by starting with $\mathcal{G}_0 = \emptyset$ and incorporating the elements from \mathcal{V} one at a time so as to maximize the gain at each step. Explicitly, at step j we do

$$\begin{aligned} \mathcal{G}_{j+1} &= \mathcal{G}_j \cup \{u\}, \\ \text{with } u &= \operatorname{argmin}_{w \in \mathcal{V} \setminus \mathcal{G}_j} f(\mathcal{G}_j \cup \{w\}). \end{aligned} \quad (7.3)$$

The recursion in (7.3) is repeated for s steps to obtain a greedy solution with s elements. If f is monotone decreasing and supermodular [169], then

$$f(\mathcal{G}_s) \leq (1 - e^{-1})f(\mathcal{X}^*). \quad (7.4)$$

The guarantee in (7.4), however, no longer applies when f is not supermodular. To provide guarantees in these cases, we leverage two measures of approximate supermodularity and derive near-optimality bounds for each of them. It is worth noting that though intuitive, such results are not trivial. In fact, [170] showed that for another measure of proximity, functions δ -close to supermodular cannot be optimized in polynomial time unless δ is small.

We start with a multiplicative relaxation of the supermodular property (7.2).

Definition 5 (α -supermodularity). A set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ is α -supermodular, for $\alpha \in \mathbb{R}$, if for

all sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and all $u \in \mathcal{V} \setminus \mathcal{B}$ it holds that

$$\Delta_u f(\mathcal{A}) \geq \alpha \Delta_u f(\mathcal{B}). \quad (7.5)$$

For $\alpha \geq 1$, (7.5) reduces the original definition of supermodularity (7.2), in which case we refer to the function simply as supermodular [166, 168]. On the other hand, when $\alpha < 1$, f is said to be *approximately supermodular*. Notice that if f is decreasing, then (7.5) always holds for $\alpha = 0$. We are therefore interested in the largest α for which (7.5) holds, i.e.,

$$\alpha = \min_{\substack{\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \\ u \in \mathcal{V} \setminus \mathcal{B}}} \frac{\Delta_u f(\mathcal{A})}{\Delta_u f(\mathcal{B})} \quad (7.6)$$

This concept first appeared in the context of auction design [171], although it has been rediscovered in the context of discrete optimization, estimation, and control [45, 55, 172, 173]. It is worth noting that α in Definition 5 is also related to the *supermodularity ratio*, a relaxation based on a different, though equivalent, definition of supermodularity. Explicitly, the supermodularity ratio is defined as the largest γ for which

$$\sum_{w \in \mathcal{W}} \Delta_w f(\mathcal{L}) \geq \gamma [f(\mathcal{L}) - f(\mathcal{L} \cup \mathcal{W})], \quad (7.7)$$

for all disjoint sets $\mathcal{W}, \mathcal{L} \subseteq \mathcal{E}$. It was introduced in [174]¹ in the context of variable selection, although the resulting guarantees depended on sparse eigenvalues that are NP-hard to compute. Explicit (P-computable) lower bounds on γ were derived in [173, 175, 176], although it is worth noting that the bounds on α obtained earlier in [45, 55] also hold for the supermodularity ratio:

Proposition 9. *Let f be an α -supermodular and denote its supermodularity ratio by γ . Then, $\alpha \leq \gamma$.*

Proof. We proceed by showing that (7.7) holds with $\gamma = \alpha$ for any α -supermodular function. To do

¹Although [174] and subsequent literature define γ in terms of submodularity, it can be recovered by taking $-f$ in (7.7). Recall that if f is supermodular, then $-f$ is submodular.

so, consider an enumeration of $\mathcal{W} = \{w_1, \dots, w_{|\mathcal{W}|}\}$ and write

$$f(\mathcal{L}) - f(\mathcal{L} \cup \mathcal{W}) = \sum_{k=1}^{|\mathcal{W}|} \Delta_{w_k} f(\mathcal{L} \cup \{w_1, \dots, w_{k-1}\}). \quad (7.8)$$

Since f is α -supermodular, we can upper bound each of the increments in (7.8) using (7.5) to obtain

$$f(\mathcal{L}) - f(\mathcal{L} \cup \mathcal{W}) \leq \alpha^{-1} \sum_{k=1}^{|\mathcal{W}|} \Delta_{w_k} f(\mathcal{L}). \quad (7.9)$$

Comparing (7.7) and (7.9) concludes the proof. ■

Although α -supermodularity yields a multiplicative approximation factor, finding meaningful bounds on α can be challenging for certain set functions. It is therefore useful to look at approximate supermodularity from a different perspective, as proposed in [177].

Definition 6 (ϵ -supermodularity). A set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ is ϵ -supermodular, for $\epsilon \in \mathbb{R}$, if for all multisets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and all $u \in \mathcal{V} \setminus \mathcal{B}$ it holds that

$$\Delta_u f(\mathcal{A}) \geq \Delta_u f(\mathcal{B}) - \epsilon. \quad (7.10)$$

Again, we say f is supermodular if $\epsilon \leq 0$ and approximately supermodular otherwise. As with α , we want the best ϵ that satisfies (7.10), which is given by

$$\epsilon = \max_{\substack{\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \\ u \in \mathcal{V} \setminus \mathcal{B}}} \Delta_u f(\mathcal{B}) - \Delta_u f(\mathcal{A}). \quad (7.11)$$

A useful lemma to reduce problems involving approximately supermodular functions shows that conic combinations preserve approximate supermodularity.

Lemma 3. Consider the set functions $f_i : 2^{\mathcal{V}} \rightarrow \mathbb{R}$, $i \in \mathbb{N}$. Then, for $\theta_i \geq 0$ and $b \in \mathbb{R}$,

- (i) if the f_i are α_i -supermodular, then $g = \sum_i \theta_i f_i + b$ is $\min(\alpha_i)$ -supermodular;
- (ii) if the f_i are ϵ_i -supermodular, then $g = \sum_i \theta_i f_i + b$ is $(\sum_i \theta_i \epsilon_i)$ -supermodular.

Proof. We proceed with the proof case-by-case:

(i) From the definition of α -supermodularity in (7.5), for $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $u \in \mathcal{V} \setminus \mathcal{B}$ it holds that

$$g(\mathcal{A}) - g(\mathcal{A} \cup \{u\}) = \sum_i \theta_i [f_i(\mathcal{A}) - f_i(\mathcal{A} \cup \{u\})] \geq \sum_i \alpha_i \theta_i [f_i(\mathcal{B}) - f_i(\mathcal{B} \cup \{u\})],$$

where the constant factors of the affine transformations cancel out. Since $\alpha_i \geq \min(\alpha_i)$, we obtain

$$g(\mathcal{A}) - g(\mathcal{A} \cup \{u\}) \geq \min(\alpha_i) \sum_i \theta_i [f_i(\mathcal{B}) - f_i(\mathcal{B} \cup \{u\})] \geq \min(\alpha_i) [g(\mathcal{B}) - g(\mathcal{B} \cup \{u\})].$$

(ii) Similarly, using the definition of ϵ -supermodularity in (7.10), for $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $u \in \mathcal{V} \setminus \mathcal{B}$ we obtain

$$g(\mathcal{A}) - g(\mathcal{A} \cup \{u\}) \geq \sum_i \theta_i [f_i(\mathcal{B}) - f_i(\mathcal{B} \cup \{u\})] - \sum_i \theta_i \epsilon_i = g(\mathcal{B}) - g(\mathcal{B} \cup \{u\}) - \sum_i \theta_i \epsilon_i,$$

where again the constant factors b canceled out. ■

In many cases, it is hard or even impossible [163] to obtain an algorithm to minimize approximately supermodular functions whose near-optimality is independent of α or ϵ . In other words, these quantities not only quantify violations of the diminishing returns property, but also the loss in optimality due to those violations. It is therefore paramount to obtain bounds on α and ϵ that are computable. One particular case of interest deals with scalarizations of a particular set matrix function that commonly appears in the least square problems, which we address next.

7.1.1 Approximately supermodular scalarizations

Consider a set of PSD matrices indexed by the elements of the ground set \mathcal{V} , i.e., $\mathbf{M}_u \succeq 0$ for $u \in \mathcal{V}$, together with the PD matrix \mathbf{M}_\emptyset . We are interested in analyzing the approximate supermodularity of scalarizations of set functions $\mathbf{Y} : 2^\mathcal{V} \rightarrow \mathbb{S}_+$ of the form

$$\mathbf{Y}(\mathcal{X}) = \left(\mathbf{M}_\emptyset + \sum_{u \in \mathcal{X}} \mathbf{M}_u \right)^{-1}. \quad (7.12)$$

In particular, we are interested in studying its trace

$$t(\mathcal{X}) = \text{Tr}[\mathbf{Y}(\mathcal{X})] \quad (7.13)$$

and its spectral norm

$$e(\mathcal{X}) = \|\mathbf{Y}(\mathcal{X})\|. \quad (7.14)$$

In estimation problems, for instance, \mathbf{M}_\emptyset denotes an *a priori* error covariance matrix that represents the error in the absence of measurements and \mathbf{M}_u carries the information obtained by taking measurement $u \in \mathcal{V}$. Then, the trace in (7.13) is related to the estimation MSE and the spectral norm in (7.14) is its robust counterpart, i.e., the worst-case error.

Before analyzing the α/ϵ -supermodularity of (7.13) and (7.14), we show that both are monotone decreasing set functions by proving that $\mathbf{Y}(\mathcal{X})$ is a decreasing set function in the PSD cone (Lemma 4). The definition of Loewner order and the monotonicity of the trace [178] imply the desired result.

Lemma 4. *The matrix-valued set function $\mathbf{Y}(\mathcal{X})$ in (7.12) is monotonically decreasing with respect to the PSD cone, i.e., $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \Rightarrow \mathbf{Y}(\mathcal{A}) \succeq \mathbf{Y}(\mathcal{B})$.*

Proof. The monotone decreasing nature of \mathbf{Y} stems from the fact that matrix inversion is an operator antitone function, i.e., that for $\mathbf{X}, \mathbf{Z} \succeq 0$, it holds that $\mathbf{X} \succeq \mathbf{Z} \Leftrightarrow \mathbf{X}^{-1} \preceq \mathbf{Z}^{-1}$ [179]. To see this, write (7.12) as $\mathbf{Y}(\mathcal{X}) = \mathbf{R}(\mathcal{X})^{-1}$, where $\mathbf{R}(\mathcal{X}) = \mathbf{M}_\emptyset + \sum_{u \in \mathcal{X}} \mathbf{M}_u$. Then, notice that since \mathbf{R} is a sum of PSD matrices, it holds that $\mathbf{R}(\mathcal{A}) \succeq 0$ for all $\mathcal{A} \subseteq \mathcal{V}$. Moreover, \mathbf{R} is a modular (additive) function, i.e., $\mathbf{R}(\mathcal{A} \cup \mathcal{B}) = \mathbf{R}(\mathcal{A}) + \mathbf{R}(\mathcal{B})$. Hence, for $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$, we obtain

$$\mathbf{R}(\mathcal{B}) = \mathbf{R}(\mathcal{A}) + \mathbf{R}(\mathcal{B} \setminus \mathcal{A}) \succeq \mathbf{R}(\mathcal{A}). \quad (7.15)$$

It is straightforward to obtain from (7.15) that for $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$

$$\mathbf{R}(\mathcal{A}) \preceq \mathbf{R}(\mathcal{B}) \Leftrightarrow \mathbf{R}(\mathcal{A})^{-1} \succeq \mathbf{R}(\mathcal{B})^{-1} \Leftrightarrow \mathbf{Y}(\mathcal{A}) \succeq \mathbf{Y}(\mathcal{B}). \quad \blacksquare$$

The remainder of this section is dedicated to showing that the trace and the spectral norm are approximately supermodular scalarizations of \mathbf{Y} , starting with the set trace function (7.13).

Theorem 6. *Let t be the set function in (7.13). Then, t is α -supermodular with*

$$\alpha \geq \frac{\mu_{\min}}{\mu_{\max}} > 0, \quad (7.16)$$

where

$$0 < \mu_{\min} \leq \lambda_{\min} [\mathbf{M}_{\emptyset}] \leq \lambda_{\max} \left[\mathbf{M}_{\emptyset} + \sum_{u \in \mathcal{V}} \mathbf{M}_u \right] \leq \mu_{\max}.$$

Remark 1. Although there exist examples for which $t(\mathcal{X})$ is not supermodular, the general statement of Theorem 6 does not allow us to claim that $\alpha < 1$. A simple counter-example involves the case in which $\mathbf{M}_{\emptyset} = \mu_0 \mathbf{I}$ and $\mathbf{M}_u = \mu_u \mathbf{I}$, $\mu_0, \mu_u \geq 0$, so that t becomes effectively a scalar function of μ_0, μ_u . Since scalar convex functions of positive modular functions are supermodular [180], we have $\alpha \geq 1$ in this case.

The proof of Theorem 6 relies on the following bounds on the variation $\Delta_u t(\mathcal{X})$:

Lemma 5. *For all $\mathcal{X} \subset \mathcal{V}$ and $u \in \mathcal{V} \setminus \mathcal{X}$, it holds that*

$$\lambda_{\min} [\mathbf{Y}(\mathcal{X})] \operatorname{Tr} [\mathbf{M}_u \mathbf{Y}(\mathcal{X} \cup \{u\})] \leq \Delta_u t(\mathcal{X}) \leq \lambda_{\max} [\mathbf{Y}(\mathcal{X})] \operatorname{Tr} [\mathbf{M}_u \mathbf{Y}(\mathcal{X} \cup \{u\})]. \quad (7.17)$$

Proof. See appendix B.2.1. ■

Theorem 6 then follows readily.

Proof of Theorem 6. Notice from (7.1) and (7.6) that α can be written as

$$\alpha = \min_{\substack{\mathcal{A} \subset \mathcal{B} \subseteq \mathcal{V} \\ u \in \mathcal{V} \setminus \mathcal{B}}} \frac{\Delta_u t(\mathcal{A})}{\Delta_u t(\mathcal{B})}.$$

From Lemma 5, we then obtain

$$\alpha \geq \frac{\lambda_{\min} [\mathbf{Y}(\mathcal{A})]}{\lambda_{\max} [\mathbf{Y}(\mathcal{B})]} \times \frac{\operatorname{Tr} [\mathbf{M}_u \mathbf{Y}(\mathcal{A} \cup \{u\})]}{\operatorname{Tr} [\mathbf{M}_u \mathbf{Y}(\mathcal{B} \cup \{u\})]}. \quad (7.18)$$

To simplify (7.18), recall from Lemma 4 that \mathbf{Y} is monotone decreasing in the PSD cone. Since $\mathcal{A} \subset \mathcal{B} \subseteq \mathcal{V}$, it holds that $\mathbf{Y}(\mathcal{A} \cup \{u\}) \succeq \mathbf{Y}(\mathcal{B} \cup \{u\})$. The ordering of the PSD cone (Loewner order)

gives us that the second term in (7.18) is always greater than one, which yields

$$\alpha \geq \frac{\lambda_{\min}[\mathbf{Y}(\mathcal{A})]}{\lambda_{\max}[\mathbf{Y}(\mathcal{B})]}.$$

The lower bound in (7.16) is readily obtained by observing that the decreasing nature of \mathbf{Y} implies that for any set $\mathcal{X} \subseteq \mathcal{V}$:

$$\lambda_{\min}[\mathbf{Y}(\mathcal{V})] \leq \lambda_{\min}[\mathbf{Y}(\mathcal{X})] \leq \lambda_{\max}[\mathbf{Y}(\mathcal{X})] \leq \lambda_{\max}[\mathbf{Y}(\emptyset)]. \quad \blacksquare$$

Theorem 6 gives a deceptively simple bound on the α -supermodularity of the set function t in (7.13) depending on the spectrum of the $\mathbf{M}_\emptyset, \mathbf{M}_i$. This bound can be interpreted geometrically in terms of the “range” of the error covariance matrix \mathbf{Y} . To see this, define the *numerical range* of the set function \mathbf{Y} as

$$W_{\mathcal{V}}(\mathbf{Y}) = W \left[\bigoplus_{\mathcal{X} \subseteq \mathcal{V}} \mathbf{Y}(\mathcal{X}) \right], \quad (7.19)$$

where $\mathbf{A} \oplus \mathbf{B} = \text{blkdiag}(\mathbf{A}, \mathbf{B})$ is the direct sum of matrices \mathbf{A} and \mathbf{B} and $W(\mathbf{M}) = \{\mathbf{x}^T \mathbf{M} \mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$ is the classical numerical range [178]. Since the numerical range is a convex set, we can define its relative diameter as

$$D = \max_{\mu, \eta \in W_{\mathcal{V}}(\mathbf{Y})} \left| \frac{\mu - \eta}{\mu} \right|. \quad (7.20)$$

Then, the following holds:

Proposition 10. *The set functions t in (7.13) is α -supermodular with*

$$\alpha \geq 1 - D,$$

where D is the relative diameter of the numerical range of \mathbf{Y} in (7.20).

Proof. Since $\mathbf{Y} \succ 0$, the numerical range in (7.19) is the bounded convex hull of the eigenvalues of $\mathbf{Y}(\mathcal{X})$ for all $\mathcal{X} \subseteq \mathcal{V}$ [178]. We can therefore simplify (7.20) using the fact that it is monotonically

increasing in μ and decreasing in η . Explicitly,

$$D = \max_{\mathcal{X}, \mathcal{Y} \subseteq \mathcal{V}} \left| \frac{\lambda_{\max}[\mathbf{Y}(\mathcal{Y})] - \lambda_{\min}[\mathbf{Y}(\mathcal{X})]}{\lambda_{\max}[\mathbf{Y}(\mathcal{Y})]} \right|.$$

Using the fact that \mathbf{Y} is monotonically decreasing (Lemma 4), this maximum is achieved for

$$D = \frac{\lambda_{\max}[\mathbf{Y}(\emptyset)] - \lambda_{\min}[\mathbf{Y}(\mathcal{V})]}{\lambda_{\max}[\mathbf{Y}(\emptyset)]}.$$

The bound in (7.16) thus becomes

$$\alpha \geq \frac{\lambda_{\min}[\mathbf{Y}(\mathcal{V})]}{\lambda_{\max}[\mathbf{Y}(\emptyset)]} = 1 - D. \quad \blacksquare$$

Hence, (7.16) bounds how much t deviates from a supermodular function (as quantified by α) in terms of the numerical range of its underlying function \mathbf{Y} . The shorter the range of \mathbf{Y} , the more supermodular-like (7.13) will be.

Proceeding, we now bound the ϵ -supermodularity of (7.14).

Theorem 7. *Let e be defined as in (7.14). Then, e is ϵ -supermodular with*

$$\epsilon \leq \frac{\lambda_{\max}(\sum_{u \in \mathcal{V}} \mathbf{M}_u)}{\lambda_{\min}[\mathbf{M}_{\emptyset}]^2}. \quad (7.21)$$

Proof. See appendix B.2.2. ■

Having established explicit bounds on the approximate supermodularity of these abstract set functions, we proceed to derive approximation guarantees for their greedy minimization under different constraints. In Section 7.4, we go back to the results of this section to obtain explicit guarantees for the greedy solution of a variety of problems, from experimental design to actuator scheduling.

7.2 Approximately supermodular minimization

Minimizing monotonically decreasing set functions is, in a sense, trivial: their minimum is achieved for the ground set \mathcal{V} . The difficulty in solving these problems arises, not from their discrete structure,

Algorithm 3 Greedy algorithm for (P-CARD)

```
Let  $\mathcal{S}_0 \leftarrow \emptyset$  and  $t \leftarrow 0$   
while  $|\mathcal{S}_t| < r$  do  
   $g \leftarrow \operatorname{argmin}_{u \in \mathcal{V}} f(\mathcal{X}_t \cup \{u\})$   
   $\mathcal{X}_{t+1} \leftarrow \mathcal{X}_t \cup \{g\}$   
   $t \leftarrow t + 1$   
end while  
 $\mathcal{X}_g \leftarrow \mathcal{X}_t$ 
```

but their combinatorial constraints. Indeed, while unconstrained set function minimization is trivial, constrained set function minimization is often NP-hard. In this section we derive guarantees for two types of constraints: cardinality and intersection of matroids. It is worth noting that while cardinality constraints are matroid constraints, their structure is amenable to better guarantees, which is why we study them separately. What we show next is that not only do α and ϵ in (7.5) and (7.10) measure violations of supermodularity, but they also quantify the loss in optimality due to those violations.

7.2.1 Cardinality constraints

We begin by providing guarantees for cardinality constrained problems. Explicitly, we are interested in problems of the form

$$\begin{aligned} \mathcal{X}^* \in \operatorname{argmin}_{\mathcal{X} \subseteq \mathcal{V}} f(\mathcal{X}) \\ \text{subject to } |\mathcal{X}| \leq s, \end{aligned} \tag{P-CARD}$$

where $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}_+$ is normalized, i.e., $f(\emptyset) = 0$, and $0 < s \leq |\mathcal{V}|$ indicates the maximum cardinality of the solution \mathcal{X} . In particular, we are interested in the performance greedy solutions \mathcal{X}_g of (P-CARD) obtained using Algorithm 3. Note that Algorithm 3 yields solutions of cardinality r possibly larger than s . In doing so, we characterize the gain in performance obtained by violating the constraint of (P-CARD).

We begin by consider the case in which f in (P-CARD) is α -supermodular.

Theorem 8. *Let f be a normalized, monotone decreasing, and α -supermodular set function (i.e., $f(\emptyset) = 0$ and $f(\mathcal{X}) \leq 0$ for all $\mathcal{X} \subseteq \mathcal{V}$). Then, the solution obtained by the greedy search in*

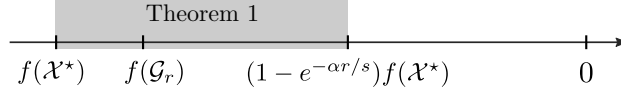


Figure 7.1: Illustration of the near-optimal guarantee from Theorem 8.

Algorithm 3 satisfies

$$f(\mathcal{X}_g) \leq (1 - e^{-\alpha r/s})f(\mathcal{X}^*). \quad (7.22)$$

Proof. See appendix B.2.3. ■

Theorem 8 bounds the suboptimality of the greedy solution of (P-CARD) when its objective is α -supermodular. Indeed, since f is a non-positive function, it guarantees that $f(\mathcal{X}_g)$ cannot be too far from the optimal value (Figure 7.1). At the same time, it quantifies the effect of relaxing the supermodularity hypothesis typically used to provide performance guarantees in these settings. In fact, if f is supermodular ($\alpha = 1$) and $r = s$, we recover the guarantee (7.4) from [169]. On the other hand, for an approximately supermodular function ($\alpha < 1$), the result in (7.22) shows that the same 63% guarantee is recovered by greedily selecting a set of size s/α . Hence, α not only measures how much f violates supermodularity, but also gives a factor by which a solution set must increase (violate the cardinality constraint) to maintain supermodular-like near-optimality. It is worth noting that, as with the original bound in [169], (7.22) is not tight and that better results are typically obtained in practice (see Section 7.4).

In contrast to α -supermodularity, we obtain an additive approximation guarantee for the greedy minimization of ϵ -supermodular functions.

Theorem 9. *Let f be a normalized, monotone decreasing, and ϵ -supermodular set function (i.e., $f(\emptyset) = 0$ and $f(\mathcal{X}) \leq 0$ for all $\mathcal{X} \subseteq \mathcal{V}$). Then, the solution obtained by the greedy search in Algorithm 3 satisfies*

$$f(\mathcal{X}_g) \leq (1 - e^{-r/s})[f(\mathcal{X}^*) + s \cdot \epsilon]. \quad (7.23)$$

Proof. See appendix B.3. ■

As before, ϵ quantifies the loss in performance guarantee due to relaxing supermodularity. Indeed, (7.23) reveals that ϵ -supermodular functions have the same guarantees as a supermodular function

up to an additive factor of $\Theta(s\epsilon)$. In fact, if $\epsilon \leq (es)^{-1}|f(\mathcal{X}^*)|$ (recall that $f(\mathcal{X}^*) \leq 0$ due to normalization), then taking $r = 3s$ recovers the supermodular 63% approximation factor. This same factor is obtained for $(\alpha \geq 1/3)$ -supermodular functions.

Notice that Theorems 8 and 9 characterize the loss in suboptimality incurred from violating supermodularity in terms of the values of α and ϵ . Contrary to classical supermodularity, these guarantees are therefore not universal. It worth noting, however, that there are cases in which this is the best approximation factor possible, i.e., in which obtaining universal approximation constants is impossible (if $P \neq NP$) [163].

7.2.2 Intersection of matroids constraints

More generally, we can consider the problem of optimizing a set function subject to matroid constraints. This generalizes the results from Section 7.2.1 in two ways. First, instead of considering cardinality constraints (i.e., a uniform matroid), we now allow the solution to be the independent set of any generic matroid. Second, we now consider the case in which the solution must satisfy multiple constraints. Explicitly, we consider the problem

$$\begin{aligned} \mathcal{X}^* \in \operatorname{argmin}_{\mathcal{S} \subseteq \mathcal{V}} \quad & f(\mathcal{X}) \\ \text{subject to} \quad & \mathcal{S} \in \mathcal{I}_p, \quad p = 1, \dots, P, \end{aligned} \tag{P-MTRD}$$

where the \mathcal{I}_p are the family of independent sets of a matroid, entities that extend the notion of linear independence in vector spaces to arbitrary algebraic structures. Formally,

Definition 7. A *matroid* $M = (\mathcal{V}, \mathcal{I})$ consists of a finite set of elements \mathcal{E} and a family $\mathcal{I} \subseteq 2^{\mathcal{V}}$ of subsets of \mathcal{V} called *independent sets* that satisfy:

(P1) $\emptyset \in \mathcal{I}$;

(P2) if $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $\mathcal{B} \in \mathcal{I}$, then $\mathcal{A} \in \mathcal{I}$;

(P3) if $\mathcal{A}, \mathcal{B} \in \mathcal{I}$ and $|\mathcal{A}| < |\mathcal{B}|$, then there exists $e \in \mathcal{B} \setminus \mathcal{A}$ such that $\mathcal{A} \cup \{e\} \in \mathcal{I}$.

Thus, the collection of valid schedules \mathcal{I} in (P-MTRD) is of the form

$$\mathcal{I} = \bigcap_{p=1}^P \mathcal{I}_p, \text{ such that } (\mathcal{E}, \mathcal{I}_p) \text{ is a matroid for all } p. \quad (7.24)$$

Matroids can be used to describe a myriad constraints common to scheduling problems, such as limiting the total number of actions, bounding the number of actions per time instant, or restrict the consecutive actions (duty-cycle constraints) (see Section 7.4). It is worth noting that matroids are not closed under intersections, so that \mathcal{I} need not be a matroid [181, Ch. 39].

The use of matroids as constraints in (P-MTRD) is attractive for two reasons. The first is a direct consequence of the inheritance property of matroids, i.e., P1 and P2 in Definition 7.

Proposition 11. *For each $\mathcal{A} \in \mathcal{I}$ with \mathcal{I} as in (7.24), there exists a chain $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \dots \subset \mathcal{A}_T = \mathcal{A}$ such that $\mathcal{A}_t \in \mathcal{I}$ for all t . In particular, there exists a chain with $T = |\mathcal{A}|$.*

Hence, every feasible set of (P-MTRD) can be constructed element-by-element. In particular, so can any optimal solution \mathcal{X}^* . This suggests an “interior point”-type algorithm that greedily minimizes the objective f at each step while maintaining feasibility, as described in Algorithm 4. The algorithm stops once no element can be added to \mathcal{X}_t without violating feasibility, i.e., when the argmin set in step 3 is empty. Denote that final iteration by T . Naturally, $T \leq |\mathcal{V}|$ so the algorithm does terminate. What is more, note from step 5 that not only is $\mathcal{X}_T \in \mathcal{I}$ by construction, but due to Proposition 11, every set $\mathcal{A} \in \mathcal{I}$ can be constructed by element-by-element as in Algorithm 4 for an appropriate choice of f (e.g., $f(\mathcal{X}) = |\mathcal{X} \cap \mathcal{A}|$). In other words, the greedy algorithm does not prune any solution from the feasibility set of (P-MTRD).

The second reason for using matroid intersections is related to the exchange property (P3 in Definition 7) and is laid out in the following proposition:

Proposition 12. *Let $\mathcal{A}, \mathcal{B} \in \mathcal{I}$ with \mathcal{I} as in (7.24). If $|\mathcal{B}| > P|\mathcal{A}|$, then there exist at least $|\mathcal{B}| - P|\mathcal{A}|$ elements b of $\mathcal{B} \setminus \mathcal{A}$ such that $\mathcal{A} \cup \{b\} \in \mathcal{I}$.*

Proof. The proof proceed by induction over the matroids in \mathcal{I} . The base case for first matroid ($p = 1$) is readily obtained from P3 in Definition 7. Indeed, take $e \in \mathcal{B} \setminus \mathcal{A}$ such that $\mathcal{A} \cup \{e\} \in \mathcal{I}_1$ and let $\mathcal{B}' = \mathcal{B} \setminus \{e\}$. This pruning process can be repeated as long as $|\mathcal{B}'| > |\mathcal{A}|$, i.e., at least $|\mathcal{B}| - |\mathcal{A}|$ times.

Algorithm 4 Greedy algorithm for (P-MTRD)

```

1: Let  $\mathcal{S}_0 \leftarrow \emptyset$ ,  $\mathcal{Z}_0 \leftarrow \mathcal{E}$ , and  $t \leftarrow 0$ 
2: while  $\mathcal{Z}_t \neq \emptyset$  do
3:    $g \leftarrow \operatorname{argmin}_{u \in \mathcal{Z}_t} V^*(\mathcal{S}_t \cup \{u\})$ 
4:    $\mathcal{Z}_{t+1} \leftarrow \mathcal{Z}_t \setminus \{g\}$ 
5:   if  $\mathcal{S}_t \cup \{g\} \in \cap_{p=1}^P \mathcal{I}_p$  then
6:      $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{g\}$ 
7:   else
8:      $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$ 
9:   end if
10:   $t \leftarrow t + 1$ 
11: end while
12:  $\mathcal{S}_g \leftarrow \mathcal{S}_t$ 

```

Now, suppose the claim holds for the first $P' - 1 < P$ independence sets, i.e., there exists a set $\mathcal{C} \subseteq \mathcal{B} \setminus \mathcal{A}$ such that

$$|\mathcal{C}| > |\mathcal{B}| - (P' - 1)|\mathcal{A}| \text{ and } \mathcal{A} \cup \{c\} \in \bigcap_{p=1}^{P'-1} \mathcal{I}_p \text{ for all } c \in \mathcal{C}. \quad (7.25)$$

Notice that $\mathcal{B} \in \mathcal{I} \Rightarrow \mathcal{B} \in \bigcap_{p=1}^{P'} \mathcal{I}_p$ and since $\mathcal{C} \subset \mathcal{B}$, the inheritance property of matroids (P2 in Definition 7) implies that $\mathcal{C} \in \bigcap_{p=1}^{P'} \mathcal{I}_p$ as well. In particular, since $\mathcal{C} \in \mathcal{I}_{P'}$, we again obtain from P3 in Definition 7 that there exist $\mathcal{C}' \subseteq \mathcal{C} \setminus \mathcal{A}$ such that

$$|\mathcal{C}'| > |\mathcal{C}| - |\mathcal{A}| \text{ and } \mathcal{A} \cup \{c\} \in \mathcal{I}_{P'} \text{ for all } c \in \mathcal{C}'. \quad (7.26)$$

Together, (7.25) and (7.26) yield that $|\mathcal{C}'| > |\mathcal{B}| - P'|\mathcal{A}|$ and that $\mathcal{A} \cup \{c\} \in \bigcap_{p=1}^{P'-1} \mathcal{I}_p \cap \mathcal{I}_{P'}$ for all $c \in \mathcal{C}'$. ■

Though more abstract, this property is fundamental to obtain near-optimal certificates for the greedy procedure in Algorithm 4. Indeed, the following noteworthy corollary is obtained for $\mathcal{B} = \mathcal{X}^*$ and $\mathcal{A} = \mathcal{X}_t$:

Corollary 3. *If $|\mathcal{X}^*| > Pt$, then there exist at least $|\mathcal{X}^*| - Pt$ elements $x \in \mathcal{X}^*$ such that $\mathcal{X}_t \cup \{x\} \in \mathcal{I}$. Hence, it must be that Algorithm 4 terminates after at least $T \geq |\mathcal{X}^*|/P$.*

At any given point, there are therefore at least $|\mathcal{X}^*| - Pt$ elements $x \in \mathcal{X}^*$ for which $f(\mathcal{X}_{t+1}) \leq f(\mathcal{X}_t \cup \{x\})$. When combined with some diminishing return property, this greedy property gives near-optimal guarantees as long as Algorithm 4 runs long enough. That is where the lower bound on T is useful. It stems directly from the fact that, if Algorithm 4 stops and returns a set with $|\mathcal{X}_T| < |\mathcal{X}^*|/P$, Proposition 12 implies there exists at least one element $x \in \mathcal{X}^*$ such that $\mathcal{X}_T \cup \{x\} \in \mathcal{I}$. This contradicts the fact that the feasibility set of step 3 must be empty for the algorithm to terminate.

Naturally, near-optimality depends not only on the feasibility set \mathcal{I} , but also on the objective f . For instance, if f is a monotone decreasing modular function, then \mathcal{X}_T is an optimal solution of (P-MTRD) with $P = 1$ in (7.24) [181, Ch. 40]; if f is a normalized, monotone decreasing supermodular function, then \mathcal{G}_T would be $1/(1+P)$ -optimal [182]. The case of α -supermodular objectives is addressed in the following theorem.

Theorem 10. *Consider (P-MTRD) where \mathcal{I}_p are independent sets of matroids and f is a normalized, monotone decreasing, α -supermodular function (i.e., $f(\mathcal{A}) \leq 0$ for all $\mathcal{A} \subseteq \mathcal{V}$). Let \mathcal{X}_g be its greedy solution from Algorithm 4. Then,*

$$f(\mathcal{X}_g) \leq \frac{\alpha}{\alpha + P} f(\mathcal{X}^*). \quad (7.27)$$

Proof. See appendix B.3.1. ■

Theorem 10 provides a near-optimal certificate for the greedy solution of α -supermodular minimization problems subject to multiple matroid constraints. Since the values of f are non-positive due to the normalization assumption, the result in (7.27) may not be straightforward to interpret. Nevertheless, it can be written equivalently in terms of improvements with respect to the empty solution, namely,

$$\frac{f(\mathcal{G}) - f(\mathcal{X}^*)}{f(\emptyset) - f(\mathcal{X}^*)} \leq \frac{P}{\alpha + P}.$$

As expected from previous results, the guarantee in (7.27) bounds the suboptimality of greedy search in terms of the α -supermodularity of the cost function. When $\alpha < 1$, (7.27) quantifies the loss in performance guarantee due to the objective f violating the diminishing returns property. Confirming our initial intuition, the larger the violations, i.e., the smaller α , the worst the guarantee.

On the other hand, (7.27) shows that the classical $1/(1+P)$ certificate for supermodular functions can be strengthened when f has stronger diminishing returns structures, i.e., when $\alpha \geq 1$. It is easy to see that Theorem 10 is consistent with previous results for single [169] and multiple [182] matroid constraints.

Observe, however, that (7.27) decreases linearly with P . In other words, the guarantees for greedy search deteriorates as more matroids are needed to represent \mathcal{I} . In a sense, this is the constraint counterpart of α -supermodularity: the further away from a pure matroid the constraints are, the worst the greedy algorithm is guaranteed to perform. Effectively, (7.27) states that the more constraints need to be satisfied, i.e., the larger P , the harder the problem becomes. Note that P in (7.27) can be replaced by the minimum number of matroids needed to represent $\bigcap_{p=1}^P \mathcal{I}_p$, so that the guarantee from Theorem 10 can sometimes be improved. Determining this minimum number, however, can be quite intricate. Still, Theorem 10 is a *worst case* bound and as is the case with the classical result for greedy supermodular minimization [169, 182], better performance is often obtained in practice. Still, pathological cases that approach the bound can be constructed (see Section 7.4).

It is worth noting that in the single matroid case ($P = 1$), Theorem 10 is strictly (though not significantly) stronger than the guarantee from [44] for $\alpha < 1$. This difference is more accentuated for small α : for instance, if $\alpha = 0.2$, (7.27) is 60% stronger than the certificate in [44]. A similar certificate was obtained in [171] with $\alpha \geq 1$ for a single matroid constraints.

7.3 The multiset case

In certain applications, we are interested in selecting multisets rather than sets. This is the case, for instance, in experimental design, where we may wish to allow an informative, yet noisy experiment to be performed multiple times thus allowing the noise to be averaged out. Directly applying the guarantees from the previous section in this setting can lead to extremely poor results due to the fact that they deteriorate the bounds on α/ϵ from Section 7.1.1. Indeed, since we would need to create a virtual ground set containing s unique copies of each element in \mathcal{V} , this can lead to guarantees that are six orders of magnitude lower than the ones put forward in this section.

To provide guarantees for multisets, we refine the definitions of approximate supermodularity from Section 7.1 to account for set cardinalities. Explicitly,

Definition 8 (α -supermodularity). A multiset function $f : \mathcal{P}(\mathcal{V}) \rightarrow \mathbb{R}$ is α -supermodular, for $\alpha : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$, if for all multisets $\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{V})$, $\mathcal{A} \subseteq \mathcal{B}$, and all $u \in \mathcal{V}$ it holds that

$$f(\mathcal{A}) - f(\mathcal{A} \cup \{u\}) \geq \alpha(|\mathcal{A}|, |\mathcal{B}|) [f(\mathcal{B}) - f(\mathcal{B} \cup \{u\})]. \quad (7.28)$$

Once again, α not only measures how much f violates supermodularity, but it also quantifies the loss in performance guarantee incurred from these violations.

Theorem 11. *Let f be a normalized, monotone decreasing, and α -supermodular multiset function. Then, for $\bar{\alpha} = \min_{a < r, b < r+s} \alpha(a, b)$ the greedy solution obtained from Algorithm (3) satisfies*

$$f(\mathcal{X}_g) \leq \left[1 - \prod_{t=0}^{r-1} \left(1 - \frac{1}{\sum_{k=0}^{s-1} \alpha(t, t+k)^{-1}} \right) \right] f(\mathcal{X}^*) \leq (1 - e^{-\bar{\alpha}r/s}) f(\mathcal{X}^*). \quad (7.29)$$

Proof. See appendix B.3.2. ■

We can proceed similarly for ϵ -supermodularity as shown in the following definition.

Definition 9 (ϵ -supermodularity). A multiset function $f : \mathcal{P}(\mathcal{V}) \rightarrow \mathbb{R}$ is ϵ -supermodular, for $\epsilon : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$, if for all multisets $\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{V})$, $\mathcal{A} \subseteq \mathcal{B}$, and all $u \in \mathcal{V}$ it holds that

$$f(\mathcal{A}) - f(\mathcal{A} \cup \{u\}) \geq f(\mathcal{B}) - f(\mathcal{B} \cup \{u\}) - \epsilon(|\mathcal{A}|, |\mathcal{B}|). \quad (7.30)$$

In contrast to α -supermodularity, we obtain an additive approximation guarantee for the greedy minimization of ϵ -supermodular functions.

Theorem 12. *Let f be a normalized, monotone decreasing, and ϵ -supermodular multiset function. Then, for $\bar{\epsilon} = \max_{a < r, b < r+s} \epsilon(a, b)$, the greedy solution obtained from Algorithm (3) satisfies*

$$\begin{aligned} f(\mathcal{X}_g) &\leq \left[1 - \left(1 - \frac{1}{s} \right)^r \right] f(\mathcal{X}^*) + \frac{1}{s} \sum_{k=0}^{s-1} \sum_{t=0}^{r-1} \epsilon(t, t+k) \left(1 - \frac{1}{s} \right)^{r-1-t} \\ &\leq (1 - e^{-r/s}) [f(\mathcal{X}^*) + s\bar{\epsilon}] \end{aligned} \quad (7.31)$$

Proof. See appendix B.3.3. ■

7.4 Applications

In this section, we illustrate the abstract near-optimality results presented so far in three applications. First, we explore the problem of downsampling graph signals, where cardinality constraints play a central role. We then proceed to analyze the control input scheduling problem, where more complex requirements involving bounds the number of actions per time instant or duty-cycle restrictions arise. These requirements can be described using an intersection of matroids. Finally, we study the experimental design problem to showcase our results on multiset optimization.

7.4.1 Graph Signal Sampling

A graph-supported signal, or *graph signal* for short, is an assignment of values to the nodes of a graph. Formally, let \mathbb{G} be a weighted graph with node set \mathcal{V} , having cardinality $|\mathcal{V}| = n$, and define a graph signal to be an injective mapping $\sigma : \mathcal{V} \rightarrow \mathbb{C}$. For an ordering of the nodes in \mathcal{V} , this signal can be represented as an $n \times 1$ vector that captures its values at each node:

$$\mathbf{x} = \begin{bmatrix} \sigma(u_1) & \cdots & \sigma(u_n) \end{bmatrix}^T, \quad u_i \in \mathcal{V}. \quad (7.32)$$

In what follows, we assume that the node ordering is fixed, so that we can index \mathbf{x} using elements of \mathcal{V} . For instance, we write $\mathbf{x}_{\{u_i, u_j, u_k\}} = [\sigma(u_i) \sigma(u_j) \sigma(u_k)]^T$.

Of interest to graph signal processing (GSP) is the spectral representation of the signal σ (or \mathbf{x}), which depends on the graph on which it is supported. Indeed, let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a matrix representation of \mathbb{G} . Usual choices include the adjacency matrix or one of the discrete Laplacians [183, 184]. Assume that \mathbf{A} is consistent with the signal vector (7.32) in the sense that they employ the same ordering of the nodes in \mathcal{V} . Furthermore, assume that \mathbf{A} is normal, i.e., that there exist $\mathbf{V} \in \mathbb{C}^{n \times n}$ unitary and $\mathbf{D} \in \mathbb{R}^{n \times n}$ diagonal such that $\mathbf{A} = \mathbf{V} \mathbf{D} \mathbf{V}^H$ [178]. Then, the *graph Fourier transform* of \mathbf{x} is given by [183, 184]

$$\bar{\mathbf{x}} = \mathbf{V}^H \mathbf{x}. \quad (7.33)$$

Observe that if \mathbf{A} is normal we obtain a spectral energy conservation property analog to Parseval's theorem in classical signal processing: it is ready to see that $\|\bar{\mathbf{x}}\|_2 = \|\mathbf{x}\|_2$ if and only if \mathbf{V} in (7.33) is unitary, which holds if and only if \mathbf{A} is normal [178].

Similar to traditional signal processing, a graph signal \mathbf{x} is said to be *spectrally sparse* (*ssparse*) when its spectral representation is sparse. Explicitly, \mathbf{x} is \mathcal{K} -*ssparse* if $\bar{\mathbf{x}}$ in (7.33) is such that $\bar{\mathbf{x}}_{\mathcal{V} \setminus \mathcal{K}}$ is a zero vector. Then,

$$\mathbf{x} = \mathbf{V}_{\mathcal{K}} \bar{\mathbf{x}}_{\mathcal{K}}. \quad (7.34)$$

Note that spectrally sparse signals are a superset of bandlimited (“low-pass”) signals. Hence, all results in this work apply to bandlimited signals regardless of the graph frequency order adopted [185–187].

The interest in \mathcal{K} -ssparse or bandlimited graph signals is motivated similarly to traditional signal processing: these signals can be sampled and interpolated without loss of information. Indeed, take sampling to be the operation of observing the value of a graph signal on $\mathcal{S} \subseteq \mathcal{V}$, the *sampling set*. Then, there exists a set \mathcal{S} of size $|\mathcal{K}|$ such that \mathbf{x} can be recovered exactly from $\mathbf{x}_{\mathcal{S}}$ [186–189]. If, however, only a corrupted version of $\mathbf{x}_{\mathcal{S}}$ is available, then \mathbf{x} can only be approximated. To do so, the next section poses noisy interpolation as a Bayesian estimation problem, from which the minimum MSE interpolation operator can be derived.

7.4.1.1 Graph signal interpolation

We study graph signal interpolation as a Bayesian estimation problem. Formally, let $\mathbf{x} \in \mathbb{C}^n$ be a graph signal and $\mathcal{S} \subseteq \mathcal{V}$ be a sampling set. We wish to estimate

$$\mathbf{z} = \mathbf{H}\mathbf{x}, \quad (7.35)$$

for some matrix $\mathbf{H} \in \mathbb{C}^{m \times n}$ based on the samples $\mathbf{y}_{\mathcal{S}}$ taken from

$$\mathbf{y} = \mathbf{x} + \mathbf{w}, \quad (7.36)$$

where $\mathbf{w} \in \mathbb{C}^n$ is a circular zero-mean noise vector. By circular we mean that its *relation matrix* vanishes, i.e., that $\mathbb{E} \mathbf{w} \mathbf{w}^T = \mathbf{0}$ [190]. Note that (7.35) accounts for scenarios in which we are not interested in the graph signal itself but on a post-processed value, such as the output of a linear classifier or estimator. The usual graph signal interpolation problem from [54, 55, 185–189] is recovered by taking $\mathbf{H} = \mathbf{I}$.

The prior distribution of \mathbf{x} reflects the fact that the graph signal is \mathcal{K} -sparse by assuming it is a circular zero-mean distribution with covariance matrix $\mathbf{\Sigma} = \mathbf{x} \mathbf{x}^H = \mathbf{V}_{\mathcal{K}} \mathbf{\Lambda} \mathbf{V}_{\mathcal{K}}^H$ for $\mathbf{\Lambda} = \text{diag}(\lambda_i)$, $\lambda_i \in \mathbb{R}_+$. We assume without loss of generality that $\mathbf{\Lambda}$ is full-rank. Otherwise, remove from \mathcal{K} any element i for which $\lambda_i = 0$. Note that this is equivalent to placing a zero-mean uncorrelated prior on $\bar{\mathbf{x}}$ in (7.33). Hence, this model can also be interpreted as the generative model for a *wide-sense stationary* random process on \mathbb{G} [191–193]. The noise prior is taken as a zero-mean circular distribution with covariance matrix $\mathbf{\Lambda}_w = \text{diag}(\lambda_{w,i})$, $\lambda_{w,i} \in \mathbb{R}_+$ and $\lambda_{w,i} > 0$.

We consider estimates of \mathbf{z} of the form

$$\hat{\mathbf{z}}(\mathcal{S}) = \mathbf{L}(\mathcal{S}) \mathbf{y}_{\mathcal{S}}, \quad (7.37)$$

for some $\mathbf{L}(\mathcal{S}) \in \mathbb{C}^{n \times |\mathcal{S}|}$. Because \mathbf{L} recovers (approximates) \mathbf{z} from the samples $\mathbf{y}_{\mathcal{S}}$, it is referred to as a *linear interpolation operator* [185–187]. An optimal interpolation operator can be found for each \mathcal{S} by minimizing the interpolation error covariance matrix as in

$$\begin{aligned} & \underset{\mathbf{L}}{\text{minimize}} && \mathbf{K}[\hat{\mathbf{z}}(\mathcal{S})] \\ & \text{subject to} && \hat{\mathbf{z}}(\mathcal{S}) = \mathbf{L} \mathbf{y}_{\mathcal{S}} \end{aligned} \quad (7.38)$$

where $\mathbf{K}[\hat{\mathbf{z}}(\mathcal{S})] = \mathbb{E} \left[(\mathbf{z} - \hat{\mathbf{z}}(\mathcal{S})) (\mathbf{z} - \hat{\mathbf{z}}(\mathcal{S}))^H \mid \mathbf{x}, \mathbf{w} \right]$ and the minimum is taken with respect to the partial ordering of the PSD cone (see Remark 2). We have omitted the dependence of \mathbf{L} on \mathcal{S} for clarity. Our interest in solving (7.38) instead of minimizing the MSE directly is that it is more general. In particular, a solution of (7.38) also minimizes any spectral function of \mathbf{K} , including the MSE and the log-determinant. The following proposition gives an explicit solution to this problem. To clarify the derivations, we define the selection matrix $\mathbf{C} \in \{0, 1\}^{|\mathcal{S}| \times N}$ composed of the identity matrix rows with indices in \mathcal{S} , so that the samples of (7.36) can be written as $\mathbf{y}_{\mathcal{S}} = \mathbf{C} \mathbf{y}$.

Proposition 13. Let $\mathbf{y} = \mathbf{x} + \mathbf{w}$ be noisy observations of a graph signal \mathbf{x} . Let the priors on \mathbf{x} and \mathbf{w} be zero-mean circular distributions with covariances $\mathbf{\Sigma} = \mathbf{V}_{\mathcal{K}}\mathbf{\Lambda}\mathbf{V}_{\mathcal{K}}^H$, $\mathbf{\Lambda} = \text{diag}(\lambda_i)$, and $\mathbf{\Lambda}_w = \text{diag}(\lambda_{w,i})$ respectively. Given a sampling set \mathcal{S} , the optimal Bayesian linear interpolator \mathbf{L}^* that solves problem (7.38) is obtained as a solution of

$$\mathbf{L}^* \mathbf{C} (\mathbf{\Sigma} + \mathbf{\Lambda}_w) \mathbf{C}^T = \mathbf{H} \mathbf{\Sigma} \mathbf{C}^T. \quad (7.39)$$

The error covariance matrix of the optimal interpolation $\hat{\mathbf{z}}^* = \mathbf{L}^* \mathbf{y}_{\mathcal{S}}$ is given by

$$\mathbf{K}^*(\mathcal{S}) = \mathbf{H} \mathbf{V}_{\mathcal{K}} \left(\mathbf{\Lambda}^{-1} + \sum_{i \in \mathcal{S}} \lambda_{w,i}^{-1} \mathbf{v}_i \mathbf{v}_i^H \right)^{-1} \mathbf{V}_{\mathcal{K}}^H \mathbf{H}^H, \quad (7.40)$$

where $\mathbf{V}_{\mathcal{K}} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_N \end{bmatrix}^H$.

Proof. Start by substituting (7.35) and (7.37) into the definition of \mathbf{K} to get

$$\mathbf{K}(\mathbf{L} \mathbf{y}_{\mathcal{S}}) = \mathbb{E} \left[(\mathbf{H} \mathbf{x} - \mathbf{L} \mathbf{C} \mathbf{y}) (\mathbf{H} \mathbf{x} - \mathbf{L} \mathbf{C} \mathbf{y})^H \mid \mathbf{x}, \mathbf{w} \right].$$

Note that we used the fact that $\mathbf{y}_{\mathcal{S}} = \mathbf{C} \mathbf{y}$. Then, using the priors on \mathbf{x} and \mathbf{w} , \mathbf{K} expands to

$$\mathbf{K}(\mathbf{L} \mathbf{y}_{\mathcal{S}}) = \mathbf{H} \mathbf{\Sigma} \mathbf{H}^H - \mathbf{L} \mathbf{C} \mathbf{\Sigma} \mathbf{H}^H - \mathbf{H} \mathbf{\Sigma} \mathbf{C}^T \mathbf{L}^H + \mathbf{L} \mathbf{C} (\mathbf{\Sigma} + \mathbf{\Lambda}_w) \mathbf{C}^T \mathbf{L}^H. \quad (7.41)$$

From the partial ordering of the PSD cone, \mathbf{L}^* can be obtained by minimizing the scalar cost function

$$J(\mathbf{L}) = \mathbf{b}^H \mathbf{K}(\mathbf{L} \mathbf{y}_{\mathcal{S}}) \mathbf{b} \quad (7.42)$$

simultaneously for all $\mathbf{b} \in \mathbb{C}^n$ [56]. Substituting (7.41) into (7.42) and setting its gradient with respect to $\mathbf{b}^H \mathbf{L}$ to zero gives

$$\frac{\partial J(\mathbf{L})}{\partial \mathbf{b}^H \mathbf{L}} = \mathbf{0} \Leftrightarrow \mathbf{C} (\mathbf{\Sigma} + \mathbf{\Lambda}_w) \mathbf{C}^T \mathbf{L}^H \mathbf{b} = \mathbf{C} \mathbf{\Sigma} \mathbf{H}^H \mathbf{b}.$$

Since this must hold for all \mathbf{b} simultaneously, we obtain (7.39).

To determine the error covariance matrix \mathbf{K}^* of the optimal interpolator, replace any \mathbf{L}^* satisfying (7.39) into (7.41) and expand $\mathbf{\Sigma} = \mathbf{V}_\mathcal{K} \mathbf{\Lambda} \mathbf{V}_\mathcal{K}^H$ to get

$$\mathbf{K}^*(\mathbf{L}^* \mathbf{y}_\mathcal{S}) = \mathbf{H} \mathbf{V}_\mathcal{K} \left\{ \mathbf{\Lambda} - \mathbf{\Lambda} \mathbf{V}_\mathcal{K}^H \mathbf{C}^T \times [\mathbf{C} (\mathbf{V}_\mathcal{K} \mathbf{\Lambda} \mathbf{V}_\mathcal{K}^H + \mathbf{\Lambda}_w) \mathbf{C}^T]^{-1} \mathbf{C} \mathbf{V}_\mathcal{K} \mathbf{\Lambda} \right\} \mathbf{V}_\mathcal{K}^H \mathbf{H}^H. \quad (7.43)$$

Note that (7.43) does not depend on \mathbf{L}^* or $\mathbf{y}_\mathcal{S}$, only on the sampling set \mathcal{S} through the selection matrix \mathbf{C} . Moreover, since $\mathbf{\Lambda}_w$ is diagonal and full rank, $(\mathbf{C} \mathbf{\Lambda}_w \mathbf{C}^T)^{-1} = \mathbf{C} \mathbf{\Lambda}_w^{-1} \mathbf{C}^T$, so that the inverse in (7.43) always exists. Therefore, using the matrix inversion lemma [178] gives

$$\mathbf{K}^*(\mathcal{S}) = \mathbf{H} \mathbf{V}_\mathcal{K} (\mathbf{\Lambda}^{-1} + \mathbf{V}_\mathcal{K}^H \mathbf{C}^T \mathbf{C} \mathbf{\Lambda}_w^{-1} \mathbf{C}^T \mathbf{C} \mathbf{V}_\mathcal{K})^{-1} \mathbf{V}_\mathcal{K}^H \mathbf{H}^H.$$

Given that $\mathbf{C}^T \mathbf{C}$ is a diagonal matrix with ones on the indices in \mathcal{S} and zeros everywhere else, we obtain (7.40) by noting that

$$\mathbf{V}_\mathcal{K}^H \mathbf{C}^T \mathbf{C} \mathbf{\Lambda}_w^{-1} \mathbf{C}^T \mathbf{C} \mathbf{V}_\mathcal{K} = \sum_{i \in \mathcal{S}} \lambda_{w,i}^{-1} \mathbf{v}_i \mathbf{v}_i^H,$$

$$\text{for } \mathbf{V}_\mathcal{K} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_N \end{bmatrix}^H.$$

■

Given prior distributions for the graph signal and noise, Proposition 13 determines the optimal linear interpolator from the samples in \mathcal{S} . If the priors on \mathbf{x} and \mathbf{w} are moreover Gaussian, then $\hat{\mathbf{z}}^* = \mathbf{L}^* \mathbf{y}_\mathcal{S}$ is also the maximum likelihood estimate of \mathbf{z} [56]. An important consequence of the Bayesian statement of Proposition 13 is that $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_w$ are taken from prior distributions on the signal and noise. Thus, their actual values need not be known exactly. Note that the optimal error covariance matrix \mathbf{K}^* now depends only on the sampling set \mathcal{S} , since it measures the error of the optimal estimator \mathbf{L}^* . Moreover, although we assume that the interpolation is performed as a single step projection, iterative procedures can also be used [194, 195].

Despite our assumption that $\mathbf{\Lambda}_w$ is full-rank, (7.39) also holds in the noiseless case ($\mathbf{\Lambda}_w = \mathbf{0}$). Its solution, however, may no longer be unique. In particular, this happens if the sampling set is not sufficient to determine \mathbf{z} , i.e., if $\mathbf{C} \mathbf{V}_\mathcal{K}$ is rank-deficient [186–189]. In contrast, when $\mathbf{\Lambda}_w \succ \mathbf{0}$, the matrix on the left-hand side of (7.39) is always invertible and \mathbf{L}^* is unique for each \mathcal{S} . This is

similar to the well-known regularization effect of noise in Kalman filtering [56]. The interpolation performance given in (7.40), however, is not the same for all sampling sets.

Remark 2. Problem (7.38) is a PSD matrix minimization problem that searches for the optimal interpolator \mathbf{L}^* that minimizes the error covariance matrix \mathbf{K} . In general, optimization problems in the PSD cone need not have a solution. Since the ordering of PSD matrices is only partial, the existence of a matrix that is smaller than all other matrices is not guaranteed [130]. As shown in Proposition 13, this is not the case here. Problem (7.38) admits a dominant solution \mathbf{L}^* in the PSD cone, i.e., it holds that $\mathbf{K}(\mathbf{L}^* \mathbf{y}_S) \preceq \mathbf{K}(\mathbf{L} \mathbf{y}_S)$ for all $\mathbf{L} \in \mathbb{C}^{n \times |\mathcal{S}|}$. This means that \mathbf{L}^* minimizes all the eigenvalues of \mathbf{K} simultaneously. Equivalently, it implies that \mathbf{L}^* is a solution to the minimization of any spectral function of \mathbf{K} . In particular, it follows that \mathbf{L}^* minimizes the MSE, since $\text{MSE}(\hat{\mathbf{z}}) := \mathbb{E} \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 = \text{Tr}[\mathbf{K}(\hat{\mathbf{z}})]$, and the $\log \det[\mathbf{K}(\hat{\mathbf{z}})]$.

7.4.1.2 Sampling set selection

Proposition 13 allows us to evaluate the optimal interpolator \mathbf{L}^* that minimizes the estimation error covariance matrix for a given sampling set. This does not guarantee, however, that there is no other sampling set of the same size for which the interpolation error is smaller. To address this issue, we investigate the *sampling set selection* problem which sets out to find the sampling set that minimizes the interpolation MSE over all sampling sets. Explicitly, we wish to solve

$$\begin{aligned} & \underset{\mathcal{S} \subseteq \mathcal{V}}{\text{minimize}} && \text{MSE}(\mathcal{S}) \\ & \text{subject to} && |\mathcal{S}| \leq k \end{aligned} \tag{7.44}$$

where $\text{MSE}(\mathcal{S}) = \text{Tr}[\mathbf{K}^*(\mathcal{S})]$.

An important fact about (7.44) is that increasing \mathcal{S} always decreases MSE. This has two important consequences. First, the unconstrained version of (7.44) is trivial, i.e., its solution is $\mathcal{S} = \mathcal{V}$. Second, it implies that the constraint in (7.44) is tight, i.e., it can be replaced by the equality constraint $|\mathcal{S}| = k$ without changing the problem solution. This property is a direct corollary of Lemma 4 and the fact that \mathbf{K}^* in (7.40) has the form 7.12.

Although Lemma 4 reduces the searching space to sampling sets of size k , (7.44) remains a

combinatorial optimization problem: $\binom{n}{k}$ sampling sets must still be checked, which is impractical even for moderately small n . In fact, due to the irregularity of the domain of graph signals, sampling set selection is NP-hard in general. It is straightforward to see that it is equivalent to the sensor placement or forward regression problems in [167, 174, 196, 197], so that the typical reduction from set cover applies [99].

Since $\text{MSE}(\mathcal{S})$ has the form of the set trace function (7.13), we know that it is α -supermodular. In fact, we can bound its α as in the following proposition:

Proposition 14. *The scalar set functions $\text{MSE}(\mathcal{S}) = \text{Tr}[\mathbf{K}^*(\mathcal{S})]$ is (i) monotone decreasing and (ii) α -supermodular with*

$$\alpha \geq \frac{\lambda_{\max}(\mathbf{\Lambda}_w)^{-1} + \mu_{\max}^{-1}}{\lambda_{\max}(\mathbf{\Lambda}_w)^{-1} + \mu_{\min}^{-1}} \frac{\mu_{\min}^2}{\kappa_2(\mathbf{W}) \mu_{\max}^2}, \quad (7.45)$$

where $\mu_{\min} \leq \lambda_{\min}[\mathbf{\Lambda}^{-1}]$, $\mu_{\max} \geq \lambda_{\max}[\mathbf{\Lambda}^{-1} + \mathbf{V}_{\mathcal{K}}^H \mathbf{\Lambda}_w^{-1} \mathbf{V}_{\mathcal{K}}]$, and $\kappa_2(\mathbf{W})$ is the 2-norm condition number of \mathbf{W} [178].

Proof. See appendix B.3.4. ■

Then, the guarantee from Theorem 8 can be combined with the characterization in (7.45) to yield:

Theorem 13. *Let \mathcal{S}^* be the solution of (7.44) and \mathcal{X}_r be the result of applying Algorithm 3 for $f(\mathcal{X}) = \text{MSE}(\mathcal{X})$. Then,*

$$\frac{\text{MSE}(\mathcal{X}_r) - \text{MSE}(\mathcal{S}^*)}{\text{MSE}(\emptyset) - \text{MSE}(\mathcal{S}^*)} \leq e^{-\alpha r/k}, \quad (7.46)$$

where

$$\alpha \geq \frac{\lambda_{\max}(\mathbf{\Lambda}_w)^{-1} + \mu_{\max}^{-1}}{\lambda_{\max}(\mathbf{\Lambda}_w)^{-1} + \mu_{\min}^{-1}} \frac{\mu_{\min}^2}{\kappa_2(\mathbf{W}) \mu_{\max}^2} \quad (7.47)$$

for $\mu_{\min} \leq \lambda_{\min}[\mathbf{\Lambda}^{-1}]$, $\mu_{\max} \geq \lambda_{\max}[\mathbf{\Lambda}^{-1} + \mathbf{V}_{\mathcal{K}}^H \mathbf{\Lambda}_w^{-1} \mathbf{V}_{\mathcal{K}}]$, and $\kappa_2(\mathbf{W})$ is the 2-norm condition number of \mathbf{W} . Assuming $\mathbf{\Lambda} = \sigma_x^2 \mathbf{I}$ and $\mathbf{\Lambda}_w = \sigma_w^2 \mathbf{I}$, (7.47) reduces to

$$\alpha \geq \frac{1 + 2\gamma}{\kappa_2(\mathbf{W}) (1 + \gamma)^4}, \quad \text{for } \gamma = \frac{\sigma_x^2}{\sigma_w^2}. \quad (7.48)$$

Theorem 13 establishes that a near-optimal solution to the sampling set selection problem in (7.44) can be obtained efficiently using greedy search. Though strong empirical evidence ex-

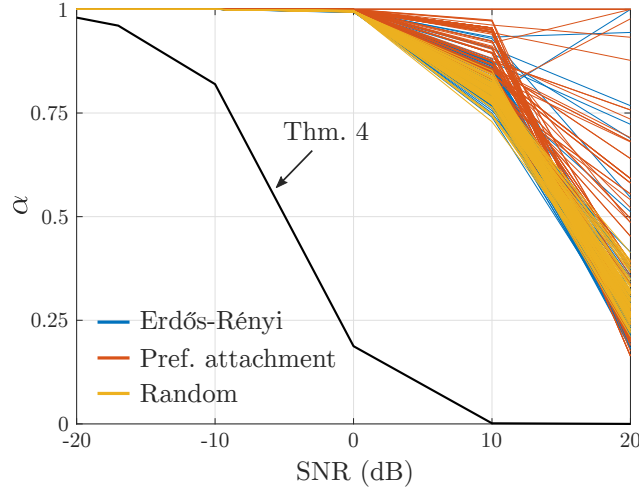


Figure 7.2: Comparison between the bound in (7.48) and α

ists that greedily minimizing the MSE yields good results in contexts such as regression, dictionary learning, and graph signal processing [174, 186–189, 198], this result is counter-intuitive given that the MSE is not supermodular in general. For instance, restrictive and often unrealistic conditions on data distribution are required to obtain supermodularity in the context of regression [174].

Theorem 13 therefore reconciles the empirical success of greedy sampling set selection and the non-supermodularity of the MSE by bounding the suboptimality of greedy sampling. In particular, (7.48) gives a simple bound on α in terms of the SNR and the condition number of \mathbf{W} that gives clear insights into its behavior. Indeed, as $\gamma \rightarrow \infty$ and we approach the noiseless case, $\alpha \rightarrow 0$. This is expected as in the noiseless case almost every set of size $|\mathcal{K}|$ achieves perfect reconstruction, so that the choice of sampling nodes is irrelevant. On the other hand, $\alpha \rightarrow 1$ as $\gamma \rightarrow 0$, i.e., the MSE becomes closer to supermodular as the SNR decreases. Given that reconstruction errors are small for high SNR, Theorem 13 guarantees that greedy sampling performs well when it is most needed. Similar trends can be observed in the more general setting of (7.47). These observations are illustrated in Figure 7.2 that compares the bound in (7.48) to the true value of α for the MSE (found by exhaustive search) in 100 realizations of random graphs (see Section 7.4.1.3 for details).

Before proceeding, the complexity issue of greedy sampling set selection must be addressed. The greedy search in Algorithm 3 requires $|\mathcal{V}|rc_f$ operations, where c_f is the cost of evaluating the objective f . As it is, problem (7.44) has $c_f = \mathcal{O}(|\mathcal{K}|^3)$. It can, however, be reduced using the matrix

Algorithm 5 Greedy graph sampling

$\mathcal{G}_0 = \{\}$ and $\mathbf{K}_0^* = \mathbf{\Lambda}$
for $j = 1, \dots, \ell$ **do**
 $u = \operatorname{argmax}_{s \in \mathcal{V} \setminus \mathcal{G}_{j-1}} \frac{\mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{W} \mathbf{K}_{j-1}^* \mathbf{v}_u}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{v}_u} \quad \triangleright \mathcal{O}(|\mathcal{V}||\mathcal{K}|^2)$
 $\mathbf{K}_j^* = \mathbf{K}_{j-1}^* - \mathbf{W} \frac{\mathbf{K}_{j-1}^* \mathbf{v}_u \mathbf{v}_u^H \mathbf{K}_{j-1}^*}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{v}_u} \quad \triangleright \mathcal{O}(|\mathcal{K}|^2)$
 $\mathcal{G}_j = \mathcal{G}_{j-1} \cup \{u\}$
end for

inversion lemma [178].

Indeed, start by noticing that the first step of the greedy approximation of problem (7.44) involves finding (see Algorithm 3)

$$u = \operatorname{argmin}_{s \in \mathcal{V}} \operatorname{Tr} \left[\mathbf{K}^* (\mathcal{G}_{j-1} \cup \{s\}) \right],$$

which, using the definition of \mathbf{K}^* in (7.40) and the circular commutation property of the trace, requires the evaluation of

$$\operatorname{Tr} \left[\mathbf{K}^* (\mathcal{G}_{j-1} \cup \{s\}) \right] = \operatorname{Tr} \left[\mathbf{W} \left(\mathbf{\Lambda}^{-1} + \sum_{i \in \mathcal{G}_{j-1}} \lambda_{w,i}^{-1} \mathbf{v}_i \mathbf{v}_i^H + \lambda_{w,s}^{-1} \mathbf{v}_s \mathbf{v}_s^H \right)^{-1} \right],$$

where once again $\mathbf{W} = \mathbf{V}_{\mathcal{K}}^H \mathbf{H}^H \mathbf{H} \mathbf{V}_{\mathcal{K}}$. Letting $\mathbf{K}_j^* = \mathbf{K}^*(\mathcal{G}_j)$ and using the matrix inversion lemma, we can reduce the update of \mathbf{K}^* to

$$\mathbf{K}^*(\mathcal{G}_j \cup \{s\}) = \mathbf{K}_{j-1}^* - \mathbf{W} \frac{\mathbf{K}_{j-1}^* \mathbf{v}_u \mathbf{v}_u^H \mathbf{K}_{j-1}^*}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{v}_u}. \quad (7.49)$$

From linearity, it is then straightforward to see that finding the minimum of the trace of (7.49) is equivalent to finding the maximum of

$$\operatorname{Tr} \left[\mathbf{W} \frac{\mathbf{K}_{j-1}^* \mathbf{v}_u \mathbf{v}_u^H \mathbf{K}_{j-1}^*}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{v}_u} \right] = \frac{\mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{W} \mathbf{K}_{j-1}^* \mathbf{v}_u}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{K}_{j-1}^* \mathbf{v}_u}. \quad (7.50)$$

The greedy sampling set selection procedure obtained by leveraging (7.49) and (7.50) is presented

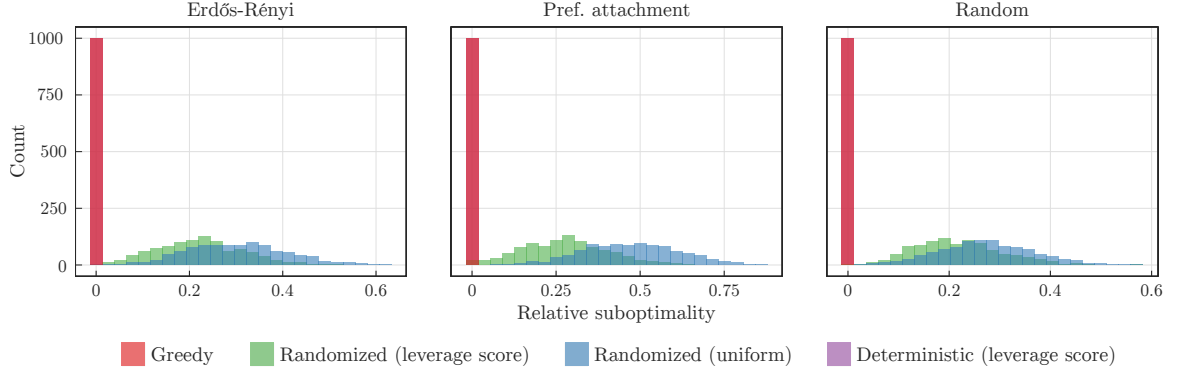


Figure 7.3: Relative suboptimality of sampling schemes for low SNR (SNR = -20 dB)

in Algorithm 5. This algorithm now requires only $\mathcal{O}(r|\mathcal{V}||\mathcal{K}|^2)$ operations.

Remark 3. Since the MSE is *not* supermodular, it is common to see surrogate supermodular figures of merit used instead, specially in statistics and experiment design [168, 174, 196, 197]. In particular, the log-determinant $\log \det[\mathbf{K}^*(\mathcal{S})]$ is a common alternative to the objective $\text{MSE}(\mathcal{S}) = \text{Tr}[\mathbf{K}^*(\mathcal{S})]$ used in (7.44). This is justified because the $\log \det[\mathbf{K}^*(\mathcal{S})]$ is proportional to the volume of the confidence ellipsoids of the estimate when the data is Gaussian [197, 199]. This choice of objective is also common in the sensor placement literature due to its relation to information theoretic measures, such as entropy and mutual information [196]. By replacing the trace operator in (7.44) by the log det, the problem becomes a supermodular function minimization that can be efficiently approximated using greedy search, as shown in [55, 200]. We remark that minimizing the log det of the error covariance matrix and the MSE are not equivalent problems.

7.4.1.3 Numerical experiments

In this section, we start by evaluating the performance greedy sampling set selection (Algorithm 5). For comparison, we also display the results obtained by the *uniform* and *leverage score* randomized methods from [185] and a *deterministic* heuristic based on sampling nodes with the highest leverage score ($\|\mathbf{v}_i\|_2^2$). In the following examples, we use undirected graphs generated using the *Erdős-Rényi* model, in which an edge is placed between two nodes with probability $p = 0.2$; the *preferential attachment* model [201], in which nodes are added one at a time and connected to a node already

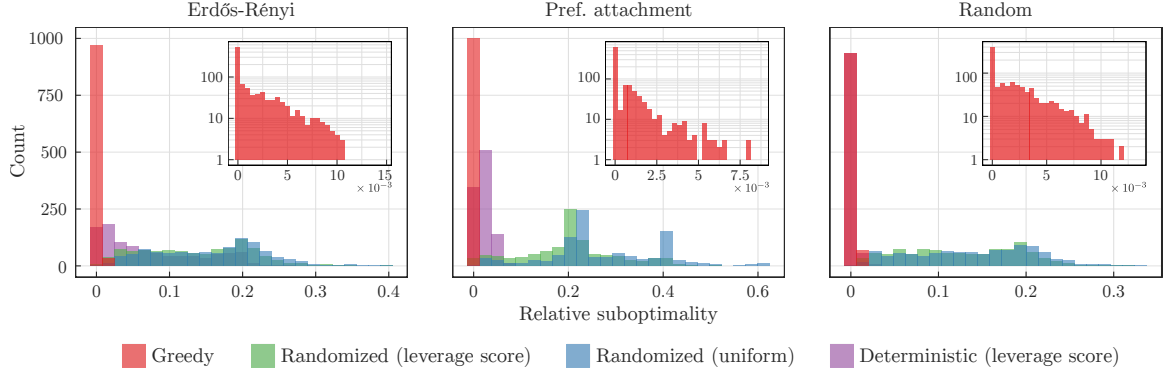


Figure 7.4: Relative suboptimality of sampling schemes for high SNR ($\text{SNR} = 20$ dB)

in the graph with probability proportional to its degree; and a *random undirected graph*, obtained by assigning a weight to all possible edges uniformly at random from $[0, 1]$.

The figure of merit in the following simulations is the *relative suboptimality* from (7.46). Since it depends on the optimal sampling set which needs to be determined by exhaustive search, we focus on graphs with $n = 20$ nodes. Bandlimited graph signals are generated by taking $\mathbf{V}_{\mathcal{K}}$ in (7.34) to be the eigenvectors of the graph adjacency matrix relative to the five eigenvalues with largest magnitude ($|\mathcal{K}| = 5$). The random vectors $\bar{\mathbf{x}}$ in (7.34) and \mathbf{w} in (7.36) are realizations of zero-mean Gaussian random variables with covariance matrices $\mathbf{\Lambda} = \mathbf{I}$ and $\mathbf{\Lambda}_w = \sigma_w^2 \mathbf{I}$, where σ_w^2 is varied to obtain different SNRs. The transform in (7.35) is taken to be the identity ($\mathbf{H} = \mathbf{I}$) and the sampling set size is chosen as $\ell = |\mathcal{K}| = 5$.

Figures 7.3 and 7.4 display histograms of the relative suboptimality for 1000 realizations of graphs and graph signals with $\sigma_w^2 = 10^2$ ($\text{SNR} = -20$ dB) and $\sigma_w^2 = 10^{-2}$ ($\text{SNR} = 20$ dB), respectively. As predicted by Theorem 13, greedy sampling set selection performs better in low SNR environments, where the optimal sampling set was obtained in more than 95% of the realizations. Nevertheless, even in high SNRs, it found the optimal sampling set almost half of the time. In fact, note that Algorithm 5 typically performs much better than the bounds in Theorem 13 (see details in Fig. 7.4). For comparison, results for greedily optimizing $\log \det [\mathbf{K}^*(\mathcal{S})]$, a supermodular function, are shown in Figure 7.5. Although the MSE is α -supermodular with $\alpha < 1$, the relative suboptimality obtained by using both cost functions is comparable.

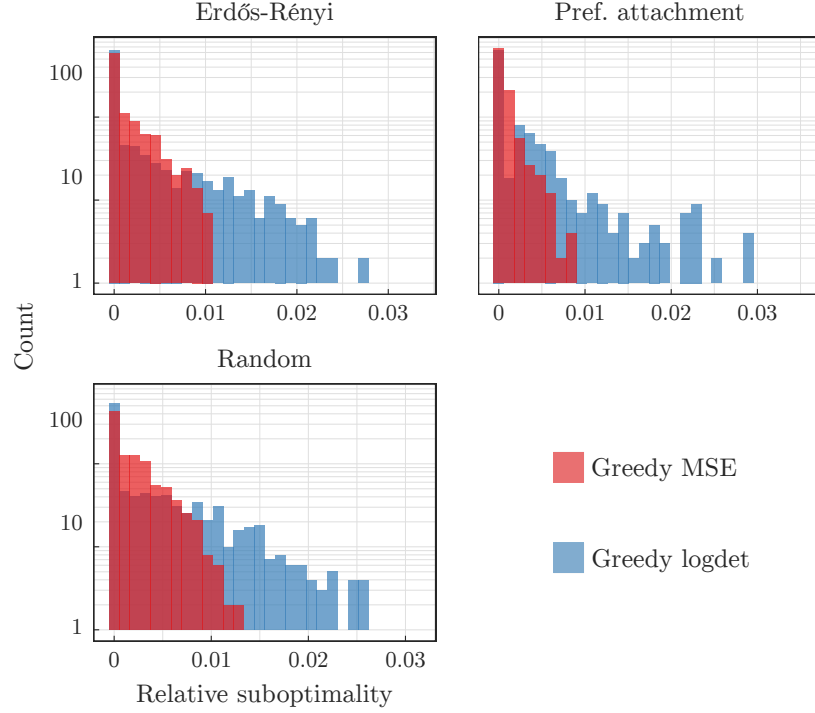


Figure 7.5: Relative suboptimality of MSE and log det (SNR = 20 dB)

It is worth noting that, although the deterministic leverage score ranking technique often yields good results, there are advantages to greedy sampling set selection, specially for higher SNR. The randomized sampling schemes, on the other hand, do not perform as well for single problem instances. To be fair, these methods are more appropriate when several sampling sets of the same graph signal are considered. Indeed, the performance measures in [185] hold in expectation over sampling realizations.

Evaluating the relative suboptimality for larger graphs is untractable. However, since these sampling set selection techniques build the sampling set sequentially, we can assess their performance in terms of the sampling set size required to obtain a given MSE reduction. Figure 7.6 displays the distribution of the sampling set size required to achieve a 90% reduction in the MSE with respect to the empty set. The plots are obtained from 1000 graphs and signals realizations with $n = 100$ nodes, $\mathbf{V}_{\mathcal{K}}$ in (7.34) composed the eigenvectors relative to the seven eigenvalues with largest magnitude ($|\mathcal{K}| = 7$), and $\sigma_w^2 = 10^{-2}$. Although Theorem 13 estimates that Algorithm 5 requires sets considerably larger to recover the same near-optimal guarantees as supermodular functions,

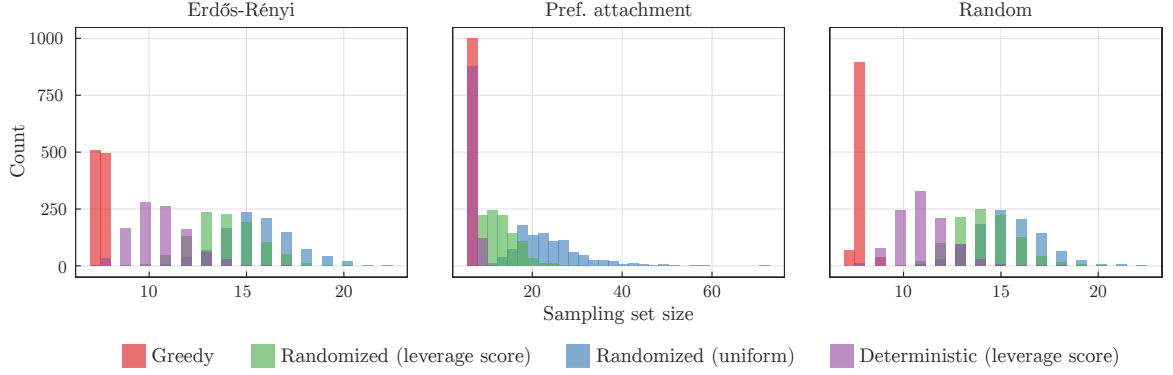


Figure 7.6: Sampling set size for 90% reduction of MSE

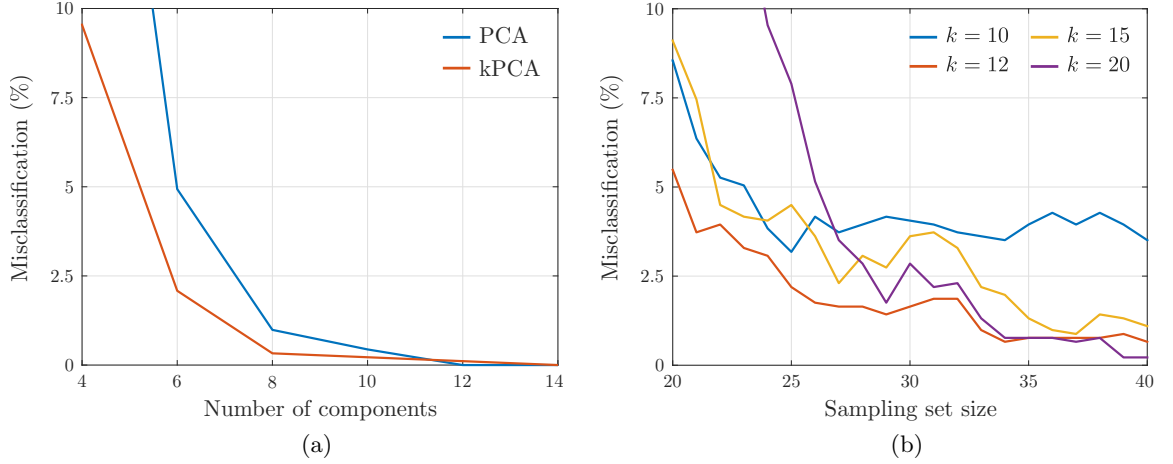


Figure 7.7: Classification performance: (a) PCA and kPCA and (b) greedy subsampled kPCA.

greedy sampling obtained a sampling set of size exactly $|\mathcal{K}|$ in more than 50% of the realizations. Moreover, as noted in [185], we can now see that leverage score sampling has similar performance to uniform sampling for Erdős-Rényi graphs, but gives better results for the preferential attachment model.

A more practically interesting application of greedy graph sampling is found in the context of kernel PCA (kPCA). Kernel PCA is a nonlinear version of PCA [202] that also identifies data subspaces by truncating the eigenvalue decomposition (EVD) of a Gram matrix Φ . However, whereas PCA uses the empirical covariance matrix, kPCA constructs Φ by evaluating inner products between data points in a higher dimensional space \mathbb{F} known as the *feature space*. Since the map $\varphi : \mathbb{R}^m \rightarrow \mathbb{F}$

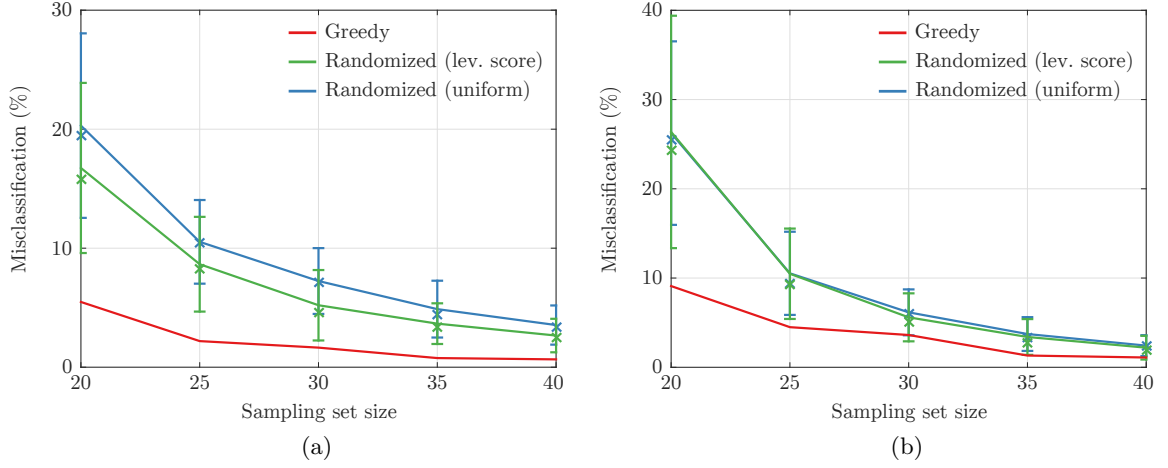


Figure 7.8: Classification performance of subsampled kPCA for different sampling schemes: (a) $k = 12$ components and (b) $k = 15$ components. Mean (solid line), median (\times), and error bars (one standard deviation) based on 100 sampling realizations.

can be nonlinear and \mathbb{F} typically has infinite dimensionality, kPCA results in richer subspaces than PCA [202–204].

Naturally, the dimensionality of \mathbb{F} poses a challenge for constructing the Gram matrix. This problem is addressed using the so called *kernel trick* [202–204]. A kernel is a function κ that allows the inner product in \mathbb{F} to be evaluated directly from vectors in \mathbb{R}^m , i.e., $\kappa(\mathbf{r}, \mathbf{s}) = \langle \varphi(\mathbf{r}), \varphi(\mathbf{s}) \rangle_{\mathbb{F}}$. We can use κ to construct Φ from a training set $\{\mathbf{u}_i\}_{i=1,\dots,n}$, $\mathbf{u}_i \in \mathbb{R}^m$, as in

$$\Phi = [\kappa(\mathbf{u}_i, \mathbf{u}_j)]_{i,j=1,\dots,n}. \quad (7.51)$$

Kernel PCA identifies the data subspace as the span of the first k eigenvectors of Φ , i.e., as $\text{colspan}(\mathbf{V}_{\mathcal{K}})$, where $\Phi = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ is the EVD of Φ with eigenvalues in decreasing order and $\mathcal{K} = 1, \dots, k$. Using the representer’s theorem [204], any data point \mathbf{y} can be projected onto this subspace by

$$\bar{\mathbf{y}} = \mathbf{V}_{\mathcal{K}}^H \tilde{\mathbf{y}}, \quad \tilde{\mathbf{y}} = [\kappa(\mathbf{u}_i, \mathbf{y})]_{i=1,\dots,n}. \quad (7.52)$$

The projection in (7.52) requires $\Theta(kn)$ operations and n KEs, making this method impractical for large data sets even if the dimension k of the subspace of interest is small. Indeed, although the training phase in (7.51) is usually performed offline, (7.52) needs to be evaluated during the

operation phase for every new data point. In [205], this issue was addressed by using a Gaussian generative model for Φ and showing that its maximum likelihood estimate depends only on a subset of the \mathbf{u}_i . Another approach is to impose sparsity on \mathbf{V} *a priori* so that it depends only on a reduced number of training points [203]. Alternatively, one can find a representative subset of the training data and apply kPCA to that subset [206]. The issue with the latter method is that finding a good data subset is known to be a hard problem [207, 208]. In fact, it is related to the problem of sampling set selection in GSP.

Indeed, since we used the same notation as in when we described graph signals, formulating kPCA in the context of GSP is straightforward. Let the graph \mathbb{G} have adjacency matrix $\mathbf{A} = \Phi$, which is symmetric and normal, so that (7.52) has the form of a (partial) graph Fourier transform (7.34). In other words, (7.52) can be interpreted as enforcing graph signals of the form $\tilde{\mathbf{y}}$ to be bandlimited on Φ . Thus, we can apply the sampling and interpolation theory from GSP to put forward a *subsampled kPCA*.

Based on the guarantees from Theorem 13, we use greedy search to obtain a sampling set \mathcal{S} and use the interpolation techniques from Section 7.4.1.1 to recover $\tilde{\mathbf{y}}$ from its samples as in

$$\tilde{\mathbf{y}} = \mathbf{L}^* \tilde{\mathbf{y}}_{\mathcal{S}}. \quad (7.53)$$

Then, (7.52) and (7.53) yield

$$\bar{\mathbf{y}} = \underbrace{\mathbf{V}_{\mathcal{K}}^H \mathbf{L}^*}_{\mathbf{P}} \tilde{\mathbf{y}}_{\mathcal{S}}. \quad (7.54)$$

Notice that \mathbf{P} is now $k \times |\mathcal{S}|$, so that the projection in (7.54) only takes $\Theta(k|\mathcal{S}|)$ operations and $|\mathcal{S}|$ KEs, leading to a considerable complexity reduction ($|\mathcal{S}|/n$) over the direct projection in (7.52). Moreover, kPCA is typically used for dimensionality reduction prior to regression or classification, so that we are actually interested in a linear transformation of $\bar{\mathbf{y}}$. Subsampled kPCA can account for this case by properly choosing \mathbf{H} in (7.35). It is worth noting that contrary to [206], the full dataset is used during the training stage to obtain $\mathbf{V}_{\mathcal{K}}$. However, once \mathbf{P} is determined, only the subset \mathcal{S} is required.

In the sequel, we illustrate this method in a face recognition application using the *faces94* data set [209]. It contains 20 pictures (200×180) of 152 individuals which were converted to black and

white and normalized so that the value of each pixel is in $[-1, 1]$. A training set is obtained by randomly choosing 14 images for each individual (70% of the data set) and the remaining pictures are used for testing. In this application, we use a polynomial kernel of degree $d = 2$ [204] and a one-against-one multiclass support vector machine (SVM) classifier, in which an SVM is trained for each pair of class and the classification is obtained by majority voting (see [210] for details on this scheme). Finally, note that since images in both training and testing sets come from the same data set there is no observation noise \mathbf{w} . Still, σ_w^2 can be used to regularize the matrix inversions in (7.39) and (7.40) [56].

Figure 7.7a shows the misclassification percentage on the test set as a function of the number of components (k) for both PCA and kPCA. Note that kPCA can achieve the same performance as PCA with less components. The results of using greedy subsampled kPCA are shown in Figure 7.7b and clearly illustrate the trade-off between complexity and performance: as the sampling set size increases, the classification errors decrease. However, since misclassification is a nonlinear function of the MSE, it may be advantageous to use more components instead of increasing the sampling set. For instance, kPCA requires $k = 7$ components to achieve a misclassification of 1%, so that evaluating the direct projection in (7.52) takes 2128 KEs and 29785 operations. Greedy subsampled kPCA, on the other hand, can achieve the same performance with $k = 12$ components and $|\mathcal{S}| = 33$, i.e., 33 KEs and 780 operations, a complexity reduction of more than 97%. Nevertheless, using 7 components, greedy subsampled kPCA would require \mathcal{S} to be almost the full training set.

Naturally, the method in (7.51)–(7.54) is not restricted to sampling sets obtained greedily. Thus, in Figure 7.8 we compare greedy sampling to other sampling methods based on their misclassification performance for 12 (Figure 7.8a) and 15 (Figure 7.8b) components. We omit the deterministic leverage score results because it performed consistently worst than the other methods. The average misclassification rates for the randomized schemes are from 1 to 15% higher than those of greedy sampling. Although some realizations yield good classification, their performance varies a lot, especially for smaller sampling sets. Comparing Figures 7.8a and 7.8b, it is ready that in this application the difference between uniform and leverage score sampling becomes less significant as the number of components increases.

Remark 4. Although we discuss the kPCA case, the same arguments apply to the classical PCA.

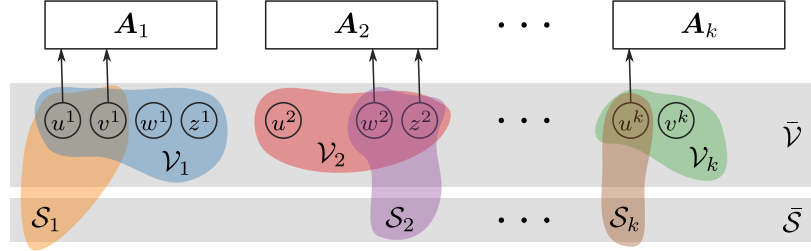


Figure 7.9: Illustration of time-specific and overall input sets and schedules.

It is therefore straightforward to derive an analog *subsampled PCA* technique using (7.51)–(7.54).

7.4.2 Control input scheduling

To showcase results in an application where matroid constraints are essential, we study the input scheduling problem in control. Consider a discrete-time, linear dynamical system and let \mathcal{V}_k denote the set of inputs available at time k as illustrated in Figure 7.9. Depending on the context, these abstract inputs can be used to represent actuators, agents, or both. Suppose that at each time instant, only a subset $\mathcal{S}_k \subseteq \mathcal{V}_k$ of inputs is used, so that the states $\mathbf{x}_k \in \mathbb{R}^n$ of the system evolve according to

$$\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \sum_{i \in \mathcal{S}_k} \mathbf{b}_{i,k} u_{i,k} + \mathbf{w}_k, \quad (7.55)$$

where, for each time k , \mathbf{A}_k denotes the state transition matrix, $\mathbf{b}_{i,k} \in \mathbb{R}^n$ is a vector representing the effect of applying the control action $u_{i,k}$ to the i -th input, and \mathbf{w}_k is a zero-mean Gaussian vector that models the process noise. We assume that $\{\mathbf{w}_j, \mathbf{w}_k\}$ are independent for $j \neq k$ and that their covariance matrices $\mathbb{E} \mathbf{w}_k \mathbf{w}_k^T = \mathbf{W}_k$ are either positive definite ($\mathbf{W}_k \succ 0$) for all k or zero ($\mathbf{W}_k = \mathbf{0}$) for all k , which accounts for the deterministic dynamics case. Let $\bar{\mathcal{V}} = \mathcal{V}_0 \cup \mathcal{V}_1 \cup \dots \cup \mathcal{V}_{N-1}$ be the set of all actuators available over the N -steps time window $[0, N-1]$ and let $\bar{\mathcal{V}} \supseteq \mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \dots \cup \mathcal{S}_{N-1}$ be called a *schedule*, as it denotes the set of active inputs at each time instant. We assume without loss of generality that $\mathcal{V}_j \cap \mathcal{V}_k = \emptyset$ for all $j \neq k$. Any input v available at more than a single time instant can then be represented by unique copies, e.g., as $v^j \in \mathcal{V}_j$ and $v^k \in \mathcal{V}_k$ (Figure 7.9).

Given a schedule $\mathcal{S} \subseteq \bar{\mathcal{V}}$, designing the control actions $u_{i,k}$ reduces to a classical optimal control problem, since (7.55) describes a well-defined, time-varying dynamical system. In particular, we

consider the linear-quadratic-Gaussian (LQG) control problem

$$V^*(\mathcal{S}) = \min_{\mathcal{U}(\mathcal{S})} \mathbb{E} \left[\sum_{k=0}^{N-1} \left(\mathbf{x}_k^T \mathbf{Q}_k \mathbf{x}_k + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k} u_{i,k}^2 \right) + \mathbf{x}_N^T \mathbf{Q}_N \mathbf{x}_N \right], \quad (7.56)$$

where $\mathcal{U}(\mathcal{S}) = \{u_{i,k} \mid i \in \mathcal{S} \cap \mathcal{V}_k\}_{k=0}^{N-1}$ is the set of valid control actions, $\mathbf{Q}_k \succ 0$ for all k are the state weights, and $r_{i,k} > 0$ for all i and k are the input weights. The relative value of these weights describe the trade-off between state regulation (\mathbf{Q}_k) and input cost ($r_{i,k}$). The expectation in (7.56) is taken with respect to the process noise sequence $\{\mathbf{w}_k\}$ and the initial state \mathbf{x}_0 , assumed to be a Gaussian random variable with mean $\bar{\mathbf{x}}_0$ and covariance $\Sigma_0 \succ 0$. When $\mathbf{w}_k = \mathbf{0}$ for all k , (7.56) reduces to the linear quadratic regulator (LQR) problem. It is useful to recall that (7.56) has a closed-form solution that we describe in the following proposition.

Proposition 15. *Given a schedule $\mathcal{S} \subseteq \bar{\mathcal{V}}$, the optimal value $V^*(\mathcal{S})$ of the LQG problem in (7.56) can be written as*

$$V^*(\mathcal{S}) = \text{Tr} [\Sigma_0 \mathbf{P}_0(\mathcal{S})] + \sum_{k=0}^{N-1} \text{Tr} [\mathbf{W}_k \mathbf{P}_{k+1}(\mathcal{S})], \quad (7.57)$$

where the $\mathbf{P}_k(\mathcal{S})$ are obtained via the backward recursion

$$\mathbf{P}_k(\mathcal{S}) = \mathbf{Q}_k + \mathbf{A}_k^T \left(\mathbf{P}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \mathbf{A}_k, \quad (7.58)$$

starting with $\mathbf{P}_N = \mathbf{Q}_N$.

Proof. See appendix B.3.5. ■

Control scheduling refers to the problem of finding a schedule \mathcal{S} that minimizes the control cost in (7.56) subject to time-input constraints, i.e.,

$$\begin{aligned} & \underset{\mathcal{S} \subseteq \bar{\mathcal{V}}}{\text{minimize}} && J(\mathcal{S}) \triangleq V^*(\mathcal{S}) - V^*(\emptyset) \\ & \text{subject to} && \mathcal{S} \in \mathcal{I}_p, \quad p = 1, \dots, P, \end{aligned} \quad (\text{PXIII})$$

where $\mathcal{I}_p \subseteq 2^{\bar{\mathcal{V}}}$ are families of subsets of $\bar{\mathcal{V}}$ that enumerates admissible schedules and $2^{\mathcal{X}}$ denotes the power set of \mathcal{X} (the collection of all finite subsets). Typical scheduling requirements include

(i) limits on the total number of control actions, (ii) limits on the number of inputs used per time instant, (iii) restrictions on the consecutive use of inputs, and combinations thereof. Observe that the constant $V^*(\emptyset)$ in the objective of (PXIII) does not affect the solution of the optimization problem. It is used so that $J(\emptyset) = 0$, which simplifies the presentation of our near-optimal certificates.

7.4.2.1 Near-optimal control scheduling

The complexity of (PXIII) is tightly related to the anatomy of its constraints and objective. Indeed, even obtaining a feasible schedules for (PXIII) can be hard depending on the structure of the \mathcal{I}_p . Not to mention obtaining a good one. In fact, (PXIII) is NP-hard even for constraints as simple as a budget on the total number of control actions ($p = 1$ and $\mathcal{I} = \{\mathcal{S} \subseteq \overline{\mathcal{V}} \mid |\mathcal{S}| \leq s\}$) [164, 165, 167, 174, 196, 197, 211]. Still, depending on the nature of the objective and constraints, there exist simple algorithms able to provide near-optimal approximate solutions. In particular, when \mathcal{I}_p are collections of independent sets of matroids and V^* is α -supermodular, (PXIII) can be solved near-optimally using greedy search (Theorem 10).

These, it turns out, are not strict conditions. Indeed, typical schedule constraints can be described in terms of matroids. Of particular interest are the uniform, partition, and transversal matroids respectively:

- bound on the total number of control actions:

$$\mathcal{I} = \{\mathcal{S} \subseteq \overline{\mathcal{V}} \mid |\mathcal{S}| \leq s\},$$

- bound on the number of inputs used per time instant:

$$\mathcal{I} = \{\mathcal{S} \subseteq \overline{\mathcal{V}} \mid |\mathcal{S} \cap \mathcal{V}_k| \leq s_k\}, \text{ and}$$

- restriction on the consecutive use of inputs:

$$\mathcal{I} = \{\mathcal{S} \subseteq \overline{\mathcal{V}} \mid v^k \notin \mathcal{S} \text{ or } v^{k+1} \notin \mathcal{S}, \text{ for } v^k, v^{k+1} \in \overline{\mathcal{V}}\},$$

Under these conditions, an approximate solution of (PXIII) can be obtained using Algorithm 4. We next show that the objective function J is α -supermodular and monotone decreasing and provide an explicit lower bound on α .

Proposition 16. *Let \mathbf{A}_k in (7.55) be full rank for all k . The normalized actuator scheduling problem objective J is (i) monotonically decreasing and (ii) α -supermodular with*

$$\alpha \geq \min_{k=0, \dots, N-1} \alpha_k, \quad (7.59)$$

where

$$\alpha_k \geq \frac{\lambda_{\min} \left[\tilde{\mathbf{P}}_{k+1}^{-1}(\emptyset) \right]}{\lambda_{\max} \left[\tilde{\mathbf{P}}_{k+1}^{-1}(\bar{\mathcal{V}}) + \sum_{i \in \mathcal{V}_k} r_{i,k}^{-1} \tilde{\mathbf{b}}_{i,k} \tilde{\mathbf{b}}_{i,k}^T \right]}, \quad (7.60)$$

for $\tilde{\mathbf{P}}_{k+1}(S) = \mathbf{H}_k^{1/2} \mathbf{P}_{k+1}(S) \mathbf{H}_k^{1/2}$, $\tilde{\mathbf{b}}_{i,k} = \mathbf{H}_k^{-1/2} \mathbf{b}_{i,k}$, $\mathbf{H}_0 = \mathbf{A}_0 \Sigma_0 \mathbf{A}_0^T$, and $\mathbf{H}_k = \mathbf{A}_k \mathbf{W}_{k-1} \mathbf{A}_k^T$ for $k \geq 1$.

Proof. See appendix B.3.6. ■

Remark 5. The full rank hypothesis on \mathbf{A}_k can be lifted using a continuity argument. However, the bound in (7.59) is trivial ($\alpha \geq 0$) for rank deficient state transition matrices, so we focus only on the case of practical significance.

Proposition 16 states that the objective of (PXIII) is α -supermodular for α as in (7.59). Notice that the lower bound on α is explicit in that it can be evaluated in terms of the parameters of the dynamical system and the weights \mathbf{Q}_k and $r_{i,k}$ in (7.56). In other words, (7.59) allows us to determine α *a priori* for the objective J and with Theorem 10, give near-optimality guarantees on the greedy solution of (PXIII). We collect this result in the theorem below.

Theorem 14. *Consider the control scheduling problem (PXIII) in which $(\bar{\mathcal{V}}, \mathcal{I}_p)$ is a matroid for each p and let \mathcal{S}^* be its optimal solution and \mathcal{S}_g be its greedy solution obtained using Algorithm 4. Then,*

$$J(\mathcal{S}_g) \leq \frac{\alpha}{\alpha + P} J(\mathcal{S}^*) \quad (7.61)$$

with $\alpha \geq \min_{k=0, \dots, N-1} \alpha_k$ for

$$\alpha_k \geq \frac{\lambda_{\min} \left[\tilde{\mathbf{P}}_{k+1}^{-1}(\emptyset) \right]}{\lambda_{\max} \left[\tilde{\mathbf{P}}_{k+1}^{-1}(\bar{\mathcal{V}}) + \sum_{i \in \mathcal{V}_k} r_{i,k}^{-1} \tilde{\mathbf{b}}_{i,k} \tilde{\mathbf{b}}_{i,k}^T \right]}, \quad (7.62)$$

where $\tilde{\mathbf{P}}_{k+1}(S) = \mathbf{H}_k^{1/2} \mathbf{P}_{k+1}(S) \mathbf{H}_k^{1/2}$, for $\mathbf{P}_k(S)$ defined in (7.58), and $\tilde{\mathbf{b}}_{i,k} = \mathbf{H}_k^{-1/2} \mathbf{b}_{i,k}$. The

transformations are the positive-definite square roots of \mathbf{H}_k , defined as $\mathbf{H}_0 = \mathbf{A}_0 \mathbf{\Sigma}_0 \mathbf{A}_0^T$ and $\mathbf{H}_k = \mathbf{A}_k \mathbf{W}_{k-1} \mathbf{A}_k^T$ for $k \geq 1$.

Theorem 14 provides a near-optimal certificate for the greedy solution of the control scheduling problem (PXIII) as a function of its parameters. Note that the larger the α , the better the guarantee, and that although J is not supermodular in general, there are situations in which its violations of the diminishing returns property are small.

To see when this is the case, observe that (7.62) is large when

$$\sum_{i \in \mathcal{V}_0} r_{i,0}^{-1} \tilde{\mathbf{b}}_i \tilde{\mathbf{b}}_i^T \text{ is small} \quad \text{and} \quad \tilde{\mathbf{P}}_1^{-1}(\bar{\mathcal{V}}) \approx \tilde{\mathbf{P}}_1^{-1}(\emptyset). \quad (7.63)$$

Conditions in (7.63) occur when $\text{diag}(r_{i,k}) \gg \mathbf{Q}_k$, i.e., when the controller (7.56) gives more weight to the input cost than state regulation. This condition is readily obtained from the definition of \mathbf{P}_k in (7.58). Figure 7.10 illustrates these observations by evaluating the bound from (7.62) for 100 random systems with $n = 100$ states controlled over a horizon of $N = 4$ steps with $\mathbf{Q} = \mathbf{I}$ and $r_{i,k} = \gamma$ (see Section 7.4.2.2 for details). Clearly, as the controller actions cost grows, i.e., as γ increases, α grows and Theorem 14 yields stronger guarantees. Note that this is a scenario of considerable practical value. It is well-known that if no restriction is imposed on the input cost, any controllable set of inputs can drive the system to any state in a single step (*dead-beat controller*). Hence, when $\text{diag}(r_{i,k}) \ll \mathbf{Q}_k$, the optimal value of the LQG only really differentiates between schedules that lead to controllable and uncontrollable input sets. On the other hand, careful scheduling can have a considerable impact when the actuation (or agent deployment) costs are high. This is also the case in which Theorem 14 provides stronger guarantees.

Observe also that since matrices are weighted by \mathbf{H}_k , the condition numbers of \mathbf{A} , $\mathbf{\Sigma}_0$, and \mathbf{W}_k play an important role in the guarantees. Indeed, if \mathbf{H}_k is poorly conditioned, there may be a large difference between the minimum and maximum eigenvalues in (7.62). This is illustrated in Figure 7.10a, which shows that when the decay rate of the system modes are considerably different, i.e., the state transition matrix has large condition number $\kappa(\mathbf{A})$, the guarantees from Theorem 14 worsen. Figure 7.10b illustrates this phenomenon for the LQG controller, showing how the value of α decreases when the process noise does not affect the states homogeneously, i.e., the condition

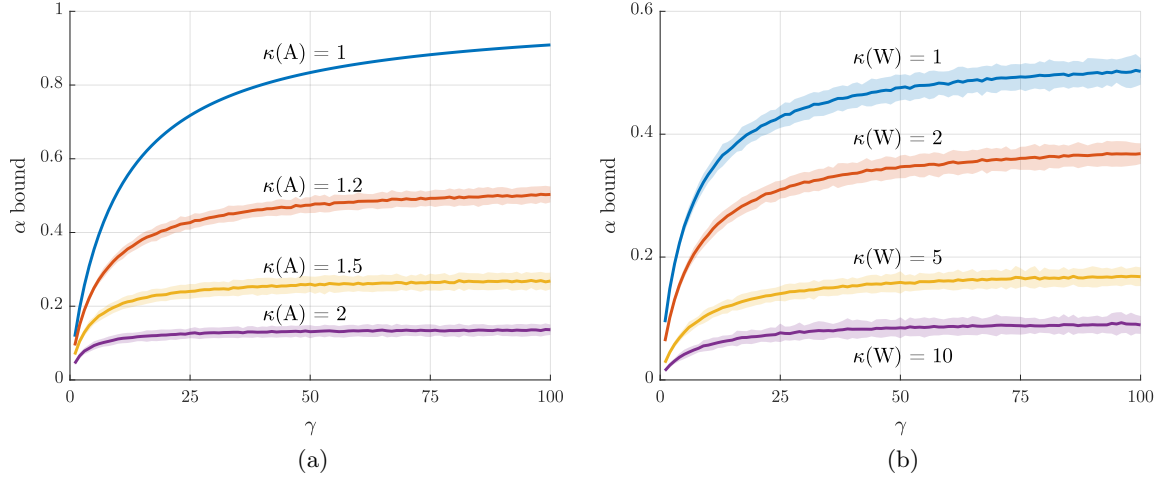


Figure 7.10: Bound on α from (7.59) for a schedule of length $N = 4$: (a) deterministic dynamical system ($\mathbf{W}_k = \mathbf{0}$, LQR) and (b) dynamical system with disturbance (LQG). Shaded regions span two standard deviations from the mean.

number of \mathbf{W}_k grows.

In the sequel, we provide details on the experiments presented in this section and illustrate the use of greedy control scheduling under multiple non-trivial constraints using an agent dispatching application.

7.4.2.2 Numerical experiments

We start by detailing the experiments in Figure 7.10. We considered dynamical systems with $n = 100$ states for which the elements of \mathbf{A} were drawn randomly from a standard Gaussian distribution, the input matrix $\mathbf{B} = \mathbf{I}$, i.e., direct state actuation, and $\mathbf{\Pi}_0 = 10^{-2}\mathbf{I}$. The state transition matrix \mathbf{A} was normalized so that its spectral radius is 1.1, i.e., the dynamical systems are unstable. In Figure 7.10a, the dynamical systems are deterministic, i.e., $\mathbf{W}_k = \mathbf{0}$. In Figure 7.10b, \mathbf{W}_k is a diagonal matrix whose elements were drawn uniformly at random from $[\kappa(\mathbf{W})^{-1}, 1]$ so that their condition number is $\kappa(\mathbf{W})$. The figures show the result for 100 independently drawn dynamical systems considering (PXIII) for $\mathbf{Q} = \mathbf{I}$, $r_{i,k} = \gamma$ for all i and k , and $N = 4$.

While Figure 7.10, along with Theorem 14, illustrate the wide range of parameters over which good performance certificates can be provided, it is worth noting that these are worst-case guarantees and that better results are common in practice. To illustrate this point, we evaluate the relative

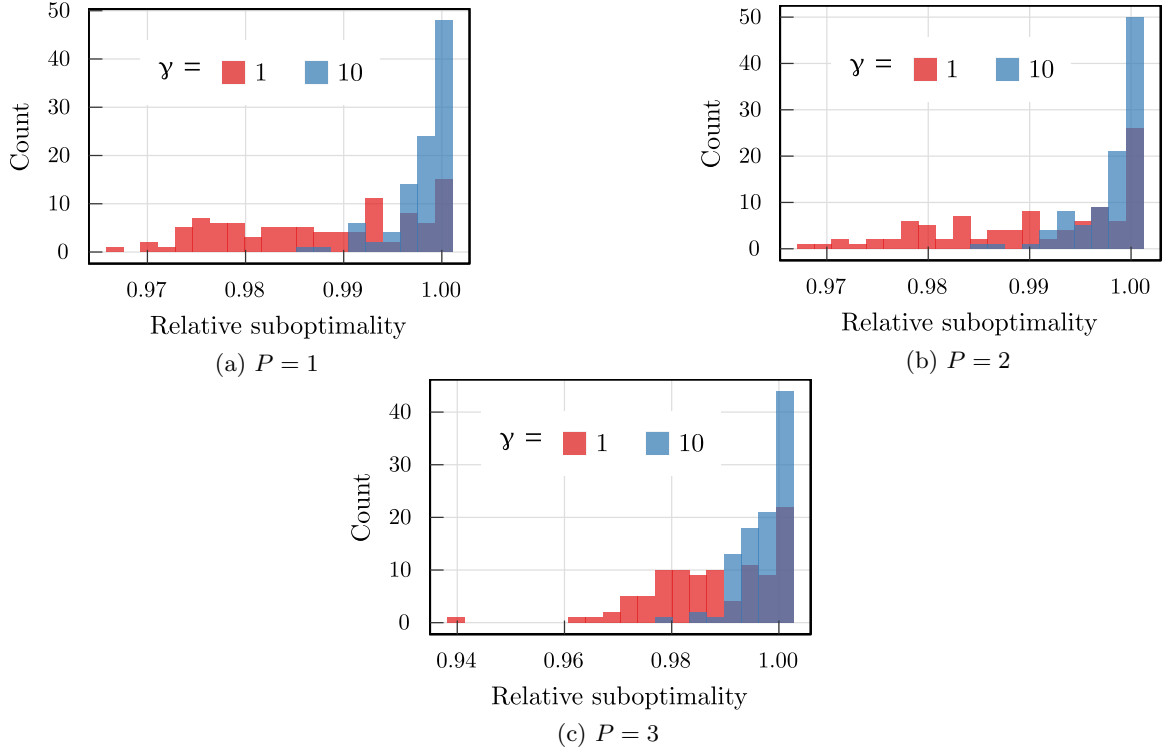


Figure 7.11: Relative suboptimality of greedy scheduling for different constraints (100 system realizations). (a) \mathcal{I}_1 (less than 2 actuators per time step); (b) \mathcal{I}_1 and \mathcal{I}_2 (less than 5 control actions over the horizon); (c) \mathcal{I}_1 , \mathcal{I}_2 , and \mathcal{I}_3 (inputs cannot be used on consecutive time slots).

suboptimality of greedily selected schedules over 100 system realizations. Explicitly, we evaluate

$$\nu^*(\mathcal{S}_g) = \frac{J(\mathcal{S}_g)}{J(\mathcal{S}^*)},$$

where \mathcal{S}_g and \mathcal{S}^* are the greedy (Algorithm 4) and optimal solutions of (PXIII) respectively. Since ν^* depends on the optimal schedule, which can only be obtained by exhaustive search, we restrict ourselves to smaller dynamical systems with $n = 7$ states. State space matrices \mathbf{A} and \mathbf{B} are as before, with $\kappa(\mathbf{A}) = 1.2$, and $\mathbf{W}_k = \mathbf{0}$ (LQR case). We again take $\mathbf{Q} = \mathbf{I}$ and $r_{i,k} = \gamma$.

In Figure 7.11, we design schedules for $N = 4$ steps horizons with different numbers of matroid constraints P . In Figure 7.11a, we select at most 2 actuators per time step ($\mathcal{I}_1 = \{\mathcal{S} \subseteq \bar{\mathcal{V}} \mid |\mathcal{S} \cap \mathcal{V}_k| \leq 2\}$). In addition to \mathcal{I}_1 , Figure 7.11b also imposes that at most 5 control actions can be taken over the horizon ($\mathcal{I}_2 = \{\mathcal{S} \subseteq \bar{\mathcal{V}} \mid |\mathcal{S}| \leq 5\}$). Finally, Figure 7.11c includes a restriction on using the

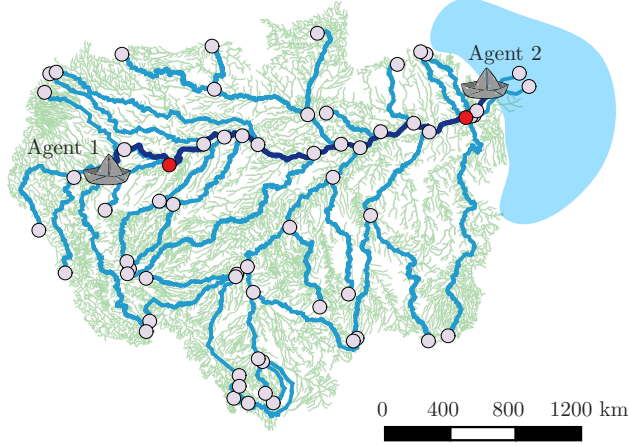


Figure 7.12: Amazon basin, amazon river (dark blue trace), system states (light grey circles), and chemical spill origins (red circles).

same input on consecutive time steps on top of \mathcal{I}_1 and \mathcal{I}_2 . Despite the fact that the lower bounds on ν^* (recall that J is a non-positive function) from Theorem 14 range from 0.03 to 0.25, Figure 7.11 show that the typical performance of greedy scheduling in practice is considerably better. Though it did not often find the optimal schedule (between 15% and 23% of the realizations for $\gamma = 1$ and 40% and 46% for $\gamma = 10$), the resulting schedule performances were at most 5% lower than the optimum. Note, however, that there exist specific dynamical systems for which greedy performs close to the guarantee in Theorem 14. Indeed, consider $\mathbf{A} = \mathbf{I}$ and

$$\mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 2 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Under constraint \mathcal{I}_1 , i.e., if we schedule up to 2 inputs per time step, over a $N = 2$ steps horizon with $\gamma = 100$, the guarantee in (7.61) yields $\nu^*(\mathcal{S}_g) \geq 0.392$, whereas in practice we achieve $\nu^*(\mathcal{S}_g) \approx 0.423$.

To provide a more practical application of these results, we illustrate the use of greedy control scheduling in an agent dispatching application to control the effect of spills on the Amazon river. Two agents navigate up and down the river (dark blue curve in Figure 7.12) over a preset route and use a chemical component to counteract damaging spills. Due to the limited capacity of each

vessel and to avoid overusage, each agent is allowed to dump the component at most 5 times. What is more, at least 2 steps must be allowed between each action so the crew has time to setup. The goal of this dispatch is to reduce how much of the spill reaches the ocean (light blue water mass in Figure 7.12).

The Amazon drainage basin, showed in green in Figure 7.12, covers 7.5 million km^2 and is composed of over 7000 tributaries. We use a simplified description of the basin (blue traces in Figure 7.12) obtained by smoothing the original map [212]. Using this river network, we construct a weighted directed tree \mathcal{G} whose vertices are the circles from in Figure 7.12 with the addition of a node midway between each circle. These vertices compose the $n = 127$ states \mathbf{x}_k of the dynamical system. In \mathcal{G} , two vertices are connected if water flows between them, i.e., if there is a blue line between them in Figure 7.12. We assume that all river flow toward the ocean, denoted in Figure 7.12 by a light blue water mass. The adjacency matrix of \mathcal{G} is then defined as

$$[\mathbf{G}]_{ij} = \begin{cases} \|\mathbf{z}_i - \mathbf{z}_j\|, & \text{if } (i, j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \quad (7.64)$$

where $\mathbf{z}_i \in \mathbb{R}^2$ is the position of node i on the map. We define its Laplacian as $\mathbf{L} = \mathbf{D} - \mathbf{G}$, where \mathbf{D} is a diagonal matrix whose elements are the sum of the columns of \mathbf{G} , and its symmetrized Laplacian as $\mathbf{L}' = \mathbf{D}' - (\mathbf{G} + \mathbf{G}^T)$, where \mathbf{D}' is a diagonal matrix whose elements are the sum of the columns of $\mathbf{G} + \mathbf{G}^T$.

Using these Laplacians, we define the state transition matrix of the dynamical system in (7.55) as

$$\mathbf{A} = 0.901 \exp(-\mathbf{L}\Delta t) + 0.099 \exp(-\mathbf{L}'\Delta t), \quad (7.65)$$

where $\Delta t = 5$ is the sampling period. The dynamics in (7.65) are a combination of two processes: the first term corresponds to the *advection* process by which water flows to the ocean and the second term corresponds to a *diffusion* process. The combination coefficients are chosen so that the system is marginally stable ($\|\mathbf{A}\| = 1$). In these experiments, we assume there is no process noise, i.e., $\mathbf{W}_k = \mathbf{0}$ for all k , and direct state actuation. The two spills are modeled by spikes in the initial state, namely \mathbf{x}_0 is zero except at the red nodes in Figure 7.12.

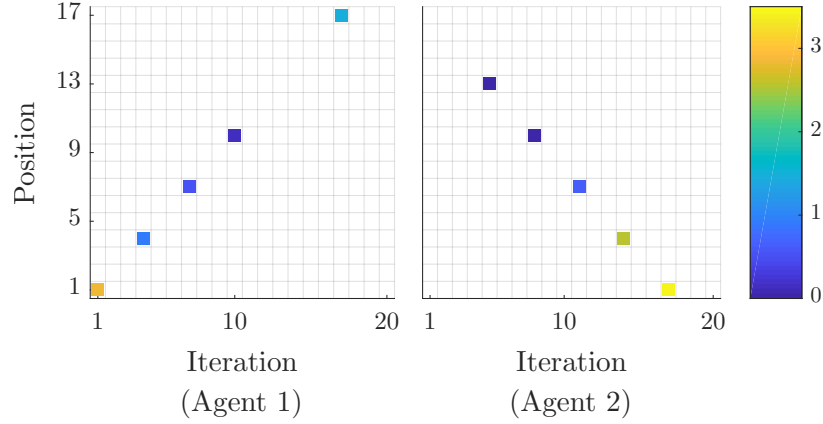


Figure 7.13: Greedy schedule and actuation energy of the spill control agents.

Agent 1 navigates the Amazon river (dark blue curve) left-to-right and Agent 2 navigates right-to-left (starting near the ocean). They can only actuate on their current position and are only allowed to do so on the states marked as circles in Figure 7.12. Thus, a centralized greedy scheduler designs a $N = 20$ steps action plan for the two agents taking into account their positions at each time step and their total number of actions and duty cycle constraints. To do so, it assumes $\mathbf{Q} = \mathbf{I}$ for both agents, but takes $r_{i,k}^{(1)} = 10$ for Agent 1 and $r_{i,k}^{(2)} = 20$ for Agent 2. In other words, Agent 1 is allowed to dump more cleaning component. The results of this dispatch are shown in Figures 7.13 and 7.14.

Notice that in the final schedule (Figure 7.13), the agents effectively act on top of the spills and towards the middle of the river to try and stop their spreading. Agent 1, in particular, saves cleaning reagent for one last action next to ocean. What is more, since it pays a lower price for actuation ($r_{i,k}^{(1)} < r_{i,k}^{(2)}$), it is able to use more reagent than Agent 2, which ends up concentrating its efforts in the beginning of its route (Figures 7.14). In the end, the vessels are able to mitigate the impact of the spills on the ocean waters, reducing contamination levels by 60% compared to the no-actuation solution. The final level achieved with these punctual actions is comparable to using an 64 agents that actuate every state at every time step with cost matrices $\mathbf{Q} = \mathbf{I}$ and $r_{i,k} = 2500$ (Figure 7.14).

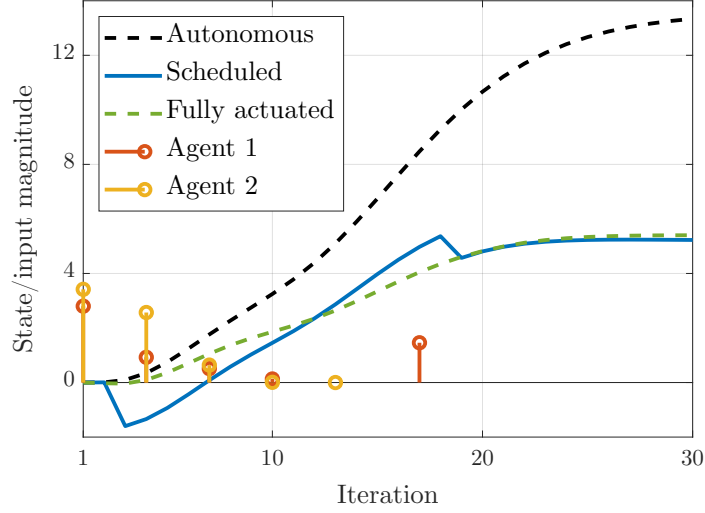


Figure 7.14: Agents actions and chemical concentrations in the ocean (light blue mass in Figure 7.12) for the autonomous system (no agent), greedy schedule, and full actuation.

7.4.3 Experimental design

We conclude this section with an experimental design application to showcase the multiset results we derived. Let \mathcal{E} be a pool of possible experiments. The outcome of experiment $e \in \mathcal{E}$ is a multivariate measurement $\mathbf{y}_e \in \mathbb{R}^{n_e}$ defined as

$$\mathbf{y}_e = \mathbf{A}_e \boldsymbol{\theta} + \mathbf{v}_e, \quad (7.66)$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a parameter vector with a prior distribution such that $\mathbb{E}[\boldsymbol{\theta}] = \bar{\boldsymbol{\theta}}$ and $\mathbb{E}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T = \mathbf{R}_\theta \succ 0$; \mathbf{A}_e is an $n_e \times p$ observation matrix; and $\mathbf{v}_e \in \mathbb{R}^{n_e}$ is a zero-mean random variable with arbitrary covariance matrix $\mathbf{R}_e = \mathbb{E} \mathbf{v}_e \mathbf{v}_e^T \succ 0$ that represents the experiment uncertainty. The $\{\mathbf{v}_e\}$ are assumed to be uncorrelated across experiments, i.e., $\mathbb{E} \mathbf{v}_e \mathbf{v}_f^T = \mathbf{0}$ for all $e \neq f$, and independent of $\boldsymbol{\theta}$. These experiments aim to estimate

$$\mathbf{z} = \mathbf{H} \boldsymbol{\theta}, \quad (7.67)$$

where \mathbf{H} is an $m \times p$ matrix. Appropriately choosing \mathbf{H} is important given that the best experiments to estimate $\boldsymbol{\theta}$ are not necessarily the best experiments to estimate \mathbf{z} . For instance, if $\boldsymbol{\theta}$ is to be used for classification, then \mathbf{H} can be chosen so as to optimize the design with respect to the output of

the classifier. Alternatively, transductive experimental design can be performed by taking \mathbf{H} to be a collection of data points from a test set [213]. Finally, $\mathbf{H} = \mathbf{I}$, the identity matrix, recovers the classical $\boldsymbol{\theta}$ -estimation case.

The experiments to be used in the estimation of \mathbf{z} are collected in a multiset \mathcal{D} called a *design*. Note that \mathcal{D} contains elements of \mathcal{E} with repetitions. Given a design \mathcal{D} , it is ready to compute an optimal Bayesian estimate $\hat{\mathbf{z}}_{\mathcal{D}}$. The estimation error of $\hat{\mathbf{z}}_{\mathcal{D}}$ is measured by the error covariance matrix $\mathbf{K}(\mathcal{D})$. An expression for the estimator and its error matrix in terms of the problem constants is given in the following proposition.

Proposition 17 (Bayesian estimator). *Let the experiments be defined as in (7.66). For $\mathbf{M}_e = \mathbf{A}_e^T \mathbf{R}_e^{-1} \mathbf{A}_e$ and a design $\mathcal{D} \in \mathcal{P}(\mathcal{E})$, the unbiased affine estimator of \mathbf{z} with the smallest error covariance matrix in the PSD cone is given by*

$$\hat{\mathbf{z}}_{\mathcal{D}} = \mathbf{H} \left[\mathbf{R}_{\theta}^{-1} + \sum_{e \in \mathcal{D}} \mathbf{M}_e \right]^{-1} \left[\sum_{e \in \mathcal{D}} \mathbf{A}_e^T \mathbf{R}_e^{-1} \mathbf{y}_e + \mathbf{R}_{\theta}^{-1} \bar{\boldsymbol{\theta}} \right]. \quad (7.68)$$

The corresponding error covariance matrix $\mathbf{K}(\mathcal{D}) = \mathbb{E}[(\mathbf{z} - \hat{\mathbf{z}}_{\mathcal{D}})(\mathbf{z} - \hat{\mathbf{z}}_{\mathcal{D}})^T \mid \boldsymbol{\theta}, \{\mathbf{M}_e\}_{e \in \mathcal{D}}]$ is given by the expression

$$\mathbf{K}(\mathcal{D}) = \mathbf{H} \left[\mathbf{R}_{\theta}^{-1} + \sum_{e \in \mathcal{D}} \mathbf{M}_e \right]^{-1} \mathbf{H}^T. \quad (7.69)$$

Proof. See appendix B.3.7. ■

The experimental design problem consists of selecting a design \mathcal{D} of cardinality at most k that minimizes the overall estimation error. This can be explicitly stated as the problem of choosing \mathcal{D} with $|\mathcal{D}| \leq k$ that minimizes the error covariance $\mathbf{K}(\mathcal{D})$ whose expression is given in (7.69). Note that (7.69) can account for unregularized (non-Bayesian) experimental design by removing \mathbf{R}_{θ} and using a pseudo-inverse [178]. However, the error covariance matrix is no longer monotone in this case—see Lemma 4. Providing guarantees for this scenario is the subject of future work.

The minimization of the PSD matrix $\mathbf{K}(\mathcal{D})$ in experimental design is typically attempted using scalarization procedures generically known as alphabetical design criteria, the most common of which are A-, D-, and E-optimal design [214]. These are tantamount to selecting different figures of merit to compare the matrices $\mathbf{K}(\mathcal{D})$. Our focus in this section is mostly on A- and E-optimal designs,

but we also consider D-optimal designs for comparison. A design \mathcal{D} with k experiments is said to be A-optimal if it minimizes the estimation MSE which is given by the trace of the covariance matrix,

$$\underset{|\mathcal{D}| \leq k}{\text{minimize}} \quad \text{Tr} \left[\mathbf{K}(\mathcal{D}) \right] - \text{Tr} \left[\mathbf{H} \mathbf{R}_\theta \mathbf{H}^T \right] \quad (\text{P-A})$$

Notice that is customary to say a design is A-optimal when $\mathbf{H} = \mathbf{I}$ in (P-A), whereas the notation V-optimal is reserved for the case when \mathbf{H} is arbitrary [214]. We do not make this distinction here for conciseness.

A design is E-optimal if instead of minimizing the MSE as in (P-A), it minimizes the largest eigenvalue of the covariance matrix $\mathbf{K}(\mathcal{D})$, i.e.,

$$\underset{|\mathcal{D}| \leq k}{\text{minimize}} \quad \lambda_{\max} \left[\mathbf{K}(\mathcal{D}) \right] - \lambda_{\max} \left[\mathbf{H} \mathbf{R}_\theta \mathbf{H}^T \right]. \quad (\text{P-E})$$

Since the trace of a matrix is the sum of its eigenvalues, we can think of (P-E) as a robust version of (P-A). While the design in (P-A) seeks to reduce the estimation error in all directions, the design in (P-E) seeks to reduce the estimation error in the worst direction. Equivalently, given that $\lambda_{\max}(\mathbf{X}) = \max_{\|\mathbf{u}\|_2=1} \mathbf{u}^T \mathbf{X} \mathbf{u}$, we can interpret (P-E) with $\mathbf{H} = \mathbf{I}$ as minimizing the MSE for an adversarial choice of \mathbf{z} .

A D-optimal design is one in which the objective is to minimize the log-determinant of the estimator's covariance matrix,

$$\underset{|\mathcal{D}| \leq k}{\text{minimize}} \quad \log \det \left[\mathbf{K}(\mathcal{D}) \right] - \log \det \left[\mathbf{H} \mathbf{R}_\theta \mathbf{H}^T \right]. \quad (\text{P-D})$$

The motivation for using the objective in (P-D) is that the log-determinant of $\mathbf{K}(\mathcal{D})$ is proportional to the volume of the confidence ellipsoid when the data are Gaussian. Note that the trace, maximum eigenvalue, and determinant of $\mathbf{H} \mathbf{R}_\theta \mathbf{H}^T$ in (P-A), (P-E), and (P-D) are constants and do not affect the respective optimization problems. They are subtracted so that the objectives vanish when $\mathcal{D} = \emptyset$, i.e., so they are normalized set functions.

Remark 6. Besides its intrinsic value as a minimizer of the volume of the confidence ellipsoid, (P-D) is often used as a surrogate for (P-A), when A-optimality (MSE) is considered the appropriate metric.

It is important to point out that this is only justified when the problem has some inherent structure that suggests the minimum volume ellipsoid is somewhat symmetric. Otherwise, since the volume of an ellipsoid can be reduced by decreasing the length of a single principal axis, using (P-D) can lead to designs that perform well—in the MSE sense—along a few directions of the parameter space and poorly along all others. Formally, this can be seen by comparing the variation of the log-determinant and trace functions with respect to the eigenvalues of the PSD matrix \mathbf{K} ,

$$\frac{\partial \log \det(\mathbf{K})}{\partial \lambda_j(\mathbf{K})} = \frac{1}{\lambda_j(\mathbf{K})} \quad \text{and} \quad \frac{\partial \text{Tr}(\mathbf{K})}{\partial \lambda_j(\mathbf{K})} = 1.$$

The gradient of the log-determinant is largest in the direction of the smallest eigenvalue of the error covariance matrix. In contrast, the MSE gives equal weight to all directions of the space. The latter therefore yields balanced, whereas the former tends to flatten the confidence ellipsoid unless the problem has a specific structure.

Although the problem formulations in (P-A), (P-E), and (P-D) are integer programs known to be NP-hard, the use of greedy methods for their solution is widespread with good performance in practice. In the case of D-optimal design, this is justified theoretically because the objective of (P-D) is supermodular, which implies greedy methods are $(1 - e^{-1})$ -optimal [166, 168, 196]. The objectives in (P-A) and (P-E), on the other hand, are *not* be supermodular in general [197, 215, 216] and it is not known why their greedy optimization yields good results in practice—conditions for the MSE to be supermodular exist but are restrictive [216]. We can settle these questions using the results from Section 7.3.

Notice that the objectives of (P-A) and (P-E) are normalized by construction and that, due since \mathbf{K} in (7.69) matches the form of (7.12), they are also monotone decreasing due to Lemma 4. To employ the results from Theorems 11 and 12, suffices to refine the bounds on α/ϵ from Theorems 6 and 7. We start by showing that the objective of (P-A) is α -supermodular as in Definition 8.

Theorem 15. *The objective of (P-A) is α -supermodular with*

$$\alpha(a, b) \geq \frac{1}{\kappa(\mathbf{H})^2} \cdot \frac{\lambda_{\min}[\mathbf{R}_{\theta}^{-1}]}{\lambda_{\max}[\mathbf{R}_{\theta}^{-1}] + a \cdot \ell_{\max}}, \quad \text{for all } b \in \mathbb{N}, \quad (7.70)$$

where $\ell_{\max} = \max_{e \in \mathcal{E}} \lambda_{\max}(\mathbf{M}_e)$, $\mathbf{M}_e = \mathbf{A}_e^T \mathbf{R}_e^{-1} \mathbf{A}_e$, and $\kappa(\mathbf{H}) = \sigma_{\max} / \sigma_{\min}$ is the ℓ_2 -norm condition number of \mathbf{H} , with σ_{\max} and σ_{\min} denoting the largest and smallest singular values of \mathbf{H} respectively.

Proof. See appendix B.3.8. ■

Theorem 15 bounds the α -supermodularity of the objective of (P-A) in terms of the condition number of \mathbf{H} , the prior covariance matrix, and the measurements SNR. To facilitate the interpretation of this result, let the SNR of the e -th experiment be $\gamma_e = \text{Tr}[\mathbf{M}_e]$ and suppose $\mathbf{R}_\theta = \sigma_\theta^2 \mathbf{I}$, $\mathbf{H} = \mathbf{I}$, and $\gamma_e \leq \gamma$ for all $e \in \mathcal{E}$. Then, for $\ell = k$ greedy iterations, (7.70) implies

$$\bar{\alpha} \geq \frac{1}{1 + 2k\sigma_\theta^2\gamma},$$

for $\bar{\alpha}$ as in Theorem 11. This deceptively simple bound reveals that the MSE behaves as a supermodular function at low SNRs. Formally, $\alpha \rightarrow 1$ as $\gamma \rightarrow 0$. In contrast, the performance guarantee from Theorem 15 degrades in high SNR scenarios. In this case, however, greedy methods are expected to give good results since designs yield similar estimation errors. The greedy solution of (P-A) also approaches the $1 - 1/e$ guarantee when the prior on $\boldsymbol{\theta}$ is concentrated ($\sigma_\theta^2 \ll 1$), i.e., when the problem is heavily regularized.

These observations also hold for a generic \mathbf{H} as long as it is well-conditioned. Even if $\kappa(\mathbf{H}) \gg 1$, we can replace \mathbf{H} by $\tilde{\mathbf{H}} = \mathbf{D}\mathbf{H}$ for some diagonal matrix $\mathbf{D} \succ 0$ without affecting the design, since \mathbf{z} is arbitrarily scaled. The scaling \mathbf{D} can be designed to minimize the condition number of $\tilde{\mathbf{H}}$ by leveraging preconditioning and balancing methods [217, 218].

Proceeding, we derive guarantees for E-optimal designs using the refined concept of ϵ -supermodularity in Definition 9.

Theorem 16. *The cost function of (P-E) is ϵ -supermodular with*

$$\epsilon(a, b) \leq (b - a) \sigma_{\max}(\mathbf{H})^2 \lambda_{\max}(\mathbf{R}_\theta)^2 \ell_{\max}, \quad (7.71)$$

where $\ell_{\max} = \max_{e \in \mathcal{E}} \lambda_{\max}(\mathbf{M}_e)$, $\mathbf{M}_e = \mathbf{A}_e^T \mathbf{R}_e^{-1} \mathbf{A}_e$, and $\sigma_{\max}(\mathbf{H})$ is the largest singular value of \mathbf{H} .

Proof. See appendix B.3.9. ■

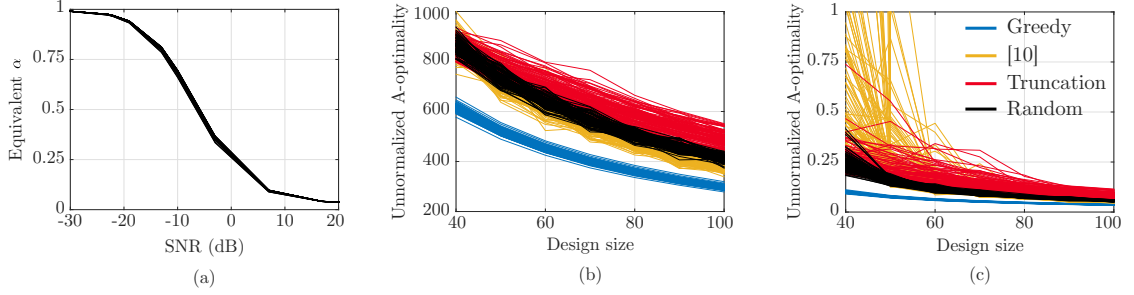


Figure 7.15: A-optimal design: (a) Thm. 15; (b) A-optimality (low SNR); (c) A-optimality (high SNR). The plots show the unnormalized A-optimality value for clarity.

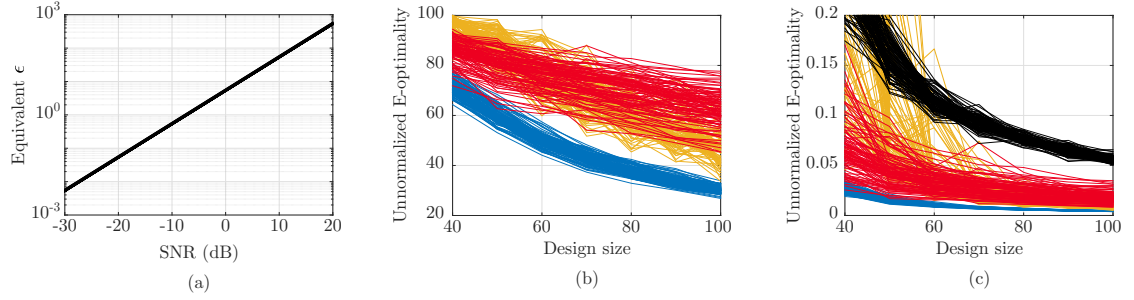


Figure 7.16: E-optimal design: (a) Thm. 16; (b) E-optimality (low SNR); (c) E-optimality (high SNR). The plots show the unnormalized E-optimality value for clarity.

Under the same assumptions as above, Theorem 16 gives

$$\bar{\epsilon} \leq 2k\sigma_\theta^4\gamma,$$

for $\bar{\epsilon}$ as in Theorem 12. Thus, $\epsilon \rightarrow 0$ as $\gamma \rightarrow 0$. In other words, the behavior of the objective of (P-E) approaches that of a supermodular function as the SNR decreases. The same holds for concentrated priors, i.e., $\lim_{\sigma_\theta^2 \rightarrow 0} \bar{\epsilon} = 0$. Once again, it is worth noting that when the SNRs of the experiments are large, almost every design has the same E-optimal performance as long as the experiments are not too correlated. Thus, greedy design is also expected to give good results under these conditions.

Notice that the proofs of Theorems 15 and 16 suggest that better bounds can be found when the designs are constructed without replacement, i.e., when only one of each experiment is allowed in the design (set function optimization case).

To illustrate these results, we draw the elements of \mathbf{A}_e from an i.i.d. zero-mean Gaussian random

variable with variance $1/p$ and $p = 20$. The noise $\{\mathbf{v}_e\}$ are also Gaussian random variables with $\mathbf{R}_e = \sigma_v^2 \mathbf{I}$. We take $\sigma_v^2 = 10^{-1}$ in high SNR and $\sigma_v^2 = 10$ in low SNR simulations. The experiment pool contains $|\mathcal{E}| = 200$ experiments.

Starting with A-optimal design, we display the bound from Theorem 15 in Figure 7.15a for multivariate measurements of size $n_e = 5$ and designs of size $k = 40$. Here, “equivalent α ” is the single $\hat{\alpha}$ that gives the same near-optimal certificate (7.22) as using (7.70). As expected, $\hat{\alpha}$ approaches 1 as the SNR decreases. In fact, for -10 dB is already close to 0.75 which means that by selecting a design of size $\ell = 55$ we would be within $1 - 1/e$ of the optimal design of size $k = 40$. Figures 7.15b and 7.15c compare greedy A-optimal designs with the convex relaxation of (P-A) in low and high SNR scenarios. The designs are obtained from the continuous solutions using the hard constraint, with replacement method of [219] and a simple design truncation as in [130]. Therefore, these simulations consider univariate measurements ($n_e = 1$). For comparison, a design sampled uniformly at random with replacement from \mathcal{E} is also presented. Note that, as mentioned before, the performance difference across designs is small for high SNR—notice the scale in Figures 7.15c and 7.16c—, so that even random designs perform well.

For the E-optimality criterion, the bound from Theorem 16 is shown in Figure 7.16a, again for multivariate measurements of size $n_e = 5$ and designs of size $k = 40$. Once again, “equivalent ϵ ” is the single value $\hat{\epsilon}$ that yields the same guarantee as using (7.71). In this case, the bound degradation in high SNR is more pronounced. This reflects the difficulty in bounding the approximate supermodularity of the E-optimality cost function. Still, Figures 7.16b and 7.16c show that greedy E-optimal designs have good performance when compared to convex relaxations or random designs. Note that, though it is not intended for E-optimal designs, we again display the results of the sampling post-processing from [219]. In Figure 7.16b, the random design is omitted due to its poor performance.

For a more concrete application, consider the problem of cold-start in recommender systems. Recommender systems use semi-supervised learning methods to predict user ratings based on few rated examples. These methods are useful, for instance, to streaming service providers who are interested in using predicted ratings of movies to provide recommendations. For new users, these systems suffer from a “cold-start problem,” which refers to the fact that it is hard to provide accurate

recommendations without knowing a user’s preference on at least a few items. For this reason, services explicitly ask users for ratings in initial surveys before emitting any recommendation. Selecting which movies should be rated to better predict a user’s preferences can be seen as an experimental design problem. In the following example, we use a subset of the EachMovie dataset [220] to illustrate how greedy experimental design can be applied to address this problem.

We randomly selected a training and test set containing 9000 and 3000 users respectively. Each experiment in \mathcal{E} represents a movie ($|\mathcal{E}| = 1622$) and the observation vector \mathbf{A}_e collects the ratings of movie e for each user in the training set. The parameter $\boldsymbol{\theta}$ is used to express the rating of a new user in term of those in the training set. Our hope is that we can extrapolate the observed ratings, i.e., $\{y_e\}_{e \in \mathcal{D}}$, to obtain the rating for a movie $f \notin \mathcal{D}$ as $\hat{y}_f = \mathbf{A}_f \hat{\boldsymbol{\theta}}$. Since the mean absolute error (MAE) is commonly used in this setting, we choose to work with the A-optimality criterion. We also let $\mathbf{H} = \mathbf{I}$ and take a non-informative prior $\bar{\boldsymbol{\theta}} = \mathbf{0}$ and $\mathbf{R}_\theta = \sigma_\theta^2 \mathbf{I}$ with $\sigma_\theta^2 = 100$.

As expected, greedy A-optimal design is able to find small sets of movies that lead to good prediction. For $k = 10$, for example, $\text{MAE} = 2.3$, steadily reducing until $\text{MAE} < 1.8$ for $k \geq 35$. These are considerably better results than a random movie selection, for which the MAE varies between 2.8 and 3.3 for k between 10 and 50. Instead of focusing on the raw ratings, we may be interested in predicting the user’s favorite genre. This is a challenging task due to the heavily skewed dataset. For instance, 32% of the movies are dramas whereas only 0.02% are animations. Still, we use the simplest possible classifier by selecting the category with highest average estimated ratings. By using greedy design, we can obtain a misclassification rate of approximately 25% by observing 100 ratings, compared to over 45% error rate for a random design.

Chapter 8

Risk-aware minimum mean square error estimation

Critical applications require that stochastic decisions be made not only on the basis of minimizing average losses, but also safeguarding against less probable, though possibly catastrophic, events. Examples appear naturally in many areas, including wireless industrial control, energy [112, 113], finance [114–116], robotics [117, 118], LIDAR [119], and networking [120]. In such cases, the ultimate goal is to obtain *risk-aware decision policies* that hedge against statistically significant extreme losses, even at the cost of slightly sacrificing performance under nominal operating conditions.

To illustrate this effect, consider the problem of estimating a random state x from observations corrupted by state-dependent noise, namely $y \mid x \sim \mathcal{N}(x, 9x^2)$ (see Section 8.4). Recall that we are now in the inference setting where we seek to estimate the value of x based on observations y and a conditional model $y \mid x$. In this setting, either small or large values of y provide highly ambiguous evidence, since they can arise from either small or large values of x . This is corroborated by Figure 8.1, which displays the posterior distribution $p_{x|y}$ for two values of y . While the MMSE estimator may incur severe losses, the risk-aware estimator we develop in this work (shown in red) hedges against this observation ambiguity, therefore avoiding extreme prediction errors. This risk-averse behavior is achieved by *biasing estimates* towards the tail of the posterior $p_{x|y}$. Although the risk-aware estimator may incur larger losses on average, it performs statistically more consistently

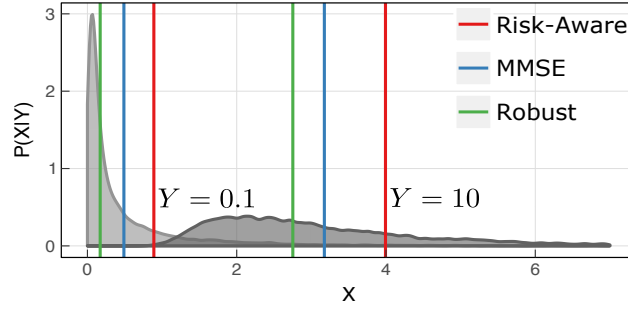


Figure 8.1: Comparison between risk-neutral and risk-aware estimates.

across realizations of y (see, e.g., risk curve in Figure 8.2). It is also worth contrasting risk-awareness with statistical robustness, whose goal is to protect against deviations from a nominal model. Robust estimators (green lines in Figure 8.1) promote insensitivity to tail events, which they designate as statistically insignificant (“outliers”). On the other hand, estimators resulting from risk-aware formulations treat these events as statistically significant, though relatively infrequent (see Table 8.1).

Over the last three decades, risk-aware optimization has grown increasingly popular and has been studied in the contexts of both decision making and learning [122, 221–226]. In risk-aware optimization, expectations are replaced by more general functionals, called *risk measures* [227], whose purpose is to quantify the statistical volatility of random losses, as well as mean performance. Popular examples include mean-variance functionals [114, 227], mean-semideviations [122], and Conditional Value-at-Risk (CVaR) [123].

In Bayesian estimation, risk awareness is typically achieved by replacing the classical quadratic cost with its exponentiation [228–232]. However, although sometimes effective, this approach is not without limitations. First, the need for finiteness of the moment generating function of the quadratic cost excludes heavy-tailed distributions, which are precisely those that incur high risk. Second, the exponential approach does not provide an interpretable way to control the trade-off between mean performance and risk, making it hard to use in settings where explicit risk levels must be met. Third, it does not result in a simple, general solution as in classical MMSE estimation, hindering its practical applicability. Finally, it does not effectively quantify *observation-induced risk*, inherent in problems where measurements provide ambiguous evidence.

Table 8.1: Classification of statistical uncertainty.

| Uncertainty | | |
|---------------|----------|-------------|
| | Frequent | Infrequent |
| Significant | Model | <i>Risk</i> |
| Insignificant | Noise | Outliers |

8.1 The risk-constrained MMSE problem

Following Section 2.2.2.3, we consider here a constrained inference approach to risk-awareness using the formulation in (P-RISK). Explicitly, let $(\Omega, \mathcal{F}, \mathbf{p})$ be a probability space and consider an arbitrary pair of random elements $\mathbf{x} : \Omega \rightarrow \mathbb{R}^d$ and $\mathbf{y} : \Omega \rightarrow \mathbb{R}^k$ on (Ω, \mathcal{F}) . Our goal is to estimate \mathbf{x} from a *single* realization of \mathbf{y} in a Bayesian setting, namely by assuming knowledge of the joint probability distribution $\mathbf{p}_{(\mathbf{x}, \mathbf{y})}$. As we illustrated before, we may conveniently think of \mathbf{y} as *observations* based on which we wish to make predictions about the *hidden state* \mathbf{x} .

An established approach to this problem is to choose an estimator ϕ as a solution to the *stochastic variational* MMSE program

$$\underset{\phi \in \mathcal{H}}{\text{minimize}} \quad \mathbb{E} \left[\|\mathbf{x} - \phi(\mathbf{y})\|^2 \right] \quad (\text{P-MMSE})$$

where \mathcal{H} denotes the space of \mathbf{y} measurable vector-valued functions $\phi : \sigma(\mathbf{y}) \rightarrow \mathbb{R}^p$. Problem (P-MMSE) is well-understood under rather general conditions. In fact, if we assume \mathbf{x} is integrable over the sub- σ -algebra of \mathcal{F} generated by \mathbf{y} , then an optimal solution to (P-MMSE) is obtained from by any conditional expectation of \mathbf{x} relative to \mathbf{y} , i.e.,

$$\phi_{\text{MMSE}}^*(\mathbf{y}) = \mathbb{E}[\mathbf{x}|\mathbf{y}] \quad (8.1)$$

However, despite the simplicity of MMSE estimation, its effectiveness is often questionable. Indeed, minimizing the squared error $\|\mathbf{x} - \phi(\mathbf{y})\|_2^2$ *in expectation* does *not* provide stability or robustness, in the sense that statistically significant variability of the resulting *optimal prediction error* is uncontrolled. In other words, the MMSE problem (P-MMSE) is *risk-neutral*. This has important consequences from a practical perspective, since the error realization $\|\mathbf{x} - \phi_{\text{MMSE}}^*(\mathbf{y})\|_2^2$

experienced in practice may be far from the expected value $\mathbb{E}[\|\mathbf{x} - \phi_{\text{MMSE}}^*(\mathbf{y})\|_2^2]$, or even the predictive statistic $\mathbb{E}\{\|\mathbf{x} - \phi_{\text{MMSE}}^*(\mathbf{y})\|_2^2 \mid \mathbf{y}\}$. It is then clear that achieving small error variability is at least as desirable as achieving minimal errors on average.

Motivated by the previous discussion, we consider a nontrivial variation of the risk-neutral MMSE problem (P-MMSE) that strikes a balance between mean performance and risk, namely (P-RISK). For ease of reference, we restate the problem here:

$$\begin{aligned} & \underset{\phi \in \mathcal{H}}{\text{minimize}} && \mathbb{E} \left[\|\mathbf{x} - \phi(\mathbf{y})\|^2 \right] \\ & \text{subject to} && \mathbb{E} \left[\text{var} \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \right) \right] \leq \varepsilon \end{aligned} \tag{P-RISK}$$

where \mathcal{H} is now the space of square-integrable, \mathbf{y} measurable functions, i.e., $\phi(\mathbf{y}) \in L_2$,

$$\text{var} \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \right) \triangleq \mathbb{E} \left[\left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 - \mathbb{E} [\|\mathbf{x} - \phi(\mathbf{y})\|^2] \right)^2 \mid \mathbf{y} \right] \tag{8.2}$$

is the predictive variance of the squared loss $\|\mathbf{x} - \phi(\mathbf{y})\|^2$ with respect to \mathbf{y} , and $\varepsilon > 0$ is a fixed risk tolerance. In words, problem (P-RISK) constrains the expected predictive variance of the quadratic cost $\|\mathbf{x} - \phi(\mathbf{y})\|^2$, known in the statistics literature as the *unexplained component* of the variance due to the law of total variance. Hence, the constraint quantifies the uncertainty of the MMSE prediction of the quadratic cost actually achieved by the estimate $\phi(\mathbf{y})$ on the basis of the observations \mathbf{y} . Naturally, $\mathbb{E} \left[\text{var} \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \right) \right]$ is a measure of risk. Therefore, we suggestively refer to the task fulfilled by problem (P-RISK) as *risk-aware MMSE estimation*.

Problem (P-RISK) confines the search for an optimal MMSE estimator within the family of estimators exhibiting risk (in the sense described above) within a tolerance ε . Naturally, an optimal solution to the risk-aware (P-RISK) in general achieves larger MSE as compared to the risk-neutral (P-MMSE). However, the statistical variability of the squared errors achieved by the former will be explicitly controlled resulting in more stable statistical prediction.

8.2 Convex Variational QCQP Reformulation

The risk-aware MMSE problem (P-RISK) is a non-convex, infinite dimensional problem, making it rather challenging to study, let alone solve. Nevertheless, it admits an equivalent Quadratically Constrained Quadratic Program (QCQP) reformulation under mild condition stated below.

Assumption 7. The conditional moment $\mathbb{E} [\|\mathbf{x}\|_2^3 \mid \mathbf{y}] \in L_2$, i.e., it is square-integrable with respect to the sub- σ -algebra of \mathcal{F} generated by \mathbf{y} .

In words, Assumption 7 states that the third-order moment filter $\mathbb{E} [\|\mathbf{x}\|_2^3 \mid \mathbf{y}]$ has finite energy. Using Assumption 7, problem(P-RISK) may be conveniently reformulated as shown next.

Lemma 6 (QCQP Reformulation of (P-RISK)). *Define the posterior covariance*

$$\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \triangleq \mathbb{E} [(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}])^T \mid \mathbf{y}] \succeq 0. \quad (8.3)$$

Under Assumption 7, (P-RISK) is well-defined and equivalent to the convex variational QCQP

$$\begin{aligned} P_{RISK}^* = \min_{\phi \in \mathcal{H}} \quad & \frac{1}{2} \mathbb{E} \left[\|\phi(\mathbf{y})\|^2 - 2 \mathbb{E}[\mathbf{x} \mid \mathbf{y}]^T \phi(\mathbf{y}) + \mathbb{E}[\|\mathbf{x}\|^2 \mid \mathbf{y}] \right] \\ \text{subject to} \quad & \mathbb{E} \left[\phi(\mathbf{y})^T \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \phi(\mathbf{y}) - (\mathbb{E}[\|\mathbf{x}\|^2 \mid \mathbf{y}] - \mathbb{E}[\|\mathbf{x}\|^2 \mid \mathbf{y}] \mathbb{E}[\mathbf{x} \mid \mathbf{y}]^T \mathbb{E}[\mathbf{x} \mid \mathbf{y}])^T \phi(\mathbf{y}) \right] \\ & \leq \frac{\varepsilon - \mathbb{E} \left[\text{var} \left(\|\mathbf{x}\|^2 \mid \mathbf{y} \right) \right]}{4} \end{aligned} \quad (\text{PXIV})$$

where all expectations and involved operations are well-defined.

Proof. See Appendix B.4.1. ■

Lemma 6 is very useful because it shows the equivalence of problem (P-RISK) to the convex QCQP (PXIV), which is well-defined and favorably structured. In particular, this reformulation allows us to effectively study (P-RISK) by looking at its variational Lagrangian dual. As we discuss next, working in the dual domain allows us to solve (P-RISK) in closed-form. Of course, such a closed form is important, not only because it provides an analytical, textbook-level solution to a functional risk-aware problem, which happens rather infrequently in such settings, but also because

the solution itself provides intuition, highlights connections, and enables comparison of (P-RISK) with its risk-neutral counterpart (P-MMSE).

8.3 Risk-Aware MMSE Estimators

In our development, we exploit a variational version of Slater's condition, which is one the most widely used constraint qualifications in both deterministic and stochastic optimization.

Assumption 8. Given $\varepsilon > 0$, there exists $\phi^\dagger \in \mathcal{H}$ such that $\mathbb{E}\{\|\mathbf{x} - \phi^\dagger(\mathbf{y})\|_2^2\} < \infty$ and

$$\mathbb{E}\left[\text{var}\left(\|\mathbf{x} - \phi^\dagger(\mathbf{y})\|^2 \mid \mathbf{y}\right)\right] < \varepsilon.$$

Under both Assumptions 7 and 8, it follows that the QCQP (PXIV) satisfies Slater's condition [130]. Then, it must be the case that $\mathbb{E}\left[\text{var}\left(\|\mathbf{x}\|^2 \mid \mathbf{y}\right)\right] < \infty$, otherwise Assumption 8 could not hold. Furthermore, (PXIV) must be feasible.

To proceed, define the *variational Lagrangian* of the *primal problem* (PXIV) as

$$\begin{aligned} L(\phi, \mu) \triangleq & \frac{1}{2}\mathbb{E}[\|\phi(\mathbf{y})\|_2^2 - 2(\mathbb{E}[\mathbf{x} \mid \mathbf{y}])^T \phi(\mathbf{y}) + \mathbb{E}[\|\mathbf{x}\|_2^2 \mid \mathbf{y}]] \\ & + \mu \mathbb{E}[\phi(\mathbf{y})^T \Sigma_{\mathbf{x}|\mathbf{y}} \phi(\mathbf{y}) - (\mathbb{E}[\|\mathbf{x}\|_2^2 \mathbf{x} \mid \mathbf{y}]] \\ & - \mathbb{E}[\|\mathbf{x}\|_2^2 \mid \mathbf{y}] \mathbb{E}[\mathbf{x} \mid \mathbf{y}])^T \phi(\mathbf{y})] - \mu \frac{\varepsilon - \mathbb{E}\left[\text{var}\left(\|\mathbf{x}\|^2 \mid \mathbf{y}\right)\right]}{4}, \end{aligned} \quad (8.4)$$

where $\mu \in \mathbb{R}_+$ is a multiplier associated with the constraint of (PXIV). Its *dual function* is accordingly defined as

$$d(\mu) \triangleq \inf_{\phi \in \mathcal{H}} L(\phi, \mu). \quad (8.5)$$

If $P_{\text{RISK}}^* \in [0, \infty]$ denotes the optimal value of problem (PXIV), it is true that $d \leq P_{\text{RISK}}^*$ on \mathbb{R}_+ .

Then, the optimal value of the concave, *dual problem*

$$D_{\text{RISK}}^* = \max_{\mu \geq 0} d(\mu) \quad (\text{PXV})$$

is the tightest under-estimate of P_{RISK}^* .

Exploiting Assumptions 7 and 8, we may now formulate the following fundamental theorem, which establishes that the convex variational problem (PXIV) exhibits zero duality gap. This essentially follows as an application of standard results in variational Lagrangian duality; see, for instance, [233, Sec. 8.3, Thm. 1]. The proof is therefore omitted.

Theorem 17 (Strong duality of (PXIV)). *Suppose Assumptions 7 and 8 are in effect. Then, strong duality holds for problem (PXIV), that is, $0 \leq D_{RISK}^* \equiv P_{RISK}^* < \infty$. Additionally, the set of dual optimal solutions, $\arg \max_{\mu \geq 0} d(\mu)$, is nonempty. Further, if $\phi^*(\mathbf{y})$ is primal optimal for (PXIV), it follows that $\phi^*(\mathbf{y}) \equiv \phi^*(\mathbf{y}) \in \arg \min_{\phi \in \mathcal{H}} L(\phi, \mu^*)$, where $0 \leq \mu^* \in \arg \max_{\mu \geq 0} d(\mu)$.*

Leveraging Theorem 17, it is possible to show that, under Assumptions 7 and 8, the QCQP (PXIV) and, therefore, the original risk-aware MMSE problem (P-RISK), admit a common closed-form solution. In this respect, we have the next theorem, which constitutes the main result of this chapter.

Theorem 18 (Closed-form solution of (PXIV)). *Suppose that Assumptions 7 and 8 are in effect. Then,*

$$\phi^*(\mathbf{y}) = \mathbf{K}^{-1}(\mathbb{E}[\mathbf{x}|\mathbf{y}] + \mathbf{b}) \quad (8.6)$$

is an optimal solution to problem (PXIV) for

$$\mathbf{K} = \mathbf{I} + 2\mu^* \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \quad (8.7a)$$

$$\mathbf{b} = \mu^* \left(\mathbb{E}[\|\mathbf{x}\|^2 \mathbf{x} | \mathbf{y}] - \mathbb{E}[\|\mathbf{x}\|^2 | \mathbf{y}] \mathbb{E}[\mathbf{x} | \mathbf{y}] \right) \quad (8.7b)$$

where μ^ is any optimal solution to the concave dual problem (PXV). Additionally, the optimal risk-aware filter $\phi^*(\mathbf{y})$ is unique \mathbf{p} -a.e.*

Theorem 18 provides the solution of (P-RISK) as an explicit expression for the risk-aware MMSE estimator $\phi^*(\mathbf{y})$, defined in terms of the dual optimal solution μ^* , a scalar value. The latter *always* exists, thanks to Theorem 17, and may be computed by leveraging our knowledge of the distribution $\mathbf{p}_{(\mathbf{x}, \mathbf{y})}$ via either some gradient-based method, probabilistic bisection, or empirically. Note that the dual function d in (8.5) is a concave function on the positive line.

The fact that a closed-form optimal solution of (PXIV) exists is remarkable and provides insight

into the intrinsic structure of *constrained* Bayesian risk-aware estimation. Most importantly, it enables an explicit comparison of the optimal risk-aware filter $\phi^*(\mathbf{y})$ with its risk-neutral counterpart. Indeed, by looking at the explicit form of the optimal risk-aware filter $\phi^*(\mathbf{y})$, we readily see that it is a function involving the MMSE estimator $\mathbb{E}[\mathbf{x}|\mathbf{y}]$, its predictive covariance matrix $\Sigma_{\mathbf{x}|\mathbf{y}}$, as well as the second and third order filters $\mathbb{E}[\|\mathbf{x}\|^2 | \mathbf{y}]$ and $\mathbb{E}[\|\mathbf{x}\|^2 \mathbf{x} | \mathbf{y}]$. All of these quantities are elementary and, in principle, can be evaluated by utilizing a single observation of \mathbf{y} and exploiting our knowledge of the conditional measure $\mathbf{p}_{\mathbf{x}|\mathbf{y}}$, just as in risk-neutral MMSE estimation.

Also, we see that $\phi^*(\mathbf{y})$ may be regarded as a *biased MMSE estimator*, drawing parallels to *James-Stein estimators*, obtained in another statistical context. While James-Stein estimators use biasing to reduce the MSE, $\phi^*(\mathbf{y})$ reduces the risk. Therefore, optimal risk aversion, in the sense of problem (P-RISK), may be interpreted as the result of bias injection in MMSE estimators.

Additionally, we observe that the solution is *regularized*, in the sense that the term $2\mu^*\Sigma_{\mathbf{x}|\mathbf{y}}$ is diagonally loaded with an identity matrix. As a result, $\phi^*(\mathbf{y})$ is always well-defined and numerically stable. In fact, if the tolerance ε is large enough, $\mu^* = 0$ and $\phi^*(\mathbf{y}) = \mathbb{E}[\mathbf{x}|\mathbf{y}]$. But this is not the only case where the two estimators match. In fact, there exists a family of models for which risk-neutral and risk-aware MMSE estimation coincide.

Theorem 19 (When do Risk-Neutral/Aware Filters Coincide?). *Suppose that the conditional measure $\mathbf{p}_{\mathbf{x}|\mathbf{y}}$ is such that*

$$\mathbb{E}\left[\left([\mathbf{x}]_i - \mathbb{E}[\mathbf{x}]_i | \mathbf{y}\right)^2 (\mathbf{x} - \mathbb{E}[\mathbf{x} | \mathbf{y}]) | \mathbf{y}\right] = \mathbf{0}, \quad \text{for } i = 1, \dots, d. \quad (8.8)$$

Then, under Assumptions 7 and 8, the risk-neutral MMSE estimator is also risk-aware for any feasible value of $\varepsilon > 0$, i.e., $\phi^(\mathbf{y}) \equiv \mathbb{E}[\mathbf{x} | \mathbf{y}]$ for any value of μ^* . This occurs, for instance, whenever $\mathbf{p}_{\mathbf{x}|\mathbf{y}}$ is joint Gaussian.*

Proof of Theorem 19. Start by expanding (8.8) to obtain (note that all involved conditional expectations in the expression above assume finite values due to Assumption 7)

$$\begin{aligned} \mathbf{0} &= \mathbb{E}\left[\left([\mathbf{x}]_i - \mathbb{E}[\mathbf{x}]_i | \mathbf{y}\right)^2 (\mathbf{x} - \mathbb{E}[\mathbf{x} | \mathbf{y}]) | \mathbf{y}\right] \\ &= \mathbb{E}\left[\left([\mathbf{x}]_i^2 \mathbf{x}\right) | \mathbf{y}\right] - \mathbb{E}[\mathbf{x} | \mathbf{y}] \mathbb{E}[\|\mathbf{x}\|_i^2 | \mathbf{y}] - 2\mathbb{E}[\mathbf{x}]_i | \mathbf{y}] \mathbb{E}[\mathbf{x}]_i | \mathbf{y}] + 2\left(\mathbb{E}[\mathbf{x}]_i | \mathbf{y}\right)^2 \mathbb{E}[\mathbf{x} | \mathbf{y}], \end{aligned}$$

which implies that

$$\mathbb{E}[(\mathbf{x}_i^2 \mathbf{x}) | \mathbf{y}] = \mathbb{E}[\mathbf{x} | \mathbf{y}] \mathbb{E}[\mathbf{x}_i^2 | \mathbf{y}] + 2\mathbb{E}[\mathbf{x}_i | \mathbf{y}] \left(\mathbb{E}[\mathbf{x} \mathbf{x}_i | \mathbf{y}] - \mathbb{E}[\mathbf{x}_i | \mathbf{y}] \mathbb{E}[\mathbf{x} | \mathbf{y}] \right). \quad (8.9)$$

Therefore, we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}\|^2 | \mathbf{y}] - \mathbb{E}[\|\mathbf{x}\|^2] \mathbb{E}[\mathbf{x} | \mathbf{y}] &= \sum_{i=1}^d \mathbb{E}[\mathbf{x}_i^2 \mathbf{x} | \mathbf{y}] - \mathbb{E}[\mathbf{x}_i^2 | \mathbf{y}] \mathbb{E}[\mathbf{x} | \mathbf{y}] \\ &= 2 \sum_{i=1}^d \mathbb{E}[\mathbf{x}_i | \mathbf{y}] \left(\mathbb{E}[\mathbf{x} \mathbf{x}_i | \mathbf{y}] - \mathbb{E}[\mathbf{x}_i | \mathbf{y}] \mathbb{E}[\mathbf{x} | \mathbf{y}] \right) \\ &= 2\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \mathbb{E}[\mathbf{x} | \mathbf{y}], \end{aligned}$$

which in turn implies that (8.7b) yields, for every $\mu^* \geq 0$, $\mathbf{b}^\dagger = 2\mu^* \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \mathbb{E}[\mathbf{x} | \mathbf{y}]$. Substituting into (8.6) we get Suppose that Assumptions 7 and 8 are in effect. Then,

$$\phi^*(\mathbf{y}) = \mathbf{K}^{-1} \left(\mathbb{E}[\mathbf{x} | \mathbf{y}] + 2\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \mathbb{E}[\mathbf{x} | \mathbf{y}] \right) = \mathbf{K}^{-1} (\mathbf{I} + 2\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \mathbb{E}[\mathbf{x} | \mathbf{y}]$$

which from (8.7a) implies that

$$\phi^*(\mathbf{y}) = \mathbb{E}[\mathbf{x} | \mathbf{y}] \quad \mathbf{p} - \text{a.e.} \quad (8.10)$$

To conclude, notice that the fact that (8.9) is true when $\mathbf{p}_{\mathbf{x}|\mathbf{y}}$ is multivariate Gaussian follows from the straightforward application of Stein's Lemma on all pairs of jointly Gaussian random variables $([\mathbf{x}]_i, [\mathbf{x}]_j)$, $i, j = 1, \dots, d$, conditioned on \mathbf{y} . ■

8.4 Applications

In this section, we illustrate the performance of the risk-aware MMSE estimator in comparison with that of the usual, risk-neutral MMSE estimator. We evaluate the behavior of the estimator in (8.6) in two scenarios. The first consists of the problem of estimating an exponentially distributed hidden state x , $\mathbb{E}[x] = 2$, from the observation $y = x + v$, where v is a zero-mean Gaussian random variable conditionally independent on x with variance $\mathbb{E}[v^2] = 9x^2$. This model is often referred to as *state-dependent noise*. In the second scenario, the goal is to jointly estimate the random vector $\mathbf{x} = [z \ h]^T$

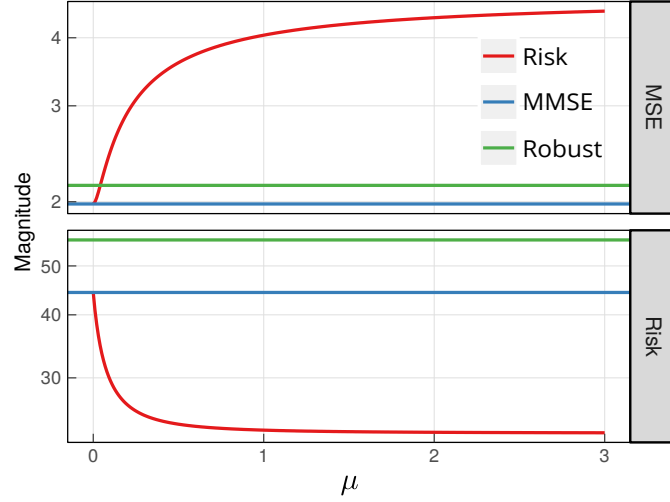


Figure 8.2: Mean squared error and risk for different values of μ in the state-dependent noise scenario.

from the observation $y = hz + w$, where z is a zero-mean Gaussian random variable with variance $\mathbb{E}[z^2] = 2$, h has a Rayleigh distribution with rate 2, and w is a zero-mean Gaussian noise with variance $\mathbb{E}[w^2] = 10^{-1}$. This scenario is prototypical for estimation problems in communications, where z is the signal of interest and h represents the channel fading. Throughout the simulations, we also show results for the risk-neutral MMSE estimator and the Minimum Mean Absolute Error (MMAE) estimator, or, equivalently, the conditional median relative to the respective observations, the latter being used as an example of a robust location parameter estimator.

In Figure 8.1, we saw that the risk-aware estimator yields larger estimates than the MMSE estimator, in order to account for the certain statistical ambiguities of the state-dependent noise model. Though this difference may seem extreme in some instances, e.g., for small values of y (as in Figure 8.1), it is in fact quite effective in reducing the conditional variance. Indeed, for $y = 0.1$, the risk-aware estimator in Figure 8.1 optimally reduces the (conditional) risk by approximately 26% as compared to the risk of the risk-neutral estimator and this is achieved by sacrificing average performance, also by a factor of 26%. Of course, this is only one of the operation points of the risk-aware estimator. In Figure 8.2, we show results for different values of μ , where we average over the distribution of y . Observe that the risk-aware estimator obtained using the constrained optimization problem (P-RISK) achieves a sharp trade-off between average performance (that is, MSE) and risk, which can be tuned according to the needs of the application. Additionally, note

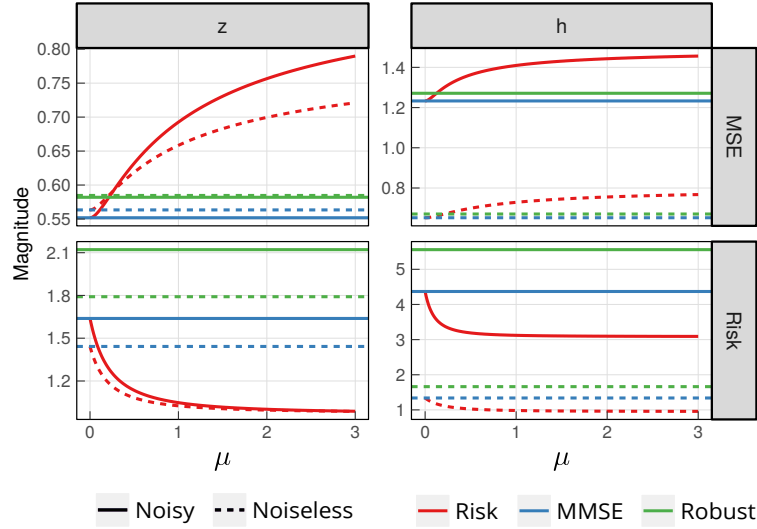


Figure 8.3: Mean squared error and risk for different values of μ in the communication scenario.

that the decrease in risk is considerably faster than the increase in MSE.

Interestingly, a similar phenomenon is observed in the communication scenario (Figure 8.3). Again, the risk-aware estimator displays a much faster initial rate of decrease with respect to μ than the rate at which the MSE increases. This is more pronounced in the estimation of the component z , for which the risk-aware estimator can provide reductions of almost 60% in risk for a 35% increase in average MSE. Note that, as per Theorem 19, the Gaussian noise has indeed no bearing on risk-awareness, as evidenced by the performance in the noiseless case, i.e., for $w = 0$ (dashed lines). To achieve the behavior of Fig. 8.3, the risk-aware estimator overestimates both z and h as compared to the MMSE estimator, as illustrated in Fig. 8.4. In fact, for small values of y , the former hedges against the event of a deep fade ($h \approx 0$) by maintaining its estimates for z away from zero.

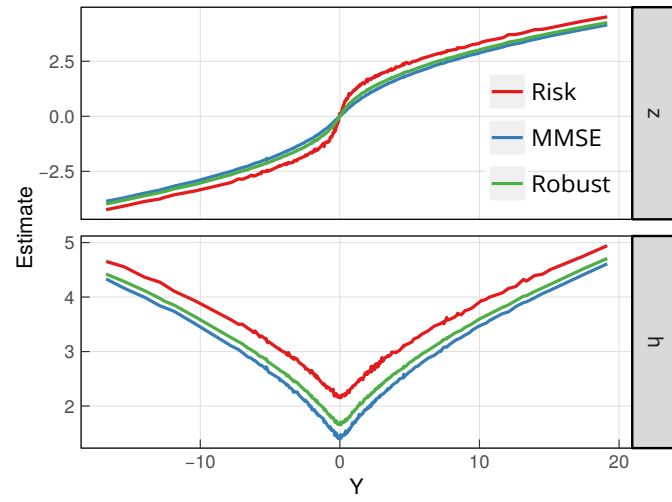


Figure 8.4: Risk-aware, MMSE, and robust estimates of z and h in the communication scenario for different values of Y .

Chapter 9

Concluding remarks

This thesis set out to tackle learning and inference under requirements by leveraging constrained statistical optimization. It started by addressing the issue of constrained learning by formalizing the notion of PACC learning and showing that, from a learning theoretical point-of-view, constrained learning is as hard as its unconstrained learning (Chapter 3). To overcome the challenges involved in solving ECRM problems, it put forward a dual learning rule and proved that it also enable PACC learning (Chapter 4). These results were illustrated in fairness and robustness applications (Chapter 5), showing that if we learning under requirements can be achieved as long as it is possible to solve classical, unconstrained learning tasks.

The second part of this thesis dealt with the issue of inference under unconventional constraints. First, it tackled the issue of fitting sparse nonlinear functional models by defining a new class of infinite dimensional, non-convex optimization problems, namely SFPs. By showing that, under mild conditions, SFPs are strongly dual, it enabled them to be solved exactly and efficiently, leading to solutions to problems such as nonlinear line spectrum estimation and robust functional regression (Chapter 6). Then, it proceeded to investigate inference problems involving quadratic objectives and combinatorial constraints on the observations, involving cardinality (maximum number of observations) or matroids (e.g., restrictions on sequential observations). By developing a theory of approximately supermodular minimization and showing that cost functions such as the MSE and the LQR/LQG objectives are approximately supermodular, it derived near-optimal certificates for

greedy solutions of these problems. These results were showcased in applications involving graph signal sampling, actuator scheduling, and experimental design (Chapter 7). Finally, it leveraged functional optimization and duality theory to obtain a closed-form solution to the problem of risk-aware minimum MSE estimation (Chapter 8).

Appendix A

Proofs of Part I

A.1 Proof of Theorem 1

Start by noticing from the definition of PACC learnability [more specifically, from (3.2) in Def. 3] that any PACC learnable class \mathcal{H} is necessarily PAC learnable.

To prove the converse, recall that if \mathcal{H} is PAC learnable, then \mathcal{H} has finite VC dimension [19, Sec. 3.4]. More precisely, for $N > C\zeta^{-1}(\epsilon, \delta, d_{\mathcal{H}})$, where C is an absolute constant and ζ^{-1} is as in (3.4), and any bounded function g it holds with probability $1 - \delta$ that

$$\left| \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [g(\phi(\mathbf{x}), y)] - \frac{1}{N} \sum_{n=1}^N g(\phi(\mathbf{x}_n), y_n) \right| \leq \epsilon \quad (\text{A.1})$$

for all function $\phi \in \mathcal{H}$, distributions \mathcal{D} , and samples $(\mathbf{x}_n, y_n) \sim \mathcal{D}$. Now, let $\hat{\phi}^*$ be a solution of (P-ECRM). From (A.1) and the boundedness hypothesis on ℓ_0 , we immediately obtain that $\hat{\phi}^*$ is probably approximately optimal as in (3.2). Additionally, $\hat{\phi}^*$ must be feasible for (P-ECRM). Hence,

$$\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(\hat{\phi}^*(\mathbf{x}_{n_i}), y_{n_i}) \leq c_i, \quad \text{for } i = 1, \dots, m, \quad \text{and} \quad (\text{A.2a})$$

$$\ell_j(\hat{\phi}^*(\mathbf{x}_{n_j}), y_{n_j}) \leq c_j, \quad \text{for all } n_j = 1, \dots, N_j \text{ and } j = m+1, \dots, m+q. \quad (\text{A.2b})$$

To show (A.2) implies that $\hat{\phi}^*$ is a probably approximately feasible, note that we can write, using (A.1),

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\hat{\phi}^*(\mathbf{x}), y)] \leq \frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(\hat{\phi}^*(\mathbf{x}_{n_i}), y_{n_i}) + \epsilon \quad \text{and} \quad (\text{A.3a})$$

$$\begin{aligned} \Pr_{(\mathbf{x}, y) \sim \mathfrak{D}_j} [\ell_j(\hat{\phi}^*(\mathbf{x}), y) \leq b_j] &= \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_j} [\mathbb{I} [\ell_j(\hat{\phi}^*(\mathbf{x}), y) \leq b_j]] \\ &\geq \frac{1}{N_j} \sum_{n_j=1}^{N_j} \mathbb{I} [\ell_j(\hat{\phi}^*(\mathbf{x}_{n_j}), y_{n_j}) \leq b_j] - \epsilon, \end{aligned} \quad (\text{A.3b})$$

each of which hold with probability $1 - \delta$ over the samples $(\mathbf{x}_{n_i}, y_{n_i})$ as long as $N_i > C\zeta^{-1}(\epsilon, \delta, d_{\mathcal{H}})$. Combining (A.2) and (A.3) we conclude that, with probability $1 - (m + q)\delta$, it holds simultaneously that

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\hat{\phi}^*(\mathbf{x}), y)] &\leq c_i + \epsilon \quad \text{and} \\ \ell_j(\hat{\phi}^*(\mathbf{x}), y) &\leq c_j \quad \text{for all } (\mathbf{x}, y) \in \mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}, \end{aligned}$$

where each \mathcal{K}_j is a set of \mathfrak{D}_j -measure at least $1 - \epsilon$.

Hence, if \mathcal{H} is PAC learnable, then there exists N such that, if $\hat{\phi}^*$ is a solution of (P-ECRM) obtained using $N_i \geq N$ samples from each \mathfrak{D}_i , then $\hat{\phi}^*$ is probably approximately optimal as in (3.2) and probably approximately feasible as in (3.3). \square

A.2 Proof of Theorem 2

As we have argued before, we cannot rely on the duality between (PIV) and ($\widehat{\text{D}}$ -CSL) to obtain this result because of its non-convexity. Hence, this proof proceeds directly from (P-CSO) by applying three transformations that yield ($\widehat{\text{D}}$ -CSL), but whose approximation and estimation errors can be controlled. First, we obtain the dual problem of (P-CSO) and show that this transformation incurs in no error. This stems from the convexity of (P-CSO) under Assumptions 1 and 2 and is a straightforward strong duality result from semi-infinite programming theory (Proposition 18). Second, we approximate the function class \mathcal{H} using the finite dimensional parametrization $f_{\boldsymbol{\theta}}$ and bound the

approximation error ϵ_0 (Proposition 19). Third, we obtain (\widehat{D} -CSL) by replacing the expectations with their empirical versions. Since the problem is now unconstrained, we can use classical learning theory to evaluate the estimation error ϵ (Proposition 20). We then combine these results to obtain Theorem 2.

We begin by showing that Assumptions 1–3 imply that (P-CSO) is strongly dual:

Proposition 18. *Under Assumptions 1–3, the semi-infinite program (P-CSO) and the saddle-point problem (D-CSL) are strongly dual, i.e., $P^* = D^*$.*

Proof. Start by noticing that (P-CSO) can be equivalently formulated as

$$\begin{aligned}
P^* = \min_{\phi \in \mathcal{H}} \quad & \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y)] \\
\text{subject to} \quad & \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] \leq c_i, \quad i = 1, \dots, m, \\
& \ell_j(\phi(\mathbf{x}), y) p_{\mathfrak{D}_j}(\mathbf{x}, y) \leq c_j p_{\mathfrak{D}_j}(\mathbf{x}, y), \quad (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, \\
& j = m + 1, \dots, n.
\end{aligned} \tag{PXVI}$$

In fact, both problems have the same objective function and feasibility set. Indeed, if $\mathfrak{D} > 0$, the transformation in the pointwise constraints is vacuous. On the other hand, when \mathfrak{D} vanishes, the constraint is not enforced in (PXVI). However, neither is it in (P-CSO) since the pointwise constraint need not hold on sets of \mathfrak{D} -measure zero. Note that this is different from satisfying the constraint with probability \mathfrak{D} .

From Assumptions 1 and 2 we obtain that (PXVI) is a semi-infinite convex program. What is more, Assumption 3 implies it has a strictly feasible solution $\phi' = f_{\theta'}$. This constraint qualification, sometimes known as *Slater's condition*, implies that it is strongly dual, i.e., that $P^* = D^*$ [234]. ■

A.2.1 The approximation gap

While there is no duality gap between (P-CSO) and (D-CSL), the latter remains a variational problem. The next step is there to approximate the functional space \mathcal{H} by $\mathcal{P} = \{f_{\theta} \mid \theta \in \mathbb{R}^p\}$, the space induced by the finite dimensional parametrization f_{θ} . Thus, (D-CSL) becomes the finite

dimensional problem

$$D_\theta^* = \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m, \lambda_j \in L_{1,+}} \min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_\theta(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \triangleq L(f_\theta, \boldsymbol{\mu}, \boldsymbol{\lambda}). \quad (\text{D}_\theta\text{-CSL})$$

Since $\mathcal{P} \subseteq \mathcal{H}$ (Assumption 2), it is clear that $D_\theta^* \geq D^* = P^*$. Yet, if the parametrization is rich enough, we should expect the gap $D_\theta^* - P^*$ to be small. This intuition is formalized in the following proposition.

Proposition 19. *Let $\boldsymbol{\theta}^*$ achieve the saddle-point in (D $_\theta$ -CSL). Under Assumptions 1–3, $f_{\boldsymbol{\theta}^*}$ is a feasible, near-optimal solution of (P-CSO). Explicitly,*

$$P^* \leq D_\theta^* \leq P^* + \left(1 + \|\tilde{\boldsymbol{\mu}}^*\|_1 + \sum_{j=m+1}^{m+q} \|\tilde{\lambda}_j^*\|_{L_1} \right) L\nu, \quad (\text{A.4})$$

for P^* and D_θ^* defined as in (P-CSO) and (D $_\theta$ -CSL) respectively and where $(\tilde{\boldsymbol{\mu}}^*, \tilde{\boldsymbol{\lambda}}^*)$ are the dual variables of (P-CSO) with the constraints tightened to $c_i - M\nu$ for $i = 0, \dots, m+q$.

Proof. See Appendix A.2.4. ■

A.2.2 The estimation gap

All that remains, is to turn the statistical Lagrangian (4.3) into the empirical (4.4). The incurred estimation error is described in the next proposition.

Proposition 20. *Let $\hat{\boldsymbol{\theta}}^*$ achieve the saddle-point in ($\hat{\text{D}}$ -CSL). Under Assumptions 1–3, it holds with probability $1 - \delta$ over the samples drawn from the distributions \mathfrak{D}_i that*

$$|D_\theta^* - \hat{D}^*| \leq V_{N_0}, \quad (\text{A.5})$$

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y)] \leq c_i + V_{N_i}, \quad \text{and} \quad (\text{A.6})$$

$$\ell_j(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y) \leq c_j \quad \text{for } (\mathbf{x}, y) \in \mathcal{K}_j, \quad (\text{A.7})$$

where $\mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}$ is a set of \mathfrak{D}_j -measure at least $1 - V_{N_j}$ for all $j = m+1, \dots, m+q$, where V_N is defined as in (4.6).

Proof. See appendix A.2.5. ■

A.2.3 The PACC solution

The proof concludes by combining the parametrization and estimation gap results from Propositions 19 and 20. Namely, notice that (A.6) and (A.7) imply that the minimizer $\hat{\boldsymbol{\theta}}^*$ that achieves the saddle-point in ($\widehat{\mathbf{D}}$ -CSL) is probably approximately feasible [see (4.2)] for (P-CSO). Then, combining (A.4) and (A.5) using the triangle inequality yields the near-PACC gap from Def. 4. Fixing N such that $V_N \leq \epsilon$ yields the result in Theorem 2. □

A.2.4 Proof of Proposition 19: The Approximation Gap

We first prove that $f_{\boldsymbol{\theta}^*}$ is feasible for (P-CSO) and then bound the gap between $D_{\boldsymbol{\theta}}^*$ and P^* .

Feasibility. Suppose that $f_{\boldsymbol{\theta}^*}$ is infeasible. Then, there exists at least one $i > 0$ such that $\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\boldsymbol{\theta}^*}(\mathbf{x}), y)] > c_i$ or $\ell_i(f_{\boldsymbol{\theta}^*}(\mathbf{x}), y) > c_i$ over some set $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{Y}$ of positive \mathfrak{D}_i -measure. Since $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are unbounded above, we obtain that $D_{\boldsymbol{\theta}}^* \rightarrow +\infty$. However, Assumptions 1 and 3 imply that $D_{\boldsymbol{\theta}}^* < +\infty$. Indeed, consider the dual function

$$\begin{aligned} d(\boldsymbol{\mu}, \boldsymbol{\lambda}) &= \min_{\boldsymbol{\theta} \in \mathcal{H}} L_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \\ &= \min_{\boldsymbol{\theta} \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(f_{\boldsymbol{\theta}}(\mathbf{x}), y)] + \sum_{i=1}^m \mu_i [\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y)] - c_i] \\ &\quad + \sum_{j=m+1}^{m+q} \int \lambda_j(\mathbf{x}, y) [\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - c_j] p_{\mathfrak{D}_j}(\mathbf{x}, y) d\mathbf{x} dy, \end{aligned}$$

for the Lagrangian defined in ($\mathbf{D}_{\boldsymbol{\theta}}$ -CSL). Using the fact that ℓ_0 is B -bounded (Assumption 1) and that there exists a strictly feasible $\boldsymbol{\theta}'$ (Assumption 3), $d(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is upper bounded by

$$\begin{aligned} d(\boldsymbol{\mu}, \boldsymbol{\lambda}) &\leq \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(f_{\boldsymbol{\theta}'}(\mathbf{x}), y)] + \sum_{i=1}^m \mu_i [\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\boldsymbol{\theta}'}(\mathbf{x}), y)] - c_i] \\ &\quad + \sum_{j=m+1}^{m+q} \int \lambda_j(\mathbf{x}, y) [\ell_j(f_{\boldsymbol{\theta}'}(\mathbf{x}), y) - c_j] p_{\mathfrak{D}_j}(\mathbf{x}, y) d\mathbf{x} dy < B, \end{aligned}$$

where we used the fact that $\mu_i \geq 0$ and $\lambda_j \geq 0$ \mathfrak{D}_j -a.e. Hence, it must be that f_{θ^*} is feasible for (P-CSO).

Near-optimality. First, recall that under Assumptions 1–3, (P-CSO)–(D-CSL) form a strongly dual pair of mathematical programs (Proposition 18). For the Lagrangian in (4.3), we therefore obtain the saddle-point relation

$$L(\phi^*, \mu', \lambda') \leq \max_{\mu, \lambda} \min_{\phi \in \mathcal{H}} L(\phi, \mu, \lambda) = D^* = P^* = \min_{\phi \in \mathcal{H}} \max_{\mu, \lambda} L(\phi, \mu, \lambda) \leq L(\phi', \mu^*, \lambda^*) \quad (\text{A.8})$$

which holds for all $\phi' \in \mathcal{H}$, $\mu' \in \mathbb{R}_+^m$, and $\lambda'_j \in L_{1,+}$, where ϕ^* is a solution of (P-CSO) and (μ^*, λ^*) are solutions of (D-CSL). We omit the spaces that (μ, λ) belong to for conciseness. Additionally, we have from (D $_{\theta}$ -CSL) that

$$D_{\theta}^* \geq \min_{\theta \in \mathbb{R}^p} L(\theta, \mu, \lambda), \quad \text{for all } \mu \in \mathbb{R}_+^m \text{ and } \lambda_j \in L_{1,+}. \quad (\text{A.9})$$

Immediately, we obtain the lower bound in (4.12). Explicitly,

$$D_{\theta}^* \geq \min_{\theta \in \mathbb{R}^p} L(\theta, \mu, \lambda) \geq \min_{\phi \in \mathcal{H}} L(\phi, \mu^*, \lambda^*) = P^*, \quad (\text{A.10})$$

where the second inequality comes from the fact that $\mathcal{P} \subseteq \mathcal{H}$ (Assumption 2).

The upper bound is obtained by relating the parameterized dual problem (D $_{\theta}$ -CSL) to a perturbed (tightened) version of the original (P-CSO). To do so, start by adding and subtracting $L(\phi, \mu, \lambda)$ from (D $_{\theta}$ -CSL) to get

$$\begin{aligned} D_{\theta}^* &= \max_{\mu, \lambda} \min_{\theta \in \mathbb{R}^p} L(\phi, \mu, \lambda) + \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(f_{\theta}(\mathbf{x}), y) - \ell_0(\phi(\mathbf{x}), y)] \\ &\quad + \sum_{i=1}^m \mu_i \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\theta}(\mathbf{x}), y) - \ell_i(\phi(\mathbf{x}), y)] \\ &\quad + \sum_{j=m+1}^{m+q} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_j} [\lambda_j(\mathbf{x}, y) (\ell_j(f_{\theta}(\mathbf{x}), y) - \ell_j(\phi(\mathbf{x}), y))], \end{aligned} \quad (\text{A.11})$$

where we wrote the integral against $p_{\mathfrak{D}_j}$ as an expectation for conciseness. Then, using the fact that ℓ_i is M -Lipschitz continuous (Assumption 1), we bound the expectations in the first two terms

of (A.11) as

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} \left[\ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_i(\phi(\mathbf{x}), y) \right] &\leq \mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} \left[\left| \ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_i(\phi(\mathbf{x}), y) \right| \right] \\ &\leq M \mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} \left[\left| f_{\boldsymbol{\theta}}(\mathbf{x}) - \phi(\mathbf{x}) \right| \right], \text{ for } i = 0, \dots, m.\end{aligned}\tag{A.12}$$

To bound the last expectation in (A.11), we first use Hölder's inequality to get

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} \left[\lambda_j(\mathbf{x}, y) \left(\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_j(\phi(\mathbf{x}), y) \right) \right] &\leq \\ &\mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} [\lambda_j(\mathbf{x}, y)] \left\| \ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_j(\phi(\mathbf{x}), y) \right\|_{L_{\infty}},\end{aligned}$$

where we recall that $\|g\|_{L_{\infty}}$ is the essential supremum of $|g|$. Then, the M -Lipschitz continuity of ℓ_j (Assumption 1) implies that

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} \left[\lambda_j(\mathbf{x}, y) \left(\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_j(\phi(\mathbf{x}), y) \right) \right] &\leq \\ &M \|f_{\boldsymbol{\theta}}(\mathbf{x}) - \phi(\mathbf{x})\|_{L_{\infty}} \mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} [\lambda_j(\mathbf{x}, y)].\end{aligned}\tag{A.13}$$

Using (A.12) and (A.13), together with the approximation property of the class \mathcal{H} (Assumption 2), we upper bound the minimum over $\boldsymbol{\theta}$ in (A.11) to obtain

$$D_{\boldsymbol{\theta}}^* \leq \max_{\boldsymbol{\mu}, \boldsymbol{\lambda}} L(\phi, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \left[1 + \sum_{i=1}^m \mu_i + \sum_{j=m+1}^{m+q} \mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} [\lambda_j(\mathbf{x}, y)] \right] M\nu.\tag{A.14}$$

Notice that since (A.14) holds uniformly for all $\phi \in \mathcal{H}$, it also holds for the minimizer

$$D_{\boldsymbol{\theta}}^* \leq \min_{\phi \in \mathcal{H}} \max_{\boldsymbol{\mu}, \boldsymbol{\lambda}} L(\phi, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \left[1 + \sum_{i=1}^m \mu_i + \sum_{j=m+1}^{m+n} \mathbb{E}_{(\mathbf{x},y) \sim \mathfrak{D}_j} [\lambda_j(\mathbf{x}, y)] \right] M\nu \triangleq \tilde{P}^*\tag{A.15}$$

and that the right-hand side of (A.15), namely \tilde{P}^* , is in fact a perturbed version of (P-CSO). Hence, we obtain another saddle-point relation similar to (A.8) relating \tilde{P}^* , and consequently $D_{\boldsymbol{\theta}}^*$, to P^* .

Formally, (A.15) can be rearranged as

$$\begin{aligned}\tilde{P}^* &= \min_{\phi \in \mathcal{H}} \max_{\boldsymbol{\mu}, \boldsymbol{\lambda}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y) + M\nu] \\ &\quad + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] - c_i + M\nu \right] \\ &\quad + \sum_{j=m+1}^{m+q} \int \lambda_j(\mathbf{x}, y) [\ell_j(\phi(\mathbf{x}), y) - c_j + M\nu] p_{\mathfrak{D}_j}(\mathbf{x}, y) d\mathbf{x}dy,\end{aligned}$$

where we recognize the optimization problem of

$$\begin{aligned}\tilde{P}^* &= \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y)] + M\nu \\ \text{subject to } &\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] \leq c_i - M\nu, \quad i = 1, \dots, m, \\ &\ell_j(\phi(\mathbf{x}), y) \leq c_j - M\nu \quad \mathfrak{D}_j\text{-a.e.}, \quad j = m+1, \dots, m+q.\end{aligned} \tag{PXVII}$$

Under Assumptions 1–3, (PXVII) is also strongly dual (Proposition 1), so that

$$\tilde{P}^* = \min_{\phi \in \mathcal{H}} L(\phi, \tilde{\boldsymbol{\mu}}^*, \tilde{\boldsymbol{\lambda}}^*) + \left[1 + \sum_{i=1}^m \tilde{\mu}_i^* + \sum_{j=m+1}^{m+q} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_j} [\tilde{\lambda}_j^*(\mathbf{x}, y)] \right] M\nu, \tag{A.16}$$

where $(\tilde{\boldsymbol{\mu}}^*, \tilde{\boldsymbol{\lambda}}^*)$ are the dual variables of (PXVII), i.e., the $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ that achieve

$$\begin{aligned}\tilde{D}^* &= \max_{\boldsymbol{\mu}, \boldsymbol{\lambda}} \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y) + M\nu] \\ &\quad + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] - c_i + M\nu \right] \\ &\quad + \sum_{j=m+1}^{m+q} \int \lambda_j(\mathbf{x}, y) [\ell_j(\phi(\mathbf{x}), y) - c_j + M\nu] p_{\mathfrak{D}_j}(\mathbf{x}, y) d\mathbf{x}dy.\end{aligned}$$

Going back to (A.15) we can now conclude the proof. First, use (A.16) to obtain

$$D_\theta^* \leq \tilde{P}^* \leq L(\phi^*, \tilde{\boldsymbol{\mu}}^*, \tilde{\boldsymbol{\lambda}}^*) + \left[1 + \|\tilde{\boldsymbol{\mu}}^*\|_1 + \sum_{j=m+1}^{m+q} \|\tilde{\lambda}_j^*\|_{L_1} \right] L\nu, \tag{A.17}$$

where we used ϕ^* , the solution of (P-CSO), as a suboptimal solution in (A.16) and exploited the

fact that the dual variables are non-negative to write their sum (integral) as an ℓ_1 -norm (L_1 -norm). The saddle point relation (A.8) gives $L(\phi^*, \tilde{\mu}^*, \tilde{\lambda}^*) \leq P^*$, from which we obtain the desired upper bound in (A.4). \square

A.2.5 Proof of Proposition 20: The Estimation Gap

Feasibility. The proof follows by first showing that $\hat{\theta}^*$ must be feasible for the parametrized ECRM (PIV) using the same argument as in Section (A.2.4). We then proceed as in the proof of Theorem 1.

Formally, suppose there exists at least one $i > 0$ such that

$$\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\hat{\theta}^*}(\mathbf{x}_{n_i}), y_{n_i}) > c_i \quad \text{or} \quad \ell_i(f_{\hat{\theta}^*}(\mathbf{x}_{n_i}), y_{n_i}) > c_i \text{ for some } n_i.$$

Then, since μ and λ_j are unbounded above, we obtain that $\hat{D}^* \rightarrow +\infty$. However, Assumptions 1 and 3 imply that $\hat{D}^* < +\infty$. Indeed, consider the empirical dual function

$$\hat{d}(\mu, \lambda_j) = \min_{\theta \in \mathbb{R}^p} \hat{L}(\theta, \mu, \lambda_j).$$

Using the fact that ℓ_0 is B -bounded (Assumption 1) and that there exists a strictly feasible θ^\dagger (Assumption 3), $\hat{d}(\mu, \lambda) < B$. Hence, it must be that

$$\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\hat{\theta}^*}(\mathbf{x}_{n_i}), y_{n_i}) \leq c_i, \quad \text{for } i = 1, \dots, m, \quad \text{and} \quad (\text{A.18a})$$

$$\ell_j(f_{\hat{\theta}^*}(\mathbf{x}_{n_j}), y_{n_j}) \leq c_j, \quad \text{for all } n_j \text{ and } j = m+1, \dots, m+q. \quad (\text{A.18b})$$

We now proceed to use the classic VC bound [19, Sec. 3.4] to show that $f_{\hat{\theta}^*}$ is a probably approximately feasible solution of (P-CSO). To do so, recall from (A.1) that since the ℓ_i are

bounded (Assumption 1) and \mathcal{P} has finite VC dimension $d_{\mathcal{P}}$, we obtain that

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y)] \leq \frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_i}), y_{n_i}) + V_{N_i} \quad \text{and} \quad (\text{A.19a})$$

$$\begin{aligned} \Pr_{(\mathbf{x}, y) \sim \mathfrak{D}_j} [\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \leq c_j] &= \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_j} [\mathbb{I}[\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \leq c_j]] \\ &\geq \frac{1}{N_j} \sum_{n_j=1}^{N_j} \mathbb{I}[\ell_j(f_{\boldsymbol{\theta}}(\mathbf{x}_{n_j}), y_{n_j}) \leq c_j] - V_{N_j} \end{aligned} \quad (\text{A.19b})$$

hold with probability $1 - \delta$ over the datasets $\{(\mathbf{x}_{n_i}), y_{n_i}\}_i$ for V_N as in (4.6). Combining (A.18) and (A.19) and using the union bound, we conclude that, with probability $1 - (m + q)\delta$,

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y)] &\leq c_i + V_{N_i} \quad \text{and} \\ \ell_j(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y) &\leq c_j \quad \text{for all } (\mathbf{x}, y) \in \mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}, \end{aligned}$$

where \mathcal{K}_j is a set of \mathfrak{D}_j -measure at least $1 - V_{N_j}$.

Near-optimality. Let $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ and $(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\mu}}^*, \hat{\boldsymbol{\lambda}}^*)$ be variables that achieve $D_{\boldsymbol{\theta}}^*$ in (D $_{\boldsymbol{\theta}}$ -CSL) and \hat{D}^* in ($\hat{\text{D}}$ -CSL) respectively. Then, it holds that

$$\mu_j^* \left(\mathbb{E}[\ell_i(f(\boldsymbol{\theta}^*, \mathbf{x}), y)] - c_j \right) = 0, \quad (\text{A.20a})$$

$$\lambda_j^*(\mathbf{x}, y) \left(\ell_j(f(\boldsymbol{\theta}^*, \mathbf{x}), y) - c_j \right) = 0, \quad \mathfrak{D}_j\text{-a.e.}, \quad (\text{A.20b})$$

$$\hat{\mu}_i \left(\frac{1}{N} \sum_{n=1}^N \ell_i(f(\hat{\boldsymbol{\theta}}^*, \mathbf{x}_n), y_n) - c_i \right) = 0, \quad \text{and} \quad (\text{A.20c})$$

$$\hat{\lambda}_{j, n_j} \left(\ell_i(f(\hat{\boldsymbol{\theta}}^*, \mathbf{x}_n), y_n) - c_j \right) = 0, \quad (\text{A.20d})$$

known as *complementary slackness* conditions. While these are part of the classical KKT conditions [130, Sec. 5.5.3], it should be noted that the non-convex nature of both (D $_{\boldsymbol{\theta}}$ -CSL) and ($\hat{\text{D}}$ -CSL) implies that these are only necessary and not sufficient for optimality. Nevertheless, feasibility is enough to establish (A.20).

Indeed, recall from Proposition 19 and (A.18) that the constraint slacks in parentheses in (A.20) are non-positive. Hence, the left-hand sides in (A.20) are also non-positive and if (A.20a) does not

hold for some i or if (A.20b) does not hold for some j and a set \mathcal{Z}_j of positive \mathfrak{D}_j measure, then letting $\mu_i^* = 0$ or making $\lambda_j(\mathbf{x}, y)$ vanish over \mathcal{Z}_j would increase the value of D_θ^* , contradicting its optimality. Note that since \mathcal{Z}_j is measurable, the modified λ_j would still be measurable. A similar argument applies to (A.20c) and (A.20d).

Immediately, (A.20) implies that both (D- θ -CSL) and ($\widehat{\text{D}}$ -CSL) reduce to

$$D_\theta^* = \mathbb{E}[\ell_0(f(\boldsymbol{\theta}^*, \mathbf{x}), y)] \triangleq F_0(\boldsymbol{\theta}^*) \quad \text{and} \quad (\text{A.21a})$$

$$\hat{D}^* = \frac{1}{N_0} \sum_{n_0=1}^{N_0} \ell_0(f(\hat{\boldsymbol{\theta}}^*, \mathbf{x}_{n_0}), y_{n_0}) \triangleq \hat{F}_0(\hat{\boldsymbol{\theta}}^*). \quad (\text{A.21b})$$

To proceed, use the optimality of $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}^*$ for F_0 and \hat{F}_0 respectively to write

$$F_0(\boldsymbol{\theta}^*) - \hat{F}_0(\boldsymbol{\theta}^*) \leq F_0(\boldsymbol{\theta}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*) \leq F_0(\hat{\boldsymbol{\theta}}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*).$$

Then, (A.21) yields the bound

$$\left| D_\theta^* - \hat{D}^* \right| = \left| F_0(\boldsymbol{\theta}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*) \right| \leq \max \left\{ \left| F_0(\boldsymbol{\theta}^*) - \hat{F}_0(\boldsymbol{\theta}^*) \right|, \left| F_0(\hat{\boldsymbol{\theta}}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*) \right| \right\} \quad (\text{A.22})$$

and applying the VC generalization bound from [19, Sec. 3.4] to (A.22), yields that, uniformly over $\boldsymbol{\theta}$,

$$\left| F_0(\boldsymbol{\theta}) - \hat{F}_0(\boldsymbol{\theta}) \right| \leq V_{N_0}, \quad (\text{A.23})$$

with probability $1 - 2\delta$ and for V_N as in (4.6). Combining (A.22) and (A.23) concludes the proof. \square

A.3 Proof of Proposition 1

Start by recalling that the dual problem (D-CSL) is a relaxation of its primal (P-CSO) and therefore provides a lower bound on its optimal value. Explicitly, $D^* \leq P^*$ [18, Chap. 5]. Hence, it suffices to prove that $D^* \geq P^*$. We do so by showing that even though (P-CSO) is a non-convex program, the range of its cost and constraints forms a convex set under the hypotheses of the proposition.

Explicitly, let the cost-constraints epigraph set be defined as

$$\mathcal{C} = \left\{ (s_0, \mathbf{s}) \in \mathbb{R}^{m+1} \mid \exists \phi \in \mathcal{H} \text{ such that } \mathbb{E}[\ell_0(\phi(\mathbf{x}), y)] \leq s_0 \text{ and } \mathbb{E}[\ell_i(\phi(\mathbf{x}), y)] \leq s_i \right\}, \quad (\text{A.24})$$

where the vector $\mathbf{s} \in \mathbb{R}^m$ collects the s_i , $i = 1, \dots, m$. For conciseness, we omit the distributions over which the expectations are taken whenever they can be inferred from the context. Then, the following holds:

Lemma 7. *If the conditional random variables $\mathbf{x}|y$ induced by the distributions \mathfrak{D}_i are non-atomic and \mathcal{Y} is finite, then the cost-constraints set \mathcal{C} in (A.24) is a non-empty convex set.*

Before proving Lemma 7, let us show how it implies strong duality for (P-CSO) by leveraging the following result from convex geometry:

Proposition 21 (Supporting hyperplane theorem [18, Prop. 1.5.1]). *Let $\mathcal{A} \subset \mathbb{R}^n$ be a nonempty convex set. If $\tilde{\mathbf{x}} \in \mathbb{R}^n$ is not in the interior of \mathcal{A} , then there exists a hyperplane passing through $\tilde{\mathbf{x}}$ such that \mathcal{A} is in one of its closed halfspaces, i.e., there exists $\mathbf{p} \neq \mathbf{0}$ such that $\mathbf{p}^T \tilde{\mathbf{x}} \leq \mathbf{p}^T \mathbf{x}$ for all $\mathbf{x} \in \mathcal{A}$.*

To proceed, observe from (P-CSO) that (P^*, \mathbf{c}) , where $\mathbf{c} \in \mathbb{R}^m$ collects the values of the c_i , cannot be in the interior of \mathcal{C} , otherwise there would exist $\epsilon > 0$ such that $(P^* - \epsilon, \mathbf{c}) \in \mathcal{C}$, violating the optimality of P^* . Proposition 21 then implies that there exists a non-zero vector $(\mu_0, \boldsymbol{\mu}) \in \mathbb{R}^{m+1}$ such that

$$\mu_0 s_0 + \boldsymbol{\mu}^T \mathbf{s} \geq \mu_0 P^* + \boldsymbol{\mu}^T \mathbf{c}, \quad \text{for all } (s_0, \mathbf{s}) \in \mathcal{C}. \quad (\text{A.25})$$

Observe that the hyperplanes in (A.25) are defined using the same notation as the dual problem (D-CSL) to foreshadow the fact that they actually span the values of the Lagrangian (4.11).

To proceed, note from (A.24) that \mathcal{C} is unbounded above, i.e., if $(s_0, \mathbf{s}) \in \mathcal{C}$ then $(s'_0, \mathbf{s}') \in \mathcal{C}$ for all $(s'_0, \mathbf{s}') \succeq (s_0, \mathbf{s})$. Hence, (A.25) can only hold if $\mu_i \geq 0$, $i = 0, \dots, m$. Otherwise, there exists a vector in \mathcal{C} such that the left-hand side of (A.25) evaluates to an arbitrarily negative number, eventually violating Proposition 21. Let us now show that furthermore $\mu_0 \neq 0$.

Indeed, suppose $\mu_0 = 0$. Then (A.25) reduces to

$$\boldsymbol{\mu}^T \mathbf{s} \geq \boldsymbol{\mu}^T \mathbf{c} \Leftrightarrow \boldsymbol{\mu}^T (\mathbf{s} - \mathbf{c}) \geq 0, \quad \text{for all } (s_0, \mathbf{s}) \in \mathcal{C}. \quad (\text{A.26})$$

Recall from Proposition 21 that there exists at least one $\boldsymbol{\lambda} \neq \mathbf{0}$ for which this inequality must hold. However, this is contradicted by the existence of the strictly feasible point ϕ' . Explicitly, for every $\boldsymbol{\mu} \neq \mathbf{0}$, there exists $(s'_0, \mathbf{s}') \in \mathcal{C}$, achieved by ϕ' from the hypotheses of the proposition, such that $s'_i < c_i$ for all i , contradicting (A.26).

However, if $\mu_0 \neq 0$, (A.25) can be written as

$$s_0 + \tilde{\boldsymbol{\mu}}^T \mathbf{s} \geq P^* + \tilde{\boldsymbol{\mu}}^T \mathbf{c}, \quad \text{for all } (s_0, \mathbf{s}) \in \mathcal{C},$$

where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\mu_0$, which from the definition of \mathcal{C} in (A.24) implies that

$$\mathbb{E}[\ell_0(\phi(\mathbf{x}), y)] + \sum_{i=1}^m \tilde{\mu}_i (\mathbb{E}[\ell_i(\phi(\mathbf{x}), y)] - c_i) \geq P^*, \quad (\text{A.27})$$

for all $\phi \in \mathcal{H}$. Note, however, that the left-hand side of (A.27) is the Lagrangian (4.11), i.e., (A.27) implies that $L(\phi, \tilde{\boldsymbol{\mu}}) \geq P^*$. In particular, this holds for the minimum of $L(\phi, \tilde{\boldsymbol{\mu}})$, implying that $D^* \geq P^*$ and therefore, that strong duality holds for (P-CSO). ■

All that remains now is proving that the cost-constraint set \mathcal{C} in (A.24) is convex.

Proof of Lemma 7. This proof follows along the lines of [38]. Let $(s_0, \mathbf{s}), (s'_0, \mathbf{s}') \in \mathcal{C}$ be arbitrary points achieved for $\phi, \phi' \in \mathcal{H}$, i.e., for $i = 0, \dots, m$,

$$\mathbb{E}[\ell_i(\phi(\mathbf{x}), y)] \leq s_i \quad \text{and} \quad \mathbb{E}[\ell_i(\phi'(\mathbf{x}), y)] \leq s'_i. \quad (\text{A.28})$$

It suffices then to show that $\lambda(s_0, \mathbf{s}) + (1 - \lambda)(s'_0, \mathbf{s}') \in \mathcal{C}$ for all $\lambda \in [0, 1]$ to obtain that \mathcal{C} is convex. Equivalently, we must obtain $\phi_\lambda \in \mathcal{H}$ such that

$$\mathbb{E}[\ell_i(\phi_\lambda(\mathbf{x}), y)] \leq \lambda s_i + (1 - \lambda)s'_i, \quad i = 0, \dots, m, \quad (\text{A.29})$$

for all $0 \leq \lambda \leq 1$. To do so, we rely on the following classical theorem about the range of non-atomic vector measures:

Theorem 20 (Lyapunov's convexity theorem [148, Chap. IX, Cor. 5]). *Let $\mathbf{v} : \mathcal{B} \rightarrow \mathbb{R}^n$ be a finite dimensional vector measure over the measurable space (Ω, \mathcal{B}) . If \mathbf{v} is non-atomic, then its range is convex, i.e., the set $\{\mathbf{v}(\mathcal{Z}) : \mathcal{Z} \in \mathcal{B}\}$ is a convex set.*

To see how Theorem 20 allows us to construct the desired ϕ_λ , let $\Omega = \mathbb{R}^d$ and \mathcal{B} be its Borel σ -algebra. Define the $2|\mathcal{Y}|(m+1) \times 1$ vector measure \mathbf{q} such that for every set $\mathcal{Z} \in \mathcal{B}$ we have

$$\mathbf{q}(\mathcal{Z}) = \begin{bmatrix} \int_{\mathcal{Z}} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \\ \int_{\mathcal{Z}} \ell_i(\phi'(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \end{bmatrix}_{i=0, \dots, m; y \in \mathcal{Y}}, \quad (\text{A.30})$$

where $f_i(\mathbf{x} | y)$ denotes the conditional density of \mathbf{x} given y induced by the joint distributions \mathfrak{D}_i . Once again, we assume that these density exist only to simplify the notation. The integrals in (A.30) can be taken against the conditional (Radon-Nikodym) measures as long as the law of \mathbf{x} is absolutely continuous with respect to the law of y [125]. Hence, the entries of \mathbf{q} contain integrals of the losses ℓ_i of ϕ and ϕ' with respect to every possible value of $y \in \mathcal{Y}$. Immediately, we note that $\mathbf{q}(\emptyset) = \mathbf{0}$ and

$$\mathbf{q}(\Omega) = \begin{bmatrix} \mathbb{E}[\ell_i(\phi(\mathbf{x}), y) | y] \\ \mathbb{E}[\ell_i(\phi'(\mathbf{x}), y) | y] \end{bmatrix}_{i=0, \dots, m; y \in \mathcal{Y}}. \quad (\text{A.31})$$

Due to the additive property of the Lebesgue integral, \mathbf{q} in (A.30) is a proper vector measure. What is more, the ℓ_i are bounded functions, so the fact that $\mathbf{x}|y$ is non-atomic implies that \mathbf{q} is also non-atomic. Hence, from Theorem 20, there exists a set $\mathcal{T}_\lambda \in \mathcal{B}$ such that

$$\mathbf{q}(\mathcal{T}_\lambda) = \lambda \mathbf{q}(\Omega) + (1 - \lambda) \mathbf{q}(\emptyset) = \lambda \mathbf{q}(\Omega), \quad (\text{A.32})$$

for $\lambda \in [0, 1]$. Since \mathcal{B} is a σ -algebra, it holds that $\Omega \setminus \mathcal{T}_\lambda \in \mathcal{B}$ and by additivity we obtain

$$\mathbf{q}(\Omega \setminus \mathcal{T}_\lambda) = (1 - \lambda) \mathbf{q}(\Omega). \quad (\text{A.33})$$

From (A.32) and (A.33), we then construct ϕ_λ as

$$\phi_\lambda(\mathbf{x}) = \begin{cases} \phi(\mathbf{x}), & \text{for } \mathbf{x} \in \mathcal{T}_\lambda \\ \phi'(\mathbf{x}), & \text{for } \mathbf{x} \in \Omega \setminus \mathcal{T}_\lambda \end{cases} \quad (\text{A.34})$$

It is straightforward from the decomposability hypothesis on \mathcal{H} , that $\phi_\lambda \in \mathcal{H}$. We claim that it also satisfies (A.29).

To see this is the case, use the construction in (A.34) to obtain

$$\begin{aligned} \mathbb{E} [\ell_i(\phi_\lambda(\mathbf{x}), y) \mid y] &= \int_{\mathcal{T}_\lambda} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} \mid y) d\mathbf{x} \\ &\quad + \int_{\Omega \setminus \mathcal{T}_\lambda} \ell_i(\phi'(\mathbf{x}), y) f_i(\mathbf{x} \mid y) d\mathbf{x}. \end{aligned}$$

Note from (A.30), that these integral can be written as entries of the vector measure \mathbf{q} , namely

$$\mathbb{E} [\ell_i(\phi_\lambda(\mathbf{x}), y) \mid y] = [\mathbf{p}(\mathcal{T}_\lambda)]_{\phi; i; y} + [\mathbf{p}(\Omega \setminus \mathcal{T}_\lambda)]_{\phi'; i; y}. \quad (\text{A.35})$$

In (A.35), we use the notation $[\mathbf{q}]_{\phi; i; y}$ to denote the entry of \mathbf{q} relative to the function ϕ , the i -th loss, and the variable $y \in \mathcal{Y}$. From (A.32) and (A.33), we know that (A.35) evaluates to

$$\begin{aligned} \mathbb{E} [\ell_i(\phi_\lambda(\mathbf{x}), y) \mid y] &= [\lambda \mathbf{p}(\Omega)]_{\phi; i; y} + [(1 - \lambda) \mathbf{p}(\Omega)]_{\phi'; i; y} \\ &= \lambda \mathbb{E} [\ell_i(\phi(\mathbf{x}), y) \mid y] \\ &\quad + (1 - \lambda) \mathbb{E} [\ell_i(\phi'(\mathbf{x}), y) \mid y], \end{aligned}$$

for all $i = 0, \dots, m$ and $y \in \mathcal{Y}$. Use the tower (total expectation) property, we immediately conclude that for $i = 0, \dots, m$,

$$\begin{aligned} \mathbb{E} [\ell_i(\phi_\lambda(\mathbf{x}), y)] &= \mathbb{E}_y \left[\lambda \mathbb{E} [\ell_i(\phi(\mathbf{x}), y) \mid y] \right. \\ &\quad \left. + (1 - \lambda) \mathbb{E} [\ell_i(\phi'(\mathbf{x}), y) \mid y] \right] \\ &= \lambda \mathbb{E} [\ell_i(\phi(\mathbf{x}), y)] + (1 - \lambda) \mathbb{E} [\ell_i(\phi'(\mathbf{x}), y)], \end{aligned}$$

which using (A.28) yields

$$\mathbb{E}[\ell_i(\phi_\lambda(\mathbf{x}), y)] \leq \lambda s_i + (1 - \lambda)s'_i.$$

Hence, there exists $\phi_\lambda \in \mathcal{H}$ such that (A.29) holds for all $\lambda \in [0, 1]$ and $(s_0, \mathbf{s}), (s'_0, \mathbf{s}') \in \mathcal{C}$. The set \mathcal{C} is therefore convex. Moreover, the strictly feasible ϕ' from the hypotheses of the proposition implies that \mathcal{C} is not empty. \blacksquare

A.4 Duality Gap for Regression

The Lyapunov convexity theorem (Theorem 20) turns out to be quite sensitive to the hypothesis that the vector measure takes values in a finite dimensional Banach space [148, Ch. IX]. When \mathcal{Y} is compact, we can overcome this issue without resorting to super-atomless (saturated) spaces by assuming the losses are continuous in y and slicing \mathcal{Y} to approximate regression by a sequence of increasingly fine classification problems. Once again, we consider the measurable space (Ω, \mathcal{B}) where $\Omega = \mathbb{R}^d$ and \mathcal{B} is a Borel σ -algebra.

Proposition 22. *Consider the dual problem (D-CSL). Under assumptions 3 and 4, (P-CSO) is strongly dual, i.e., $P^* = D^*$, if \mathcal{Y} is compact, the conditional distributions $\mathbf{x}|y$ induced by the \mathfrak{D}_i are non-atomic, and $y \mapsto \ell_i(\phi(\cdot), y)f_i(\cdot | y)$ is continuous in the total variation topology for each $\phi \in \mathcal{H}$, where $f_i(\mathbf{x} | y)$ denotes the density of the conditional random variable induced by \mathfrak{D}_i . Explicitly, for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $|y - \tilde{y}| \leq \delta$ it holds that*

$$\sup_{Z \in \mathcal{B}} \int_Z |\ell_i(\phi(\mathbf{x}), y)f_i(\mathbf{x} | y) - \ell_i(\phi(\mathbf{x}), \tilde{y})f_i(\mathbf{x} | \tilde{y})| d\mathbf{x} \leq \epsilon,$$

The proof of Proposition 22 follows that of the finite \mathcal{Y} case in Appendix A.3 by replacing Lemma 7 with

Lemma 8. *Under the assumptions of Proposition 22, the cost-constraints set \mathcal{C} in (A.24) is a non-empty convex set.*

Proof of Lemma 8. Without loss of generality, assume $\mathcal{Y} = [0, 1]$. Once again, let $(s_0, \mathbf{s}), (s'_0, \mathbf{s}') \in \mathcal{C}$ be achieved by $\phi, \phi' \in \mathcal{H}$. Our goal, as before, is to construct ϕ_λ such that (A.29) holds for all $\lambda \in [0, 1]$.

To do so, fix $\epsilon > 0$ and let δ be such that

$$\sup_{\mathbf{z} \in \mathcal{B}} \int_{\mathcal{Z}} \left| \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) - \ell_i(\phi(\mathbf{x}), \tilde{y}) f_i(\mathbf{x} | \tilde{y}) \right| d\mathbf{x} \leq \frac{\epsilon}{3}$$

for all $|y - \tilde{y}| \leq \delta$. The continuity hypothesis guarantees that such a δ exists. Then, partition \mathcal{Y} into $\lceil 1/\delta \rceil$ intervals $\mathcal{I}_k = [(k-1)\delta, k\delta]$ and let

$$\tilde{\mathcal{Y}} = \left\{ \tilde{y}_k \triangleq (k-1/2)\delta \text{ for } k = 1, \dots, \lceil 1/\delta \rceil \right\}. \quad (\text{A.36})$$

To proceed, construct the $2|\tilde{\mathcal{Y}}|(m+1) \times 1$ vector measure

$$\mathbf{q}_\epsilon(\mathcal{Z}) = \begin{bmatrix} \int_{\mathcal{Z}} \ell_i(\phi(\mathbf{x}), \tilde{y}) f_i(\mathbf{x} | \tilde{y}) d\mathbf{x} \\ \int_{\mathcal{Z}} \ell_i(\phi'(\mathbf{x}), \tilde{y}) f_i(\mathbf{x} | \tilde{y}) d\mathbf{x} \end{bmatrix}_{i=0, \dots, m; \tilde{y} \in \tilde{\mathcal{Y}}} \quad (\text{A.37})$$

and, using the non-atomicity of \mathbf{q}_ϵ , obtain from Theorem 20 a set $\mathcal{T}_{\lambda, \epsilon} \in \mathcal{B}$ such that

$$\mathbf{q}(\mathcal{T}_{\lambda, \epsilon}) = \lambda \mathbf{q}_\epsilon(\Omega) \quad \text{and} \quad \mathbf{q}(\Omega \setminus \mathcal{T}_{\lambda, \epsilon}) = (1 - \lambda) \mathbf{q}_\epsilon(\Omega). \quad (\text{A.38})$$

From (A.38), we construct $\phi_{\lambda, \epsilon}$ as

$$\phi_{\lambda, \epsilon}(\mathbf{x}) = \begin{cases} \phi(\mathbf{x}), & \text{for } \mathbf{x} \in \mathcal{T}_{\lambda, \epsilon} \\ \phi'(\mathbf{x}), & \text{for } \mathbf{x} \in \Omega \setminus \mathcal{T}_{\lambda, \epsilon} \end{cases} \quad (\text{A.39})$$

Since \mathcal{H} is decomposable, we again have $\phi_{\lambda, \epsilon} \in \mathcal{H}$. Let us show that it satisfies (A.29) up to an additive error ϵ .

Indeed, notice from (A.39) that

$$\begin{aligned} \mathbb{E}[\ell_i(\phi_{\lambda, \epsilon}(\mathbf{x}), y)] &= \mathbb{E}_y \left[\mathbb{E}[\ell_i(\phi_{\lambda, \epsilon}(\mathbf{x}), y) | y] \right] \\ &= \mathbb{E}_y \left[\int_{\mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \right] \\ &\quad + \mathbb{E}_y \left[\int_{\Omega \setminus \mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi'(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \right]. \end{aligned} \quad (\text{A.40})$$

Focusing on the first expectation, we start by constructing a simple function approximation of the integrand using the intervals from (A.36). From the continuity of $\ell_i(\phi(\mathbf{x}), y)f_i(\mathbf{x} | y)$, we then obtain the bound

$$\mathbb{E}_y \left[\int_{\mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \right] \leq \sum_{k=0}^{|\tilde{\mathcal{Y}}|} \mathbb{E}_y [\mathbb{I}[y \in \mathcal{I}_k]] \int_{\mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi(\mathbf{x}), \tilde{y}_k) f_i(\mathbf{x} | \tilde{y}_k) d\mathbf{x} + \frac{\epsilon}{3}. \quad (\text{A.41})$$

Notice from the definition of the vector measure in (A.37) that the value of the integrals in (A.41) are entries of the vector $\mathbf{q}_\epsilon(\mathcal{T}_{\lambda, \epsilon})$. Using the property of $\mathcal{T}_{\lambda, \epsilon}$ in (A.38), we then get

$$\mathbb{E}_y \left[\int_{\mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \right] \leq \lambda \sum_{k=0}^{|\tilde{\mathcal{Y}}|} \mathbb{E}_y [\mathbb{I}[y \in \mathcal{I}_k]] \int \ell_i(\phi(\mathbf{x}), \tilde{y}_k) f(\mathbf{x} | \tilde{y}_k) d\mathbf{x} + \frac{\epsilon}{3}. \quad (\text{A.42})$$

Using once again the continuity hypothesis and the fact that since \mathfrak{D}_i is a probability measure, $\sum_{k=0}^{|\tilde{\mathcal{Y}}|} \mathbb{E}_y [\mathbb{I}[y \in \mathcal{I}_k]] = 1$, yields

$$\begin{aligned} \mathbb{E}_y \left[\int_{\mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \right] &\leq \lambda \mathbb{E}_y [\mathbb{E}[\ell_i(\phi(\mathbf{x}), y) | y]] + \frac{(1 + \lambda)\epsilon}{3} \\ &= \lambda \mathbb{E}[\ell_i(\phi(\mathbf{x}), y)] + \frac{(1 + \lambda)\epsilon}{3}. \end{aligned} \quad (\text{A.43})$$

A similar argument yields

$$\mathbb{E}_y \left[\int_{\Omega \setminus \mathcal{T}_{\lambda, \epsilon}} \ell_i(\phi(\mathbf{x}), y) f_i(\mathbf{x} | y) d\mathbf{x} \right] \leq (1 - \lambda) \mathbb{E}[\ell_i(\phi'(\mathbf{x}), y)] + \frac{(2 - \lambda)\epsilon}{3}. \quad (\text{A.44})$$

Combining (A.43) and (A.44) in (A.40), we obtain that for all $\epsilon > 0$ and $\lambda \in [0, 1]$, there exists $\phi_{\lambda, \epsilon} \in \mathcal{H}$ such that

$$\mathbb{E}[\ell_i(\phi_{\lambda, \epsilon}(\mathbf{x}), y)] \leq \lambda s_i + (1 - \lambda) s'_i + \epsilon, \quad (\text{A.45})$$

for all $i = 0, \dots, m$.

Suppose now that there is no $\phi_\lambda \in \mathcal{H}$ such that $\mathbb{E}[\ell_i(\phi_\lambda(\mathbf{x}), y)] \leq \lambda s_i + (1 - \lambda) s'_i$. Then, there exists $\tau > 0$ such that

$$\mathbb{E}[\ell_i(\phi(\mathbf{x}), y)] > \lambda s_i + (1 - \lambda) s'_i + \tau$$

for $\phi \in \mathcal{H}$. However, this violates (A.45) for $\epsilon = \tau$ leading to a contradiction. Since \mathcal{H} is closed, we

therefore obtain that for all $(s_0, \mathbf{s}), (s'_0, \mathbf{s}') \in \mathcal{C}$ and $\lambda \in [0, 1]$, there exists $\phi_\lambda \in \mathcal{H}$ such that (A.29), showing that \mathcal{C} is convex. The strictly feasible ϕ' from assumption 3 implies that \mathcal{C} is also not empty. \blacksquare

A.5 Proof of Proposition 2: The Approximation Gap

We first prove that f_{θ^*} is feasible for (P-CSO) and then bound the gap between D_θ^* and P^* .

Feasibility. Suppose that f_{θ^*} is infeasible. Then, there exists at least one index i for which $\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\theta^*}(\mathbf{x}), y)] > c_i$. Since $\boldsymbol{\mu}$ is unbounded above, we obtain that $D_\theta^* \rightarrow +\infty$. However, assumption 3 implies that $D_\theta^* < +\infty$. Indeed, consider the dual function

$$\begin{aligned} d(\boldsymbol{\mu}) &= \min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_\theta(\boldsymbol{\theta}, \boldsymbol{\mu}) \\ &= \min_{\boldsymbol{\theta} \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(f_\theta(\mathbf{x}), y)] \\ &\quad + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_\theta(\mathbf{x}), y)] - c_i \right], \end{aligned} \tag{A.46}$$

for the parametrized Lagrangian L_θ defined in (D $_\theta$ -CSL). Using the fact that ℓ_0 is B -bounded and that there exists a strictly feasible $\boldsymbol{\theta}'$ (assumption 3), $d(\boldsymbol{\mu})$ is upper bounded by

$$\begin{aligned} d(\boldsymbol{\mu}) &\leq \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(f_{\boldsymbol{\theta}'}(\mathbf{x}), y)] \\ &\quad + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(f_{\boldsymbol{\theta}'}(\mathbf{x}), y)] - c_i \right] \leq B, \end{aligned} \tag{A.47}$$

where we used the fact that $\mu_i \geq 0$. Hence, it must be that f_{θ^*} is feasible for (P-CSO).

Near-optimality. First, recall from Proposition 1 that (P-CSO)–(D-CSL) form a strongly dual pair of mathematical programs under the assumptions of Proposition 2. The Lagrangian in (4.11) therefore obeys the saddle-point relation

$$L(\phi^*, \boldsymbol{\mu}') \leq \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} \min_{\phi \in \mathcal{H}} L(\phi, \boldsymbol{\mu}) = D^* = P^* = \min_{\phi \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} L(\phi, \boldsymbol{\mu}) \leq L(\phi', \boldsymbol{\mu}^*) \tag{A.48}$$

for all $\phi' \in \mathcal{H}$ and $\boldsymbol{\mu}' \in \mathbb{R}_+^m$, where ϕ^* is a solution of (P-CSO) and $\boldsymbol{\mu}^*$ is a solution of (D-CSL). Additionally, we have from (D $_{\theta}$ -CSL) that

$$D_{\theta}^* \geq \min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_{\theta}(\boldsymbol{\theta}, \boldsymbol{\mu}), \quad \text{for all } \boldsymbol{\mu} \in \mathbb{R}_+^m. \quad (\text{A.49})$$

Immediately, we obtain the lower bound in (4.12). Explicitly,

$$D_{\theta}^* \geq \min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_{\theta}(\boldsymbol{\theta}, \boldsymbol{\mu}^*) \geq \min_{\phi \in \mathcal{H}} L(\phi, \boldsymbol{\mu}^*) = P^*, \quad (\text{A.50})$$

where the second inequality comes from the fact that $\mathcal{P} \subseteq \mathcal{H}$ (Assumption 5).

The upper bound is obtained by relating the parameterized dual problem (D $_{\theta}$ -CSL) to a perturbed (tightened) version of the original (P-CSO). To do so, start by adding and subtracting $L(\phi, \boldsymbol{\mu})$ from (D $_{\theta}$ -CSL) to get

$$\begin{aligned} D_{\theta}^* &= \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} \min_{\boldsymbol{\theta} \in \mathbb{R}^p} L(\phi, \boldsymbol{\mu}) \\ &\quad + \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_0} \left[\ell_0(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_0(\phi(\mathbf{x}), y) \right] \\ &\quad + \sum_{i=1}^m \mu_i \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left[\ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_i(\phi(\mathbf{x}), y) \right]. \end{aligned} \quad (\text{A.51})$$

Then, use the fact that ℓ_i is M -Lipschitz continuous (assumption 4) to bound the expectations in (A.51) as

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left[\ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_i(\phi(\mathbf{x}), y) \right] \\ \leq \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left[\left| \ell_i(f_{\boldsymbol{\theta}}(\mathbf{x}), y) - \ell_i(\phi(\mathbf{x}), y) \right| \right] \\ \leq M \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left[\left| f_{\boldsymbol{\theta}}(\mathbf{x}) - \phi(\mathbf{x}) \right| \right], \end{aligned} \quad (\text{A.52})$$

for all $i = 0, \dots, m$. Using (A.52) and the approximation property of \mathcal{P} (assumption 5), the minimization in (A.51) can be upper bounded to obtain

$$D_{\theta}^* \leq \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} L(\phi, \boldsymbol{\mu}) + (1 + \|\boldsymbol{\mu}\|_1) M \nu, \quad (\text{A.53})$$

where we used the fact that $\mu_i \geq 0$ to write $\sum \mu_i = \|\boldsymbol{\mu}\|_1$.

Notice that since (A.53) holds uniformly for all $\phi \in \mathcal{H}$, it also holds for the minimizer

$$D_\theta^* \leq \min_{\phi \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} L(\phi, \boldsymbol{\mu}) + (1 + \|\boldsymbol{\mu}\|_1) M\nu \triangleq \tilde{P}^* \quad (\text{A.54})$$

and that the right-hand side of (A.54), namely \tilde{P}^* , is in fact a perturbed version of (P-CSO). Hence, we obtain another saddle-point relation similar to (A.48) relating \tilde{P}^* , and consequently D_θ^* , to P^* .

Formally, (A.54) can be rearranged as

$$\begin{aligned} \tilde{P}^* &= \min_{\phi \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y) + M\nu] \\ &\quad + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] - c_i + M\nu \right] \end{aligned} \quad (\text{A.55})$$

where we recognize the primal optimization problem

$$\begin{aligned} \tilde{P}^* &= \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y)] + M\nu \\ \text{subject to} \quad &\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] \leq c_i - M\nu, \\ &i = 1, \dots, m. \end{aligned} \quad (\text{PXVIII})$$

Notice from Proposition 1 and assumption 3 that (PXVIII) is also strongly dual, so that

$$\tilde{P}^* = \min_{\phi \in \mathcal{H}} L(\phi, \tilde{\boldsymbol{\mu}}^*) + (1 + \|\tilde{\boldsymbol{\mu}}^*\|_1) M\nu, \quad (\text{A.56})$$

where $\tilde{\boldsymbol{\mu}}^*$ are the dual variables of (PXVIII), i.e., the $\boldsymbol{\mu}$ that achieve

$$\begin{aligned} \tilde{D}^* &= \max_{\boldsymbol{\mu} \in \mathbb{R}_+^m} \min_{\phi \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_0} [\ell_0(\phi(\mathbf{x}), y) + M\nu] \\ &\quad + \sum_{i=1}^m \mu_i \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} [\ell_i(\phi(\mathbf{x}), y)] - c_i + M\nu \right]. \end{aligned} \quad (\text{A.57})$$

Going back to (A.54) we can now conclude the proof. First, use (A.56) to obtain

$$D_\theta^* \leq \tilde{P}^* \leq L(\phi^*, \tilde{\boldsymbol{\mu}}^*) + (1 + \|\tilde{\boldsymbol{\mu}}^*\|_1) L\nu, \quad (\text{A.58})$$

where we used ϕ^* , the solution of (P-CSO), as a suboptimal solution in (A.56). The saddle point relation (A.48) gives $L(\phi^*, \tilde{\mu}^*) \leq P^*$, from which we obtain the upper bound in (4.12). ■

A.6 Proof of Proposition 3: The Estimation Gap

Feasibility. The proof follows by first showing that $\hat{\theta}^*$ must be feasible for the parametrized ECRM (PIV) using the same argument as in Section A.5. Then, leveraging the fact that \mathcal{P} is PAC learnable (Assumption 5), we can apply the generalization bounds from classical learning theory.

Formally, suppose there exists at least one $i > 0$ such that

$$\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\hat{\theta}^*}(\mathbf{x}_{n_i}), y_{n_i}) > c_i,$$

for the samples $(\mathbf{x}_{n_i}, y_{n_i}) \sim \mathfrak{D}_i$. Then, since μ is unbounded above, we obtain that $\hat{D}^* \rightarrow +\infty$. However, assumption 3 implies that $\hat{D}^* < +\infty$. Indeed, consider the empirical dual function

$$\hat{d}(\mu) = \min_{\theta \in \mathbb{R}^p} \hat{L}(\theta, \mu) \tag{A.59}$$

for the empirical Lagrangian \hat{L} from (4.4). Using the fact that ℓ_0 is B -bounded and that there exists a strictly feasible θ^\dagger (assumption 3), we can use the same argument leading to (A.47) to obtain that $\hat{d}(\mu) < B$. Hence, it must be that

$$\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\hat{\theta}^*}(\mathbf{x}_{n_i}), y_{n_i}) \leq c_i, \quad \text{for } i = 1, \dots, m. \tag{A.60}$$

We now proceed to use the classical VC bound [19, Sec. 3.4] to show that $f_{\hat{\theta}^*}$ must approximately satisfy each constraint of (P-CSO) with high probability. To do so, recall that since \mathcal{P} is PAC learnable, it must have a finite VC dimension $d_{\mathcal{P}}$ [20, Thm. 6.7]. Since the ℓ_i are bounded, it holds with probability $1 - \delta$ over the samples $(\mathbf{x}_{n_i}, y_{n_i})$ that [19, Sec. 3.4]

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} \left[\ell_i(f_{\theta}(\mathbf{x}), y) \right] \leq \frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\theta}(\mathbf{x}_{n_i}), y_{n_i}) + V_{N_i} \tag{A.61}$$

for each $i = 1, \dots, m$, where V_N is as in (4.6).

Near-optimality. Let $(\tilde{\boldsymbol{\theta}}^*, \tilde{\boldsymbol{\mu}}^*)$ and $(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\mu}}^*)$ be variables that achieve D_θ^* in (D $_\theta$ -CSL) and \hat{D}^* in (\hat{D} -CSL) respectively. Then, it holds that

$$\tilde{\mu}_i^* \left(\mathbb{E} \left[\ell_i(f(\tilde{\boldsymbol{\theta}}^*, \mathbf{x}), y) \right] - c_j \right) = 0 \quad \text{and} \quad (\text{A.62a})$$

$$\hat{\mu}_i^* \left(\frac{1}{N} \sum_{n=1}^N \ell_i(f(\hat{\boldsymbol{\theta}}^*, \mathbf{x}_n), y_n) - c_i \right) = 0, \quad (\text{A.62b})$$

known as *complementary slackness* conditions. While these are part of the classical KKT conditions [130, Sec. 5.5.3], it should be noted that the non-convex nature of both (D $_\theta$ -CSL) and (\hat{D} -CSL) implies that these are only necessary and not sufficient for optimality. Nevertheless, feasibility is enough to establish (A.62).

Indeed, recall from Proposition 2 and (A.60) that the constraint slacks in parentheses in (A.62) are non-positive. Hence, the left-hand sides in (A.62) are also non-positive and if (A.62a) does not hold for some i , then letting $\tilde{\mu}_i^* = 0$ would increase the value of D_θ^* , contradicting its optimality. A similar argument applies to (A.62b).

Immediately, (A.62) implies that both (D $_\theta$ -CSL) and (\hat{D} -CSL) reduce to

$$D_\theta^* = \mathbb{E} \left[\ell_0 \left(f(\tilde{\boldsymbol{\theta}}^*, \mathbf{x}), y \right) \right] \triangleq F_0(\tilde{\boldsymbol{\theta}}^*) \quad \text{and} \quad (\text{A.63a})$$

$$\hat{D}^* = \frac{1}{N_0} \sum_{n_0=1}^{N_0} \ell_0 \left(f(\hat{\boldsymbol{\theta}}^*, \mathbf{x}_{n_0}), y_{n_0} \right) \triangleq \hat{F}_0(\hat{\boldsymbol{\theta}}^*). \quad (\text{A.63b})$$

To proceed, use the optimality of $\tilde{\boldsymbol{\theta}}^*$ and $\hat{\boldsymbol{\theta}}^*$ for F_0 and \hat{F}_0 respectively to write

$$F_0(\tilde{\boldsymbol{\theta}}^*) - \hat{F}_0(\tilde{\boldsymbol{\theta}}^*) \leq F_0(\tilde{\boldsymbol{\theta}}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*) \leq F_0(\hat{\boldsymbol{\theta}}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*).$$

Then, (A.63) yields the bound

$$\left| D_\theta^* - \hat{D}^* \right| = \left| F_0(\tilde{\boldsymbol{\theta}}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*) \right| \leq \max \left\{ \left| F_0(\tilde{\boldsymbol{\theta}}^*) - \hat{F}_0(\tilde{\boldsymbol{\theta}}^*) \right|, \left| F_0(\hat{\boldsymbol{\theta}}^*) - \hat{F}_0(\hat{\boldsymbol{\theta}}^*) \right| \right\} \quad (\text{A.64})$$

and applying the VC generalization bound from [19, Sec. 3.4] to (A.64), yields that, uniformly

over $\boldsymbol{\theta}$,

$$\left| F_0(\boldsymbol{\theta}) - \hat{F}_0(\boldsymbol{\theta}) \right| \leq V_{N_0}, \quad (\text{A.65})$$

with probability $1 - 2\delta$ and for V_N as in (4.6).

Union bound. To conclude, we use the union bound to combine (A.61) for $i = 1, \dots, m$ with (A.64) and (A.65). Doing so, we obtain that

$$\begin{aligned} \left| D_\theta^* - \hat{D}^* \right| &\leq V_{N_0} \\ \mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{D}_i} \left[\ell_i(f_{\hat{\boldsymbol{\theta}}^*}(\mathbf{x}), y) \right] &\leq c_i + V_{N_i} \end{aligned}$$

occur simultaneously with probability at least $1 - (m + 2)\delta$. Choosing the appropriate confidence level δ concludes the proof. \blacksquare

A.7 Proof of Theorem 4

We proceed by proving that

$$\hat{D}^* - \rho - \frac{\eta}{2} S - \beta \leq \hat{L}(\boldsymbol{\theta}^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\lambda}^{(T)}) \leq \hat{D}^* + \rho, \quad (\text{A.66})$$

for a fixed precision $\beta > 0$, from which we obtain (4.15) by recalling that \hat{D}^* is near-PACC (Theorems 2 and 3). The result stated in Theorem 4 is based on taking $\beta = M\nu$.

Start by defining the empirical dual function of (P-CSO) for $q \geq 0$

$$\hat{d}(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) \triangleq \min_{\boldsymbol{\theta}} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j). \quad (\text{A.67})$$

The upper bound in (A.66) then holds trivially from the fact that $\hat{d}(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) \leq \hat{D}^*$ for all $(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$. Then, from the characteristics of the approximate minimizer $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\dagger(\boldsymbol{\mu}^{(t)}, \boldsymbol{\lambda}^{(t)})$ in Assumption 6 we obtain that

$$\hat{L}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\lambda}^{(t)}) \leq \hat{D}^* + \rho, \quad \text{for all } t \geq 0. \quad (\text{A.68})$$

For the lower bound, we rely on the following relaxation of Dankin's classical theorem [77, Ch. 3]:

Lemma 9. Let $\boldsymbol{\theta}^\dagger$ be the approximate minimizer of the empirical Lagrangian (4.4) at $(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ from Assumption 6. Then, the constraint slacks are approximate subgradients of the dual function (A.67), i.e.,

$$\begin{aligned} \hat{d}(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) &\geq \hat{d}(\boldsymbol{\mu}', \boldsymbol{\lambda}_j') + \sum_{i=1}^m (\mu_i - \mu'_i) \left[\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}^\dagger}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right] \\ &\quad + \sum_{j=m+1}^{m+q} \left[\frac{1}{N_j} \sum_{n_j=1}^{N_j} (\lambda_{j,n_j} - \lambda'_{j,n_j}) (\ell_j(f_{\boldsymbol{\theta}^\dagger}(\mathbf{x}_{n_j}), y_{n_j}) - c_j) \right] - \rho \end{aligned} \quad (\text{A.69})$$

for all $(\boldsymbol{\mu}', \boldsymbol{\lambda}_j')$.

Proof. From Assumption 6, we obtain that

$$d(\boldsymbol{\mu}', \boldsymbol{\lambda}_j') \leq d(\boldsymbol{\mu}', \boldsymbol{\lambda}_j') + d(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) - \hat{L}(\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j), \boldsymbol{\mu}, \boldsymbol{\lambda}_j) + \rho. \quad (\text{A.70})$$

Additionally, we can upper bound (A.70) by replacing the optimal minimizer in $d(\boldsymbol{\mu}', \boldsymbol{\lambda}_j')$ by any $\boldsymbol{\theta}$.

In particular, we can choose $\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ to get

$$d(\boldsymbol{\mu}', \boldsymbol{\lambda}_j') \leq \hat{L}(\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j), \boldsymbol{\mu}', \boldsymbol{\lambda}_j') + d(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) - \hat{L}(\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j), \boldsymbol{\mu}, \boldsymbol{\lambda}_j) + \rho. \quad (\text{A.71})$$

Notice from (4.4) that the first term of the Lagrangians in (A.71) are identical. By expanding them, (A.71) can then be rearranged as in (A.69). ■

To proceed, let $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}_j^*)$ be solutions of the dual problem ($\widehat{\mathbf{D}}$ -CSL). We show next that for at least $T = O(1/\beta)$, the total distance

$$U_t = \left\| \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^* \right\|^2 + \sum_{j=m+1}^{m+q} \left\| \boldsymbol{\lambda}_j^{(t)} - \boldsymbol{\lambda}_j^* \right\|^2 \quad (\text{A.72})$$

decreases to at least $O(\beta)$. To do so, use the updates from Algorithm 1 to write (A.72) as

$$\begin{aligned} U_t = \sum_{i=1}^m &\left\{ \left[\mu_i^{(t-1)} + \eta \left(\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right) \right]_+ - \mu_i^* \right\}^2 \\ &+ \sum_{j=m+1}^{m+q} \sum_{n_j=1}^{N_j} \left\{ \left[\lambda_{j,n_j}^{(t-1)} + \frac{\eta}{N_j} (\ell_j(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_j}), y_{n_j}) - c_j) \right]_+ - \lambda_{j,n_j}^* \right\}^2. \end{aligned}$$

Since both $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^*$ belong to the non-negative orthant, we can then use the non-expansiveness of the projection $[\cdot]_+$ [18] to obtain

$$U_t = \sum_{i=1}^m \left[\mu_i^{(t-1)} + \eta \left(\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right) - \mu_i^* \right]^2 + \sum_{j=m+1}^{m+q} \sum_{n_j=1}^{N_j} \left[\lambda_{j,n_j}^{(t-1)} + \frac{\eta}{N_j} \left(\ell_j(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_j}), y_{n_j}) - c_j \right) - \lambda_{j,n_j}^* \right]^2. \quad (\text{A.73})$$

By expanding the norms in (A.73), we get that

$$U_t \leq U_{t-1} + 2\eta \left[\sum_i \left(\mu_i^{(t-1)} - \mu_i^* \right) \left(\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right) + \sum_j \sum_{n_j=1}^{N_j} \frac{1}{N_j} \left(\lambda_{j,n_j}^{(t-1)} - \lambda_{j,n_j}^* \right) \left(\ell_j(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_j}), y_{n_j}) - c_j \right) \right] + \eta^2 \left[\sum_{i=1}^m \left[\frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_i}), y_{n_i}) - c_i \right]^2 + \sum_{j=m+1}^{m+q} \sum_{n_j=1}^{N_j} \frac{1}{N_j^2} \left[\ell_j(f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x}_{n_j}), y_{n_j}) - c_j \right]^2 \right]. \quad (\text{A.74})$$

Using the fact that the ℓ_i are bounded, the last term in (A.74) is upper bounded by

$$S = \sum_{i=1}^m (B - c_i)^2 + \sum_{j=m+1}^{m+q} \frac{1}{N_j} (B - c_j)^2 = O(B^2).$$

What is more, Lemma 9 can be used to bound the second term in (A.74) and write

$$U_t \leq U_{t-1} + 2\eta \left[\hat{d}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}) - \hat{D}^* + \rho \right] + \eta^2 S,$$

where we used the fact that $\hat{D}^* = \hat{d}(\boldsymbol{\mu}^*, \boldsymbol{\lambda}_j^*)$. Solving the recursion then yields

$$U_t \leq U_0 + 2\eta \sum_{t=0}^{t-1} \Delta_t, \quad (\text{A.75})$$

for

$$\Delta_t = \hat{d}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}) - \hat{D}^* + \rho + \frac{\eta}{2} S. \quad (\text{A.76})$$

To conclude, notice that $\hat{d}(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) \leq \hat{D}^*$ for all $(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$. Hence, when $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\lambda}_j^{(t)}$ are sufficiently far from the optimum and the step size η is sufficiently small, we have $\Delta_t \leq 0$ and (A.75) shows that the distance to the optimum U_t decreases. Formally, fix a precision $\beta > 0$ and let $T = \min\{t \mid \Delta_t > -\beta\}$. Then, from the definition of Δ_t we obtain the desired lower bound

$$\Delta_T > -\beta \Leftrightarrow \hat{d}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}) > \hat{D}^* - \rho - \frac{\eta}{2}S - \beta$$

What is more, (A.75) yields

$$T \leq \frac{U_0}{2\eta\beta} + 1 = O(\beta^{-1}).$$

■

Appendix B

Proofs of Part II

B.1 Chapter 6: Nonconvex functional models

B.1.1 Proof of Lemma 2

Let $(c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I), (c', \mathbf{u}', \mathbf{k}'_R, \mathbf{k}'_I)$ be arbitrary points in \mathcal{C} achieved for $(X, \mathbf{z}), (X', \mathbf{z}') \in \mathcal{X} \times \mathbb{C}^p$. In other words, it holds that $X(\beta), X'(\beta) \in \mathcal{P}$ a.e., $f_0(X) \leq c, f_0(X') \leq c', g_i(\mathbf{z}) \leq [\mathbf{u}]_i, g_i(\mathbf{z}') \leq [\mathbf{u}']_i$,

$$\begin{aligned} \int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta - \mathbf{z} &= \mathbf{k}_R + j\mathbf{k}_I \triangleq \mathbf{k}, \\ \int_{\Omega} \mathbf{F}[X'(\beta), \beta] d\beta - \mathbf{z}' &= \mathbf{k}'_R + j\mathbf{k}'_I \triangleq \mathbf{k}', \end{aligned}$$

where we have defined the shorthands \mathbf{k} and \mathbf{k}' for conciseness. To show that \mathcal{C} is convex, it suffices to show that $\theta(c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I) + (1 - \theta)(c', \mathbf{u}', \mathbf{k}'_R, \mathbf{k}'_I) \in \mathcal{C}$ for any $\theta \in [0, 1]$. Equivalently, we must obtain $(X_{\theta}, \mathbf{z}_{\theta}) \in \mathcal{X} \times \mathbb{C}^p$ such that $X_{\theta}(\beta) \in \mathcal{P}$ a.e.,

$$f_0(X_{\theta}) \leq \theta c + (1 - \theta)c', \tag{B.1a}$$

$$g_i(\mathbf{z}_{\theta}) \leq [\theta \mathbf{u} + (1 - \theta)\mathbf{u}']_i, \tag{B.1b}$$

$$\int_{\Omega} \mathbf{F}[X_{\theta}(\beta), \beta] d\beta - \mathbf{z}_{\theta} = \theta \mathbf{k} + (1 - \theta)\mathbf{k}', \tag{B.1c}$$

for any $0 \leq \theta \leq 1$. To do so, we will rely on the following classical theorem about the range of non-atomic vector measures:

Theorem 21 (Lyapunov's convexity theorem [148]). *Let $\mathbf{v} : \mathcal{B} \rightarrow \mathbb{C}^n$ be a vector measure over the measurable space (Ω, \mathcal{B}) . If \mathbf{v} is non-atomic, then its range is convex, i.e., the set $\{\mathbf{v}(\mathcal{A}) : \mathcal{A} \in \mathcal{B}\}$ is a convex set.*

To see how Theorem 21 allows us to construct the desired X_θ , start by defining a $2(p+1) \times 1$ vector measure \mathbf{q} over (Ω, \mathcal{B}) such that for every set $\mathcal{Z} \in \mathcal{B}$ we have

$$\mathbf{q}(\mathcal{Z}) = \begin{bmatrix} \int_{\mathcal{Z}} \mathbf{F}[X(\beta), \beta] d\beta \\ \int_{\mathcal{Z}} \mathbf{F}[X'(\beta), \beta] d\beta \\ \int_{\mathcal{Z}} [F_0(X(\beta), \beta) + \lambda \mathbb{I}(X(\beta) \neq 0)] d\beta \\ \int_{\mathcal{Z}} [F_0(X'(\beta), \beta) + \lambda \mathbb{I}(X'(\beta) \neq 0)] d\beta \end{bmatrix}. \quad (\text{B.2})$$

Notice that \mathbf{q} is a proper vector measure, so that $\mathbf{q}(\emptyset) = \mathbf{0}$. Also, observe that evaluating \mathbf{q} on the whole space Ω yields

$$\mathbf{q}(\Omega) = \begin{bmatrix} \int_{\Omega} \mathbf{F}[X(\beta), \beta] d\beta \\ \int_{\Omega} \mathbf{F}[X'(\beta), \beta] d\beta \\ f_0(X) \\ f_0(X') \end{bmatrix} \Rightarrow \mathbf{q}(\Omega) = \begin{bmatrix} \mathbf{k} + \mathbf{z} \\ \mathbf{k}' + \mathbf{z}' \\ f_0(X) \\ f_0(X') \end{bmatrix}. \quad (\text{B.3})$$

Finally, observe that since F_0 and \mathbf{F} do not contain Dirac deltas, they induce non-atomic measures. Consequently, \mathbf{q} is non-atomic.

To proceed, use Theorem 21 to find a set $\mathcal{T}_\theta \in \mathcal{B}$ such that

$$\mathbf{q}(\mathcal{T}_\theta) = \theta \mathbf{q}(\Omega) + (1 - \theta) \mathbf{q}(\emptyset) = \theta \mathbf{q}(\Omega) \quad (\text{B.4})$$

for $\theta \in [0, 1]$. Since \mathcal{B} is a σ -algebra, it holds that $\Omega \setminus \mathcal{T}_\theta \in \mathcal{B}$ and by the additivity of measures we get

$$\mathbf{q}(\Omega \setminus \mathcal{T}_\theta) = (1 - \theta)\mathbf{q}(\Omega). \quad (\text{B.5})$$

From (B.4) and (B.5), construct X_θ as

$$X_\theta(\beta) = \begin{cases} X(\beta), & \text{for } \beta \in \mathcal{T}_\theta \\ X'(\beta), & \text{for } \beta \in \Omega \setminus \mathcal{T}_\theta \end{cases} \quad (\text{B.6})$$

and let $\mathbf{z}_\theta = \theta\mathbf{z} + (1 - \theta)\mathbf{z}'$. We claim that this pair satisfies (B.1). It is straightforward from the fact that \mathcal{X} is decomposable and that \mathcal{P} is a pointwise constraint, that $X_\theta \in \mathcal{X}$ and $X_\theta(\beta) \in \mathcal{P}$ a.e.

Let us start by showing that X_θ satisfies (B.1a). Evaluating f_0 at X_θ yields

$$\begin{aligned} f_0(X_\theta) &= \int_{\Omega} [F_0(X_\theta(\beta), \beta) + \lambda \mathbb{I}(X_\theta(\beta) \neq 0)] d\beta \\ &= \int_{\mathcal{T}_\theta} [F_0(X(\beta), \beta) + \lambda \mathbb{I}(X(\beta) \neq 0)] d\beta \\ &\quad + \int_{\Omega \setminus \mathcal{T}_\theta} [F_0(X'(\beta), \beta) + \lambda \mathbb{I}(X'(\beta) \neq 0)] d\beta. \end{aligned}$$

From (B.2), we can write these terms using the last two rows of the vector measure \mathbf{q} as

$$f_0(X_\theta) = [\mathbf{p}(\mathcal{T}_\theta)]_{2p+1} + [\mathbf{p}(\Omega \setminus \mathcal{T}_\theta)]_{2p+2}. \quad (\text{B.7})$$

Then, using (B.4) and (B.5) we obtain that

$$\begin{aligned} f_0(X_\theta) &= [\theta\mathbf{p}(\Omega)]_{2p+1} + [(1 - \theta)\mathbf{p}(\Omega)]_{2p+2} \\ &= \theta f_0(X) + (1 - \theta)f_0(X') \leq \theta c + (1 - \theta)c'. \end{aligned}$$

To proceed, notice that since the g_i are convex functions, (B.1b) obtains immediately. Explicitly,

$$\begin{aligned} g_i(\mathbf{z}_\theta) &= g_i(\theta\mathbf{z} + (1 - \theta)\mathbf{z}') \leq \theta g_i(\mathbf{z}) + (1 - \theta)g_i(\mathbf{z}') \\ &\leq [\theta\mathbf{u} + (1 - \theta)\mathbf{u}']_i. \end{aligned}$$

Finally, we can use the same machinery as in (B.7) to obtain (B.1c). Indeed,

$$\begin{aligned}
\int_{\Omega} \mathbf{F}[X_{\theta}(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} &= \int_{\mathcal{T}_{\theta}} \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} \\
&\quad + \int_{\Omega \setminus \mathcal{T}_{\theta}} \mathbf{F}[X'(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} \\
&= [\mathbf{q}(\mathcal{T}_{\theta})]_{\mathcal{S}_1} + [\mathbf{q}(\Omega \setminus \mathcal{T}_{\theta})]_{\mathcal{S}_2} \\
&= [\theta \mathbf{q}(\Omega)]_{\mathcal{S}_1} + [(1 - \theta) \mathbf{q}(\Omega)]_{\mathcal{S}_2} \\
&= \theta \mathbf{k} + (1 - \theta) \mathbf{k}' + \mathbf{z}_{\theta},
\end{aligned}$$

where $\mathcal{S}_1 = \{1, \dots, p\}$ and $\mathcal{S}_2 = \{p + 1, \dots, 2p\}$ select rows 1 through p and $p + 1$ through $2p$, respectively, of the vector measure \mathbf{q} .

To conclude, since there exists a pair $(X_{\theta}, \mathbf{z}_{\theta}) \in \mathcal{X} \times \mathbb{C}^p$ with $X_{\theta}(\boldsymbol{\beta}) \in \mathcal{P}$ a.e. and such that (B.1) holds for any $\theta \in [0, 1]$ and $(c, \mathbf{u}, \mathbf{k}_R, \mathbf{k}_I), (c', \mathbf{u}', \mathbf{k}'_R, \mathbf{k}'_I) \in \mathcal{C}$, the set \mathcal{C} is convex. Moreover, the strictly feasible pair (X', \mathbf{z}') from the hypotheses implies that \mathcal{C} cannot be empty. ■

B.1.2 A step-by-step guide to solving SFPs

Start with a problem of the form (P-SFP). Initialize $\boldsymbol{\mu}_0$ and $\nu_{i,0} > 0$; compute $d_{\mathbf{z},0} = \min_{\mathbf{z}} \sum_i \nu_{i,0} g_i(\mathbf{z}) - \mathbb{R}e[\boldsymbol{\mu}_0^H \mathbf{z}]$ and let \mathbf{z}_0 be its minimizer; evaluate

$$\gamma_0^o(\boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathbb{R}e[\boldsymbol{\mu}_0^H \mathbf{F}(x, \boldsymbol{\beta})] \quad (\text{B.8})$$

and let $\bar{X}_0(\boldsymbol{\beta})$ be its minimizer; define the initial solution support to be $\mathcal{S}_0 = \{\boldsymbol{\beta} \in \Omega : \gamma_0^o(\boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}_0, \boldsymbol{\beta}) - \lambda\}$ for $\gamma^{(0)}$ defined as in Proposition 4; obtain the primal solution

$$X_0(\boldsymbol{\beta}) = \begin{cases} \bar{X}_0(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{S}_0 \\ 0, & \text{otherwise} \end{cases}$$

and evaluate the initial dual objective using

$$d_0 = d_{\mathbf{z},0} + I \left[(\lambda + \gamma_0^o(\boldsymbol{\beta})) \times \mathbb{I}(\boldsymbol{\beta} \in \mathcal{S}_0) \right] \\ + I \left[\gamma^{(0)}(\boldsymbol{\mu}_0, \boldsymbol{\beta}) \times \mathbb{I}(\boldsymbol{\beta} \in \Omega \setminus \mathcal{S}_0) \right],$$

where I denotes a numerical integration method. Then, proceed using one of the following solvers.

Approximate supergradient ascent. Consider a numerical integration procedure represented by $I(\cdot)$ such that

$$\left| I(f) - \int_{\Omega} f(\boldsymbol{\beta}) d\boldsymbol{\beta} \right| \leq \delta, \quad (\text{B.9})$$

for $\delta > 0$ and assume that I applies element-wise to vectors. Let $X_0^* = X_0$ and for $t = 1, \dots, T$:

i) compute the supergradients

$$\mathbf{p}_{\boldsymbol{\mu},t-1} = I \left[\mathbf{F}(X_{t-1}(\boldsymbol{\beta}), \boldsymbol{\beta}) \right] - \mathbf{z}_{t-1} \\ \mathbf{p}_{\nu_i,t-1} = g_i[\mathbf{z}_{t-1}],$$

ii) update the dual variables

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \eta_t \mathbf{p}_{\boldsymbol{\mu},t-1} \\ \nu_{i,t} = [\nu_{i,t-1} + \eta_t g_i(\mathbf{z}_{t-1})]_+,$$

iii) evaluate $d_{\mathbf{z},t} = \min_{\mathbf{z}} \sum_i \nu_{i,t} g_i(\mathbf{z}) - \mathbb{R}e[\boldsymbol{\mu}_t^H \mathbf{z}]$ and let \mathbf{z}_t be its minimizer,

iv) evaluate

$$\gamma_t^o(\boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathbb{R}e[\boldsymbol{\mu}_t^H \mathbf{F}(x, \boldsymbol{\beta})],$$

and let $\bar{X}_t(\boldsymbol{\beta})$ be its minimizers,

v) evaluate the dual function

$$d_t = d_{\mathbf{z},t} + I \left[(\lambda + \gamma_t^o(\boldsymbol{\beta})) \times \mathbb{I}(\boldsymbol{\beta} \in \mathcal{S}_t) \right] \\ + I \left[\gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) \times \mathbb{I}(\boldsymbol{\beta} \in \Omega \setminus \mathcal{S}_t) \right],$$

for $\mathcal{S}_t = \{\boldsymbol{\beta} \in \Omega : \gamma_t^o(\boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) - \lambda\}$, and

vi) if $d_t > d_{t-1} + 2\delta$, obtain the primal solution

$$X_t(\boldsymbol{\beta}) = \begin{cases} \bar{X}_t(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{S}_t \\ 0, & \text{otherwise} \end{cases}$$

and let $X_t^* = X_t$. Otherwise, $X_t^* = X_{t-1}^*$.

The solution of (P-SFP) is given by X_T^* .

Stochastic supergradient ascent. Choose the mini-batch size $N \geq 1$ and initialize the solution set $\mathcal{X}_0 = \emptyset$. For $t = 1, \dots, T$:

i) draw $\{\boldsymbol{\beta}_j\}$, $j = 1, \dots, N$, uniformly at random from Ω and compute the stochastic supergradients

$$\mathbf{p}_{\boldsymbol{\mu},t-1} = \frac{1}{N} \sum_{j=1}^N \mathbf{F}[X_{t-1}(\boldsymbol{\beta}_j), \boldsymbol{\beta}] d\boldsymbol{\beta} - \mathbf{z}_{t-1}$$

$$\mathbf{p}_{\nu_i,t-1} = g_i[\mathbf{z}_{t-1}],$$

ii) update the dual variables

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \eta_t \mathbf{p}_{\boldsymbol{\mu},t-1}$$

$$\nu_{i,t} = [\nu_{i,t-1} + \eta_t g_i(\mathbf{z}_{t-1})]_+,$$

iii) evaluate $d_{\mathbf{z},t} = \min_{\mathbf{z}} \sum_i \nu_{i,t} g_i(\mathbf{z}) - \text{Re}[\boldsymbol{\mu}_t^H \mathbf{z}]$ and let \mathbf{z}_t be its minimizer,

iv) evaluate

$$\gamma_t^o(\boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathbb{R}e \left[\boldsymbol{\mu}_t^H \mathbf{F}(x, \boldsymbol{\beta}) \right]$$

and let $\bar{X}_t(\boldsymbol{\beta})$ be its minimizer,

v) evaluate the dual function

$$d_t = d_{\mathbf{z},t} + \int_{\mathcal{S}_t} [\lambda + \gamma_t^o(\boldsymbol{\beta})] d\boldsymbol{\beta} + \int_{\Omega \setminus \mathcal{S}_t} \gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) d\boldsymbol{\beta},$$

for $\mathcal{S}_t = \{\boldsymbol{\beta} \in \Omega : \gamma_t^o(\boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) - \lambda\}$, and

vi) if $d_t > d_{t-1}$, obtain the primal solution

$$X_{t-1}(\boldsymbol{\beta}) = \begin{cases} \bar{X}_{t-1}(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{S}_{t-1} \\ 0, & \text{otherwise} \end{cases}$$

and let $\mathcal{X}_t = \mathcal{X}_{t-1} \cup X_t$. Otherwise, $\mathcal{X}_t = \mathcal{X}_{t-1}$.

The final solution is obtained by averaging the elements of \mathcal{X}_T , i.e.,

$$X^*(\boldsymbol{\beta}) = \frac{1}{|\mathcal{X}_T|} \sum_{X \in \mathcal{X}_T} X(\boldsymbol{\beta}).$$

B.1.3 Proof of Proposition 7

We actually prove the following quantitative version of Proposition 7:

Proposition 23. *Under the conditions of Proposition 7, $|P^* - P_\delta^*| \leq c\delta + o(\delta^2)$ for*

$$c = \frac{\bar{F}_0 + \lambda \mathbf{m}(\Omega)}{\alpha \epsilon} \max \left(\left| \sum_i g_i(-\alpha \mathbf{1}) \right|, \left| \sum_i g_i(\alpha \mathbf{1}) \right| \right), \quad (\text{B.10})$$

where P^* is the optimal value of (P-SFP) and P_δ^* is the value of the solution obtained by Algorithm 2 when evaluating the integral in the supergradient (6.23a) with approximation error $0 < \delta \ll 1$ [as in (B.9)].

Proof. Start by noticing that evaluating the integral in (6.23a) numerically introduces an error term in the supergradient. Explicitly, (6.23a) becomes

$$\tilde{g}_{\boldsymbol{\mu}}(\boldsymbol{\mu}', \nu'_i) = \int_{\Omega} \mathbf{F}[X_d(\boldsymbol{\mu}', \boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} - \mathbf{z}_d(\boldsymbol{\mu}', \nu'_i) + \boldsymbol{\delta}, \quad (\text{B.11})$$

where $\boldsymbol{\delta}$ is an error vector whose magnitude is bounded by δ , i.e., $|\boldsymbol{\delta}|_i < \delta$. Then, observe that (B.11) is the supergradient of the dual function of a perturbed version of (P-SFP), namely

$$\begin{aligned} & \text{minimize} && \int_{\Omega} F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \lambda \|X\|_{L_0} \\ & \text{subject to} && g_i(\mathbf{z}) \leq 0 \\ & && \mathbf{z} = \int_{\Omega} \mathbf{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \boldsymbol{\delta} \\ & && X \in \mathcal{X}. \end{aligned} \quad (\text{PXIX})$$

Hence, the value P_{δ}^* of the solution obtained by the using approximate supergradient in Algorithm 2 is the optimal value of (PXIX). We can therefore use perturbation theory to relate the values of P_{δ}^* and P^* .

Formally, using the fact that the perturbation function of (P-SFP) is differentiable around zero [hypothesis (i)], we obtain the Taylor expansion $P_{\delta}^* = P^* - \boldsymbol{\mu}^{*T} \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2^2)$, where $o(t)$ is a term such that $o(t)/t \rightarrow 0$ as $t \rightarrow 0$ [130]. Hence, using the triangle inequality and the upper bound on the elements of $|\boldsymbol{\delta}|$, we can write

$$|P^* - P_{\delta}^*| = \left| \boldsymbol{\mu}^{*T} \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2^2) \right| \leq \left| \boldsymbol{\mu}^{*T} \mathbf{1} \right| \delta + o(\delta^2). \quad (\text{B.12})$$

It suffices now to bound $\left| \boldsymbol{\mu}^{*T} \mathbf{1} \right|$, which we do in two steps.

First, we obtain an upper bound on $\boldsymbol{\mu}^{*T} \mathbf{1}$ by recalling from (6.4) that the dual function d is the value of a minimization problem. Thus, taking the suboptimal $X \equiv 0$ and $\mathbf{z} = \alpha \mathbf{1}$, $\alpha > 0$, under hypothesis (ii) yields

$$d(\boldsymbol{\mu}^*, \nu_i^*) \leq \sum_i \nu_i^* g_i(\alpha \mathbf{1}) - \alpha \boldsymbol{\mu}^{*T} \mathbf{1}.$$

From Theorem 5, $d(\boldsymbol{\mu}^*, \nu_i^*) = P^* \geq 0$, which gives

$$\boldsymbol{\mu}^{*T} \mathbf{1} \leq \frac{\sum_i \nu_i^* g_i(\alpha \mathbf{1})}{\alpha}. \quad (\text{B.13})$$

Proceeding in a similar manner, we derive a lower bound by taking $X \equiv 0$ and $\mathbf{z} = -\alpha \mathbf{1}$ in (6.4), leading to

$$\boldsymbol{\mu}^{*T} \mathbf{1} \geq -\frac{\sum_i \nu_i^* g_i(-\alpha \mathbf{1})}{\alpha}. \quad (\text{B.14})$$

Using the Cauchy-Schwartz inequality, the bounds in (B.13) and (B.14) yield

$$\left| \boldsymbol{\mu}^{*T} \mathbf{1} \right| \leq \frac{\|\boldsymbol{\nu}^*\|_1}{\alpha} \max \left(\left| \sum_i g_i(-\alpha \mathbf{1}) \right|, \left| \sum_i g_i(\alpha \mathbf{1}) \right| \right), \quad (\text{B.15})$$

where $\boldsymbol{\nu}^* = [\nu_i^*]$ is a vector that collects the optimal dual variables ν_i^* . Note that since $\nu_i^* \geq 0$, we have that $|\sum_i \nu_i^*| = \|\boldsymbol{\nu}^*\|_1$. All that remains to evaluate (B.15) is to bound $\|\boldsymbol{\nu}^*\|_1$ using a classical result from optimization theory.

Explicitly, consider the strictly feasible pair $(X^\dagger, \mathbf{z}^\dagger)$ from hypothesis (iv) and recall that $g_i(\mathbf{z}^\dagger) \leq -\epsilon$ for some $\epsilon > 0$. Plugging these suboptimal values in (6.4) yields

$$d(\boldsymbol{\mu}^*, \nu_i^*) \leq \int_{\Omega} F_0 [X^\dagger(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \lambda \|X^\dagger\|_{L_0} - \sum_i \nu_i^* \epsilon. \quad (\text{B.16})$$

Recall that $\nu_i^* \geq 0$, $\epsilon > 0$, and $d(\boldsymbol{\mu}^*, \nu_i^*) \geq 0$ (from Theorem 5). Thus, using the fact that $\|X^\dagger\|_{L_0} \leq \mathbf{m}(\Omega)$, we readily obtain from (B.16) that

$$\|\boldsymbol{\nu}^*\|_1 \leq \frac{\int_{\Omega} F_0 [X^\dagger(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \lambda \mathbf{m}(\Omega)}{\epsilon}. \quad (\text{B.17})$$

Combining (B.15) and (B.17) in (B.12) we obtain that $|P^* - P_\delta^*| \leq c\delta + o(\delta^2)$ for c as in (B.10). Furthermore, hypotheses (iii) and (iv), together with $\mathbf{m}(\Omega) < \infty$ (since Ω is compact), imply that $c < \infty$, so that indeed $|P^* - P_\delta^*| \leq \mathcal{O}(\delta)$. ■

B.2 Chapter 7: Combinatorial constraints

B.2.1 Proof of Lemma 5

Start by simplifying $\Delta_u t(\mathcal{X}) = t(\mathcal{X}) - t(\mathcal{X} \cup \{u\})$ using the matrix inversion lemma. To do so, write

$$\mathbf{Y}(\mathcal{X} \cup \{u\}) = [\mathbf{R}(\mathcal{X}) + \mathbf{M}_u]^{-1}, \quad (\text{B.18})$$

letting once again $\mathbf{R}(\mathcal{X}) = \mathbf{M}_\emptyset + \sum_{u \in \mathcal{X}} \mathbf{M}_u$. Since $\mathbf{R}(\mathcal{X}) \succ 0$, but the matrices \mathbf{M}_u need not be invertible, we use an alternative form of the matrix inversion lemma [235] to get

$$\mathbf{Y}(\mathcal{X} \cup \{u\}) = \mathbf{Y}(\mathcal{X}) - \mathbf{Y}(\mathcal{X}) \mathbf{M}_u \mathbf{Y}(\mathcal{X} \cup \{u\}). \quad (\text{B.19})$$

where we used (B.18) to write $\mathbf{Y}(\mathcal{X}) = \mathbf{R}(\mathcal{X})^{-1}$. Using (B.19) and the linearity of the trace [178], we can write $\Delta_u t(\mathcal{X})$ as

$$\Delta_u t(\mathcal{X}) = \text{Tr}[\mathbf{Y}(\mathcal{X}) \mathbf{M}_u \mathbf{Y}(\mathcal{X} \cup \{u\})]. \quad (\text{B.20})$$

To proceed, let $\tilde{\mathbf{M}}_u = \mathbf{M}_u + \epsilon \mathbf{I} \succ 0$ for $\epsilon > 0$ and define the perturbed version of (B.20) as

$$\tilde{\Delta}_u t(\mathcal{X}) = \text{Tr} \left[\mathbf{Y}(\mathcal{X}) \tilde{\mathbf{M}}_u \left(\mathbf{R}(\mathcal{X}) + \tilde{\mathbf{M}}_u \right)^{-1} \right]. \quad (\text{B.21})$$

Notice that $\tilde{\Delta}_u t \rightarrow \Delta_u t$ as $\epsilon \rightarrow 0$. Using the invertibility of $\tilde{\mathbf{M}}_u$ and $\mathbf{R}(\mathcal{X})$, we obtain

$$\tilde{\Delta}_u t(\mathcal{X}) = \text{Tr} \left[\mathbf{Y}(\mathcal{X}) \left(\mathbf{Y}(\mathcal{X}) + \tilde{\mathbf{M}}_u^{-1} \right)^{-1} \mathbf{Y}(\mathcal{X}) \right].$$

Since $\mathbf{Y}(\mathcal{X}) \succ 0$, its square-root $\mathbf{Y}(\mathcal{X})^{1/2}$ is well-defined and unique [178]. We can therefore use the circular commutation property of the trace to get

$$\tilde{\Delta}_u t(\mathcal{X}) = \text{Tr}[\mathbf{Y}(\mathcal{X}) \mathbf{Z}(\mathcal{X}, u)], \quad (\text{B.22})$$

with $\mathbf{Z}(\mathcal{X}, u) = \mathbf{Y}(\mathcal{X})^{1/2} \left[\mathbf{Y}(\mathcal{X}) + \tilde{\mathbf{M}}_u^{-1} \right]^{-1} \mathbf{Y}(\mathcal{X})^{1/2}$. Since both matrices in (B.22) are positive

definite, we can use the bound from [236] to get

$$\lambda_{\min} [\mathbf{Y}(\mathcal{X})] \operatorname{Tr}(\mathbf{Z}) \leq \tilde{\Delta}_u t(\mathcal{X}) \leq \lambda_{\max} [\mathbf{Y}(\mathcal{X})] \operatorname{Tr}(\mathbf{Z}).$$

Reversing the manipulations used to obtain (B.22) yields

$$\lambda_{\min} [\mathbf{Y}(\mathcal{X})] \operatorname{Tr} \left[\tilde{\mathbf{M}}_u \left(\mathbf{Y}(\mathcal{X}) + \tilde{\mathbf{M}}_u \right)^{-1} \right] \leq \tilde{\Delta}_u t(\mathcal{X}) \leq \lambda_{\max} [\mathbf{Y}(\mathcal{X})^{-1}] \operatorname{Tr} \left[\tilde{\mathbf{M}}_u \left(\mathbf{Y}(\mathcal{X}) + \tilde{\mathbf{M}}_u \right)^{-1} \right].$$

The result in (7.17) is obtained by continuity as $\epsilon \rightarrow 0$. ■

B.2.2 Proof of Lemma 7

The proof follows a homotopy argument, i.e., we define a continuous map between $\Delta_u e(\mathcal{A})$ and $\Delta_u e(\mathcal{B})$ and bound its derivative using spectral bounds on the sum of Hermitian matrices. The inequality in (7.21) then follows from the fundamental theorem of calculus.

Start by defining the homotopy, with $t \in [0, 1]$,

$$h_{\mathcal{AB}}(t) = \left\| \mathbf{Z}(t)^{-1} \right\| - \left\| (\mathbf{Z}(t) + \mathbf{M}_u)^{-1} \right\| \quad (\text{B.23})$$

for $\mathcal{A} \subset \mathcal{B} \subseteq \mathcal{V}$, $\mathbf{Z}(t) = \mathbf{R}(\mathcal{A}) + t[\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})]$, and $\mathbf{R}(\mathcal{X}) = \mathbf{M}_\emptyset + \sum_{u \in \mathcal{X}} \mathbf{M}_u$. Note that since $\mathbf{Y}(\mathcal{X}) = \mathbf{R}(\mathcal{X})^{-1}$, we have $h_{\mathcal{AB}}(0) = \Delta_u e(\mathcal{A})$ and $h_{\mathcal{AB}}(1) = \Delta_u e(\mathcal{B})$. If $\dot{h}_{\mathcal{AB}}(t)$ is the derivative of $h_{\mathcal{AB}}$ with respect to t , it therefore holds that

$$\Delta_u e(\mathcal{B}) = \Delta_u e(\mathcal{A}) + \int_0^1 \dot{h}_{\mathcal{AB}}(t) dt. \quad (\text{B.24})$$

Comparing (B.24) to the definition of ϵ in (7.11), we obtain

$$\epsilon = \max_{\substack{\mathcal{A} \subset \mathcal{B} \subseteq \mathcal{V} \\ u \in \mathcal{V} \setminus \mathcal{B}}} \int_0^1 \dot{h}_{\mathcal{AB}}(t) dt. \quad (\text{B.25})$$

We now proceed by evaluating \dot{h} . We omit the dependence on \mathcal{A} and \mathcal{B} for conciseness. First,

recall from matrix analysis that for any $\mathbf{X} \succ 0$ we have

$$\frac{d}{dt} \|\mathbf{X}(t)^{-1}\| = \mathbf{q}(t)^T \left[\frac{d}{dt} \mathbf{X}(t)^{-1} \right] \mathbf{q}(t) = -\mathbf{q}(t)^T \mathbf{X}(t)^{-1} \dot{\mathbf{X}}(t) \mathbf{X}(t)^{-1} \mathbf{q}(t), \quad (\text{B.26})$$

where $\mathbf{q}(t)$ is the eigenvector relative to the maximum eigenvalue of $\mathbf{X}(t)$. To obtain (B.26), we used the fact that $\|\mathbf{X}\| = \lambda_{\max}(\mathbf{X})$ for $\mathbf{X} \succeq 0$ and $\frac{d}{dt} \lambda_{\max}[\mathbf{X}(t)] = \mathbf{q}(t)^T \dot{\mathbf{X}}(t) \mathbf{q}(t)$. We then used $\frac{d}{dt} \mathbf{X}(t)^{-1} = -\mathbf{X}(t)^{-1} \dot{\mathbf{X}}(t) \mathbf{X}(t)^{-1}$ [179]. In view of (B.23) and (B.26), we obtain

$$\begin{aligned} \dot{h}(t) &= \mathbf{w}(t)^T [\mathbf{Z}(t) + \mathbf{M}_u]^{-1} [\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})] [\mathbf{Z}(t) + \mathbf{M}_u]^{-1} \mathbf{w}(t) \\ &\quad - \mathbf{u}(t)^T \mathbf{Z}(t)^{-1} [\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})] \mathbf{Z}(t)^{-1} \mathbf{u}(t), \end{aligned} \quad (\text{B.27})$$

where $\mathbf{u}(t)$ and $\mathbf{w}(t)$ are the eigenvectors relative to the maximum eigenvalues of $\mathbf{Z}(t)^{-1}$ and $[\mathbf{Z}(t) + \mathbf{M}_u]^{-1}$ respectively.

We now proceed by finding a bound for \dot{h} in (B.27) that is independent of t . To do so, observe that $\mathbf{R}(\mathcal{B}) \succeq \mathbf{R}(\mathcal{A})$, since $\mathcal{A} \subseteq \mathcal{B}$ and \mathbf{R} is monotone increasing (Lemma 4). By the Loewner ordering, the second term in (B.27) is therefore non-positive. Thus, using Rayleigh's spectral inequality and the fact $\|\mathbf{w}(t)\| = 1$ for all t yields

$$\dot{h}(t) \leq \left\| (\mathbf{Z}(t) + \mathbf{M}_u)^{-1} [\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})] (\mathbf{Z}(t) + \mathbf{M}_u)^{-1} \right\|.$$

Next, we once again use the fact that $\mathbf{R}(\mathcal{B}) \succeq \mathbf{R}(\mathcal{A})$ to obtain $\mathbf{Z}(t) \succeq \mathbf{Z}(0) = \mathbf{R}(\mathcal{A})$, effectively removing the dependence on t . Using Cauchy-Schwartz then yields

$$\dot{h}(t) \leq \lambda_{\max} \left[(\mathbf{R}(\mathcal{A}) + \mathbf{M}_u)^{-2} \right] \lambda_{\max} [\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})],$$

which can be used in (B.25) to get

$$\epsilon \leq \max_{\substack{\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \\ u \in \mathcal{V} \setminus \mathcal{B}}} \frac{\lambda_{\max} [\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})]}{\lambda_{\min} [\mathbf{R}(\mathcal{A}) + \mathbf{M}_u]^2}, \quad (\text{B.28})$$

The inequality in (7.21) can then be obtained by using $\|\mathbf{R}(\mathcal{B}) - \mathbf{R}(\mathcal{A})\| \leq \lambda_{\max}(\sum_{u \in \mathcal{V}} \mathbf{M}_u)$ and $\lambda_{\min} [\mathbf{R}(\mathcal{A}) + \mathbf{M}_u] \geq \lambda_{\min} [\mathbf{M}_\emptyset]$. ■

B.2.3 Proof of Theorem 8

Since f is monotone decreasing, it holds for every set \mathcal{X}_t that

$$\begin{aligned} f(\mathcal{X}^*) &\geq f(\mathcal{X}^* \cup \mathcal{X}_t) \\ &= f(\mathcal{X}_t) - \sum_{i=1}^s [f(\mathcal{T}_{i-1}) - f(\mathcal{T}_{i-1} \cup v_i^*)], \end{aligned} \quad (\text{B.29})$$

where $\mathcal{T}_0 = \mathcal{X}_t$, $\mathcal{T}_i = \mathcal{X}_t \cup \{v_1^*, \dots, v_i^*\}$, $i = 1, \dots, s$, and v_i^* is the i -th element of \mathcal{X}^* . Notice that this holds regardless of the order in which the v_i^* are taken. Since f is α -supermodular and $\mathcal{X}_t \subseteq \mathcal{T}_i$ for all i , the incremental gains in (B.29) can be bounded using (7.2) to get

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) - \alpha^{-1} \sum_{i=1}^s [f(\mathcal{X}_t) - f(\mathcal{X}_t \cup v_i^*)]. \quad (\text{B.30})$$

To proceed, we use the fact that $\mathcal{X}_{t+1} = \mathcal{X}_t \cup \{u\}$ is constructed by the greedy procedure in Algorithm 3 so as to maximize the incremental gains in (B.30). It therefore holds that

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) - \alpha^{-1} s [f(\mathcal{X}_t) - f(\mathcal{X}_{t+1})]. \quad (\text{B.31})$$

By adding and subtracting $f(\mathcal{X}^*)$ in the brackets of (B.31), (B.31) yields a recursion for the distance to the optimal value $\delta_t = f(\mathcal{X}^*) - f(\mathcal{X}_t)$ given by

$$\delta_t \geq \alpha^{-1} s [\delta_t - \delta_{t+1}] \Rightarrow \delta_{t+1} \geq \left(1 - \frac{1}{\alpha^{-1} s}\right) \delta_t.$$

Note that since f is non-positive, $\delta_t \leq 0$ for all t by the optimality of \mathcal{X}^* .

The expression in (7.22) yields directly from this recursion by noticing that since f is normalized, we have that $\delta_0 = f(\mathcal{X}^*) - f(\emptyset) = f(\mathcal{X}^*)$. Immediately, since Algorithm 3 is used for r iterations, we obtain

$$f(\mathcal{X}^*) - f(\mathcal{X}_r) \geq \left(1 - \frac{\alpha}{s}\right)^r f(\mathcal{X}^*). \quad (\text{B.32})$$

Using the fact that $1 - x \leq e^{-x}$, the expression in (B.32) can be rearranged into (7.22). ■

B.3 Proof of Theorem 9

Using the assumption that f is monotone decreasing, we write

$$\begin{aligned} f(\mathcal{X}^*) &\geq f(\mathcal{X}^* \cup \mathcal{X}_t) \\ &= f(\mathcal{X}_t) - \sum_{i=1}^s [f(\mathcal{T}_{i-1}) - f(\mathcal{T}_{i-1} \cup \{v_i^*\})], \end{aligned} \quad (\text{B.33})$$

where again $\mathcal{T}_0 = \mathcal{X}_t$, $\mathcal{T}_i = \mathcal{X}_t \cup \{v_1^*, \dots, v_i^*\}$, for $i = 1, \dots, s$, and v_i^* is the i -th element of \mathcal{X}^* .

Observe that the order in which the v_i^* occur in the telescopic sum is irrelevant. Since f is ϵ -supermodular and $\mathcal{X}_t \subseteq \mathcal{T}_i$ for all i , (7.10) can be used to bound the incremental gains in (B.33).

Explicitly,

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) - \sum_{i=1}^s [f(\mathcal{X}_t) - f(\mathcal{X}_t \cup \{v_i^*\}) + \epsilon].$$

Since $\mathcal{X}_{t+1} = \mathcal{X}_t \cup \{u\}$ is constructed by Algorithm 3 so as to minimize $f(\mathcal{X}_{t+1})$, it holds that

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) - s [f(\mathcal{X}_t) - f(\mathcal{X}_{t+1}) + \epsilon]. \quad (\text{B.34})$$

Finally, a recursion for the optimality gap $\delta'_t = f(\mathcal{X}_t) - f(\mathcal{X}^*)$ is obtained from (B.34) by adding and subtracting $f(\mathcal{X}^*)$ in the brackets. Explicitly,

$$\delta'_t \leq s (\delta'_t - \delta'_{t+1} + \epsilon) \Rightarrow \delta'_{t+1} \leq \left(1 - \frac{1}{s}\right) \delta'_t + \epsilon. \quad (\text{B.35})$$

Notice that since f is non-positive, $\delta'_t \geq 0$ for all j due to the optimality of \mathcal{X}^* . The solution of (B.35) after r steps is

$$\delta'_r \leq \left(1 - \frac{1}{s}\right)^r \delta'_0 + \epsilon \sum_{j=0}^{r-1} \left(1 - \frac{1}{s}\right)^j \quad (\text{B.36})$$

Evaluating the geometric series in (B.36) yields

$$\delta'_r \leq \left(1 - \frac{1}{s}\right)^r \delta'_0 + s \left[1 - \left(1 - \frac{1}{s}\right)^r\right] \epsilon$$

and under the assumption that f is normalized, i.e., for $\delta'_0 = f(\emptyset) - f(\mathcal{X}^*) = -f(\mathcal{X}^*) \geq 0$, we obtain

$$f(\mathcal{X}_r) \leq \left[1 - \left(1 - \frac{1}{s}\right)^r\right] [f(\mathcal{X}^*) + s \cdot \epsilon].$$

Using once again the fact that $1 - x \leq e^{-x}$ yields (7.23). ■

B.3.1 Proof of Theorem 10

This proof follows similarly to the one for the cardinality constraint (uniform matroid) problem, but relying on the exchange property of matroids through Proposition 12. It may be informative to recall the proof technique in that simpler case first, e.g., by referring to [37, Thm. 1].

Let $|\mathcal{X}^*| = r$. Partition the optimal solution in (P-MTRD) such that $\mathcal{X}^* = \bigcup_{j=0}^{\lfloor r/P \rfloor} \mathcal{X}_j^*$, $|\mathcal{X}_j^*| \leq P$, and

$$\mathcal{X}_t \cup \{x^*\} \in \mathcal{I} \text{ for } x^* \in \mathcal{X}_t^* \text{ for all } t = 0, \dots, \left\lfloor \frac{r}{P} \right\rfloor. \quad (\text{B.37})$$

Observe that such a partition exists due to the matroid exchange property from Proposition 12. Fix an arbitrary enumeration of each partition $\mathcal{X}_t^* = \{x_{t1}^*, \dots, x_{tp}^*\}$. The following proof is independent of the specific enumeration.

Use the fact that f is monotone decreasing to write

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_T \cup \mathcal{X}^*) = f(\mathcal{X}_T) - \sum_{t=0}^{\lfloor r/P \rfloor} \left[\sum_{j=1}^{|\mathcal{X}_t^*|} \Delta_{x_{tj}^*} f(\mathcal{T}_t^{j-1}) \right], \quad (\text{B.38})$$

where the set \mathcal{T}_t^j contains the greedy solution \mathcal{X}_T , the elements of all partitions \mathcal{X}_i^* up to $t-1$, and the first j elements of partition \mathcal{X}_t^* . They can be defined recursively as $\mathcal{T}_0^0 = \mathcal{X}_T$, $\mathcal{T}_t^0 = \mathcal{T}_{t-1}^0 \cup \mathcal{X}_t^*$, and $\mathcal{T}_t^j = \mathcal{T}_{t-1}^0 \cup \{x_{t1}^*, \dots, x_{tj}^*\}$. Then, we can bound (B.38) by distinguishing two cases: (i) if $x_{tj}^* \notin \mathcal{T}_t^{j-1}$, the α -supermodularity of f yields

$$\Delta_{x_{tj}^*} f(\mathcal{T}_t^{j-1}) \leq \alpha^{-1} \Delta_{x_{tj}^*} f(\mathcal{X}_t), \quad (\text{B.39})$$

since $\mathcal{X}_t \subseteq \mathcal{X}_T \subseteq \mathcal{T}_t^j$ for all t and j ; (ii) if $x_{tj}^* \in \mathcal{T}_t^{j-1}$, then $\Delta_{x_{tj}^*} f(\mathcal{T}_t^{j-1}) = 0$ and (B.39) holds

trivially. Using (B.39) in (B.38) gives

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_T) - \alpha^{-1} \sum_{t=0}^{\lfloor r/P \rfloor} \left[\sum_{j=1}^{|\mathcal{X}_t^*|} \Delta_{x_{t,j}^*} f(\mathcal{X}_t) \right]. \quad (\text{B.40})$$

The bound in (B.40) can be simplified using the greedy nature of Algorithm (4). Indeed, observe that (B.37) implies that $\Delta_{x_{t,j}^*} f(\mathcal{X}_t) \leq \Delta_{g_t} f(\mathcal{X}_t)$ for g_t is the t -th element selected as in step 3. Hence,

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_T) - \alpha^{-1} \sum_{t=0}^{\lfloor r/P \rfloor} \left[\sum_{j=1}^{|\mathcal{X}_t^*|} \Delta_{g_t} f(\mathcal{X}_t) \right]. \quad (\text{B.41})$$

Using the fact that the increments are always non-negative [see (7.20)] and since $|\mathcal{X}_t^*| \leq P$, (B.41) reduces to

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_T) - \alpha^{-1} P \sum_{t=0}^{\lfloor r/P \rfloor} \Delta_{g_t} f(\mathcal{X}_t). \quad (\text{B.42})$$

To conclude, observe from (7.20) that $f(\mathcal{X}_T) \leq f(\mathcal{X}_t) = -\sum_{t=0}^{T-1} \Delta_{g_t} f(\mathcal{X}_t)$ and recall from Proposition 12 that $T \geq r/P$. Thus,

$$f(\mathcal{X}^*) \geq (1 + \alpha^{-1} P) f(\mathcal{X}_T),$$

which can be rearranged as in (7.27). ■

B.3.2 Proof of Theorem 11

Since f is monotone decreasing, it holds for every \mathcal{X}_t that

$$f(\mathcal{X}^*) \geq f(\mathcal{D}^* \cup \mathcal{X}_t) = f(\mathcal{X}_t) + \sum_{k=0}^{s-1} f(\mathcal{T}_k \cup \{e_k^*\}) - f(\mathcal{T}_k), \quad (\text{B.43})$$

where $\mathcal{T}_k = \mathcal{X}_t \cup \{e_0^*, \dots, e_{k-1}^*\}$, with $\mathcal{T}_0 = \mathcal{X}_t$, and e_k^* is the k -th element of \mathcal{X}^* . The equality comes from expressing the set function as a telescopic sum. Since f is α -supermodular and $\mathcal{X}_t \subseteq \mathcal{T}_k$ for all k , the incremental gains in (B.43) can be bounded using (7.28) to get

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) + \sum_{k=0}^{s-1} \alpha(t, t+k)^{-1} [f(\mathcal{X}_t \cup \{e_k^*\}) - f(\mathcal{X}_t)].$$

Given that \mathcal{X}_{t+1} is constructed from \mathcal{X}_t so as to minimize $f(\mathcal{X}_{t+1})$ [as per step 3 of Algorithm (3)], it holds that

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) + [f(\mathcal{X}_{t+1}) - f(\mathcal{X}_t)] \sum_{k=0}^{s-1} \alpha(t, t+k)^{-1}. \quad (\text{B.44})$$

A recursion is obtained by taking $\delta_t = f(\mathcal{X}_t) - f(\mathcal{X}^*)$, so that (B.44) can be written as

$$\delta_t \leq \alpha'(t) (\delta_t - \delta_{t+1}) \Rightarrow \delta_{t+1} \leq \left(1 - \frac{1}{\alpha'(t)}\right) \delta_t,$$

with $\alpha'(t) = \sum_{k=0}^{s-1} \alpha(t, t+k)^{-1}$. Considering that f is normalized, $\delta_0 = -f(\mathcal{X}^*)$ and the solution of this recursion is

$$f(\mathcal{X}_r) \leq \left[1 - \prod_{t=0}^{r-1} \left(1 - \frac{1}{\alpha'(t)}\right)\right] f(\mathcal{X}^*).$$

Since $\bar{\alpha} \leq \alpha(t, t+k)$ for $t < r$ and $k < s$, it holds that $\alpha'(t) \leq \bar{\alpha}^{-1}t$. Then, using the fact that $1 - x \leq e^{-x}$ yields (7.29). ■

B.3.3 Proof of Theorem 12

Given that f is monotone decreasing,

$$f(\mathcal{X}^*) \geq f(\mathcal{D}^* \cup \mathcal{X}_t) = f(\mathcal{X}_t) + \sum_{k=0}^{s-1} f(\mathcal{T}_k \cup \{e_k^*\}) - f(\mathcal{T}_k), \quad (\text{B.45})$$

where $\mathcal{T}_k = \mathcal{X}_t \cup \{e_0^*, \dots, e_{k-1}^*\}$, with $\mathcal{T}_0 = \mathcal{X}_t$, and e_k^* is the k -th experiment in \mathcal{D}^* . Since f is ϵ -supermodular and $\mathcal{X}_t \subseteq \mathcal{T}_k$ for all k , (7.30) can be used to bound the incremental gains in (B.45).

Explicitly,

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) + \sum_{k=0}^{s-1} [f(\mathcal{X}_t \cup \{e_k^*\}) - f(\mathcal{X}_t) + \epsilon(t, t+k)].$$

Since \mathcal{X}_{t+1} is chosen greedily to minimize $f(\mathcal{X}_{t+1})$ [see step 3 of Algorithm (3)], it holds that

$$f(\mathcal{X}^*) \geq f(\mathcal{X}_t) + s [f(\mathcal{X}_{t+1}) - f(\mathcal{X}_t)] + \sum_{k=0}^{s-1} \epsilon(t, t+k). \quad (\text{B.46})$$

The following recursion is obtained from (B.46) by letting $\delta_t = f(\mathcal{X}_t) - f(\mathcal{X}^*)$:

$$\delta_t \leq s(\delta_t - \delta_{t+1}) - \epsilon'(t) \Rightarrow \delta_{t+1} \leq \left(1 - \frac{1}{s}\right) \delta_t + \frac{\epsilon'(t)}{s}$$

with $\epsilon'(t) = \sum_{k=0}^{s-1} \epsilon(t, t+k)$. Since f is normalized, $\delta_0 = -f(\mathcal{X}^*)$, and solving this recursion yields

$$f(\mathcal{X}_r) \leq \left[1 - \left(1 - \frac{1}{s}\right)^r\right] f(\mathcal{X}^*) + \frac{1}{s} \sum_{t=0}^{r-1} \epsilon'(t) \left(1 - \frac{1}{s}\right)^{r-1-t}.$$

Using the fact that $\bar{\epsilon} \geq \epsilon(t, t+k)$ for $t < r$ and $k < s$ then gives

$$f(\mathcal{X}_r) \leq \left[1 - \left(1 - \frac{1}{s}\right)^r\right] f(\mathcal{X}^*) + \bar{\epsilon} \sum_{t=0}^{r-1} \left(1 - \frac{1}{s}\right)^t,$$

from which (7.31) obtains using the closed form of the geometric series and $1 - x \leq e^{-x}$. ■

B.3.4 Proof of Proposition 14

Part (i) is a corollary of Lemma 4. To prove part (ii), let $\mathbf{Z}(\mathcal{A}) = \mathbf{\Lambda}^{-1} + \sum_{i \in \mathcal{A}} \lambda_{w,i}^{-1} \mathbf{v}_i \mathbf{v}_i^H$ so that the increment in (7.6) reads

$$f(\mathcal{A} \cup \{u\}) - f(\mathcal{A}) = \text{Tr} \left[\mathbf{W} (\mathbf{Z}(\mathcal{A}) + \lambda_{w,u}^{-1} \mathbf{v}_u \mathbf{v}_u^H)^{-1} - \mathbf{W} \mathbf{Z}(\mathcal{A})^{-1} \right],$$

which using the matrix inversion lemma simplifies to

$$f(\mathcal{A} \cup \{u\}) - f(\mathcal{A}) = -\text{Tr} \left[\mathbf{W} \frac{\mathbf{Z}(\mathcal{A})^{-1} \mathbf{v}_u \mathbf{v}_u^H \mathbf{Z}(\mathcal{A})^{-1}}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{Z}(\mathcal{A})^{-1} \mathbf{v}_u} \right] = -\frac{\mathbf{v}_u^H \mathbf{Z}(\mathcal{A})^{-1} \mathbf{W} \mathbf{Z}(\mathcal{A})^{-1} \mathbf{v}_u}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{Z}(\mathcal{A})^{-1} \mathbf{v}_u}.$$

Using this expression, the expression for α in (7.6) becomes

$$\alpha = \min_{\substack{\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \\ u \notin \mathcal{B}}} \frac{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{Z}(\mathcal{B})^{-1} \mathbf{v}_u}{\lambda_{w,u}^{-1} + \mathbf{v}_u^H \mathbf{Z}(\mathcal{A})^{-1} \mathbf{v}_u} \frac{\mathbf{v}_u^H \mathbf{Z}(\mathcal{A})^{-1} \mathbf{W} \mathbf{Z}(\mathcal{A})^{-1} \mathbf{v}_u}{\mathbf{v}_u^H \mathbf{Z}(\mathcal{B})^{-1} \mathbf{W} \mathbf{Z}(\mathcal{B})^{-1} \mathbf{v}_u}. \quad (\text{B.47})$$

We now bound (B.47) by noticing that for any set $\mathcal{X} \subseteq \mathcal{V}$

$$\mu_{\min} \leq \lambda_{\min} [\mathbf{\Lambda}^{-1}] \leq \lambda_{\min} [\mathbf{Z}(\mathcal{X})] \leq \lambda_{\max} [\mathbf{Z}(\mathcal{X})] \leq \lambda_{\max} [\mathbf{\Lambda}^{-1} + \mathbf{V}_{\mathcal{K}}^H \mathbf{\Lambda}_w^{-1} \mathbf{V}_{\mathcal{K}}] \leq \mu_{\max}.$$

Thus, using the Rayleigh quotient inequalities leads to

$$\alpha \geq \frac{\lambda_{w,u}^{-1} + \|\mathbf{v}_u\|_2^2 \lambda_{\max}[\mathbf{Z}(\mathcal{B})]^{-1} \lambda_{\min}[\mathbf{Z}(\mathcal{A})^{-1} \mathbf{W} \mathbf{Z}(\mathcal{A})^{-1}]}{\lambda_{w,u}^{-1} + \|\mathbf{v}_u\|_2^2 \lambda_{\min}[\mathbf{Z}(\mathcal{A})]^{-1} \lambda_{\max}[\mathbf{Z}(\mathcal{B})^{-1} \mathbf{W} \mathbf{Z}(\mathcal{B})^{-1}]},$$

which can be simplified using the singular value bounds in [237, Thm. 9.H.1, p. 338] to yield

$$\alpha \geq \frac{\lambda_{w,u}^{-1} + \|\mathbf{v}_u\|_2^2 \mu_{\max}^{-1}}{\lambda_{w,u}^{-1} + \|\mathbf{v}_u\|_2^2 \mu_{\min}^{-1}} \frac{\mu_{\max}^{-2}}{\kappa_2(\mathbf{W}) \mu_{\min}^{-2}} \triangleq \alpha', \quad (\text{B.48})$$

where again $\kappa_2(\mathbf{W}) = \lambda_{\max}(\mathbf{W})/\lambda_{\min}(\mathbf{W})$ is the 2-norm condition number of \mathbf{W} . To obtain the expression in (7.45), notice that (B.48) is decreasing with respect to $\|\mathbf{v}_u\|_2^2$ and $\lambda_{w,u}^{-1}$. Indeed,

$$\begin{aligned} \frac{\partial \alpha'}{\partial \|\mathbf{v}_u\|_2^2} &= \frac{\mu_{\max}^{-2}}{\kappa_2(\mathbf{W}) \mu_{\min}^{-2}} \frac{\lambda_{w,u}^{-1} (\mu_{\max}^{-1} - \mu_{\min}^{-1})}{\left(\lambda_{w,u}^{-1} + \|\mathbf{v}_u\|_2^2 \mu_{\min}^{-1}\right)^2} \leq 0 \\ \frac{\partial \alpha'}{\partial \lambda_{w,u}^{-1}} &= \frac{\mu_{\max}^{-2}}{\kappa_2(\mathbf{W}) \mu_{\min}^{-2}} \frac{\lambda_{w,u}^{-2} \|\mathbf{v}_u\|_2^2 (\mu_{\max}^{-1} - \mu_{\min}^{-1})}{\left(\lambda_{w,u}^{-1} + \|\mathbf{v}_u\|_2^2 \mu_{\min}^{-1}\right)^2} \leq 0 \end{aligned}$$

are both non-positive because $0 < \mu_{\min} \leq \mu_{\max}$ and $\kappa_2(\mathbf{W}) \geq 1$ [178]. We then use the fact that $\|\mathbf{v}_u\|_2^2 \leq 1$ to get (7.45). ■

B.3.5 Proof of Proposition 15

This result follows directly from the classical dynamic programming argument for the LQG by considering only the inputs in \mathcal{S} (see, e.g., [238]). We display the derivations here for ease of reference. Explicitly, we proceed by backward induction, first defining the cost-to-go function

$$V_j(\mathcal{S}) = \min_{\mathcal{U}_j(\mathcal{S})} \mathbb{E} \left[\sum_{k=j}^{N-1} \left(\mathbf{x}_k^T \mathbf{Q}_k \mathbf{x}_k + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k} u_{i,k}^2 \right) + \mathbf{x}_N^T \mathbf{Q}_N \mathbf{x}_N \right], \quad (\text{B.49})$$

where $\mathcal{U}_j(\mathcal{S}) = \{u_{i,k} | i \in \mathcal{S} \cap \mathcal{V}_k\}_{k=j}^{N-1}$ is the subsequence of control actions from time j to the end of the control horizon. Notice due to the additivity of (B.49), it admits the equivalent recursive definition

$$V_j(\mathcal{S}) = \min_{\mathcal{U}_j(\mathcal{S})} \mathbb{E} \left[V_{j+1}(\mathcal{S}) + \mathbf{x}_j^T \mathbf{Q}_j \mathbf{x}_j + \sum_{i \in \mathcal{S} \cap \mathcal{V}_j} r_{i,j} u_{i,j}^2 \right]. \quad (\text{B.50})$$

To proceed, we postulate that (B.49) has the form

$$V_j(\mathcal{S}) = \mathbf{x}_j^T \mathbf{P}_j(\mathcal{S}) \mathbf{x}_j + q_j, \quad (\text{B.51})$$

for some $\mathbf{P}_j(\mathcal{S}) \succ 0$ and $q_j > 0$ and show that this is indeed the case by recursion. To do so, observe that for the base case $j = N$, (B.49) becomes $V_N(\mathcal{S}) = \mathbf{x}_N^T \mathbf{Q}_N \mathbf{x}_N$ and (B.51) holds trivially by taking $\mathbf{P}_N(\mathcal{S}) = \mathbf{Q}_N \succ 0$ and $q_N = 0$. Now, assume that (B.51) holds for $j + 1$. Using the system dynamics (7.55) in (B.51), we can expand $V_j(\mathcal{S})$ to read

$$\begin{aligned} V_j(\mathcal{S}) = \min_{\mathcal{U}_j(\mathcal{S})} \mathbb{E} & \left[\mathbf{x}_j^T (\mathbf{Q}_j + \mathbf{A}_j^T \mathbf{P}_{j+1} \mathbf{A}_j) \mathbf{x}_j + \left(\sum_{i \in \mathcal{S} \cap \mathcal{V}_j} \mathbf{b}_{i,j} u_{i,j} \right)^T \mathbf{P}_{j+1}(\mathcal{S}) \left(\sum_{i \in \mathcal{S} \cap \mathcal{V}_j} \mathbf{b}_{i,j} u_{i,j} \right) \right. \\ & + 2\mathbf{x}_j^T \mathbf{A}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \left(\sum_{i \in \mathcal{S} \cap \mathcal{V}_j} \mathbf{b}_{i,j} u_{i,j} \right) + 2\mathbf{w}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \left(\sum_{i \in \mathcal{S} \cap \mathcal{V}_j} \mathbf{b}_{i,j} u_{i,j} \right) \\ & \left. + 2\mathbf{x}_j^T \mathbf{A}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \mathbf{w}_j + \mathbf{w}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \mathbf{w}_j + \sum_{i \in \mathcal{S} \cap \mathcal{V}_j} r_{i,j} u_{i,j}^2 + q_{j+1} \right]. \quad (\text{B.52}) \end{aligned}$$

Recall that the expected value is taken over the process noises noise \mathbf{w}_k and the initial state \mathbf{x}_0 . Note that the terms linear in \mathbf{w}_j vanish since it is zero-mean and that (B.52) is actually a quadratic optimization problem in the $\{u_{i,j}\}$. This is straightforward to see by rewriting (B.52) in matrix form:

$$\begin{aligned} V_j(\mathcal{S}) = \min_{\mathbf{u}_j} & \mathbf{u}_j^T (\mathbf{R}_j + \mathbf{B}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \mathbf{B}_j) \mathbf{u}_j + 2\mathbf{x}_j^T \mathbf{A}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \mathbf{B}_j \mathbf{u}_j \\ & + \mathbf{x}_j^T (\mathbf{Q}_j + \mathbf{A}_j^T \mathbf{P}_{j+1} \mathbf{A}_j) \mathbf{x}_j + \text{Tr}(\mathbf{P}_{j+1}(\mathcal{S}) \mathbf{W}_j) + q_{j+1}, \end{aligned} \quad (\text{B.53})$$

where $\mathbf{u}_j = [u_{i,j}]_{i \in \mathcal{S} \cap \mathcal{V}_j}$ is a $|\mathcal{S} \cap \mathcal{V}_j| \times 1$ vector that collects the control actions, $\mathbf{B}_j = [\mathbf{b}_{i,j}]_{i \in \mathcal{S} \cap \mathcal{V}_j}$ is an $n \times |\mathcal{S} \cap \mathcal{V}_j|$ matrix whose columns contain the input vectors corresponding to each control action, $\mathbf{R}_j = \text{diag}(r_{i,j})$, and we used the that $\mathbb{E}[\mathbf{w}_j^T \mathbf{P}_{j+1}(\mathcal{S}) \mathbf{w}_j] = \text{Tr}(\mathbf{P}_{j+1}(\mathcal{S}) \mathbf{W}_j)$. Observe that (B.53) is classical LQR problem, up to constant terms [238]. Its minimum is then of the form (B.51) with

$$\mathbf{P}_j(\mathcal{S}) = \mathbf{Q}_j + \mathbf{A}_j^T \left(\mathbf{P}_{j+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_j} r_{i,j}^{-1} \mathbf{b}_{i,j} \mathbf{b}_{i,j}^T \right)^{-1} \mathbf{A}_j,$$

and $q_j = \text{Tr}(\mathbf{P}_{j+1}(\mathcal{S}) \mathbf{W}_j) + q_{j+1}$. Unrolling the recursion and using the fact that $\mathbb{E}[\mathbf{x}_0 \mathbf{x}_0^T] = \mathbf{\Sigma}_0$

yields the result in (7.57). ■

B.3.6 Proof of Proposition 16

Start by defining

$$J_k(\mathcal{S}) = \text{Tr} [\mathbf{\Pi}_{k-1} \mathbf{P}_k(\mathcal{S})] \quad (\text{B.54})$$

with $\mathbf{\Pi}_{-1} = \mathbf{\Sigma}_0$ and $\mathbf{\Pi}_k = \mathbf{W}_k$ for $k = 0, \dots, N-2$ and from (7.57), notice that the objective J of (PXIII) can be written as

$$J(\mathcal{S}) = \sum_{k=0}^{N-1} J_k(\mathcal{S}) + \text{Tr} [\mathbf{W}_{N-1} \mathbf{Q}_N] - V^*(\emptyset). \quad (\text{B.55})$$

Hence, using the non-negative affine combination property in Proposition 3, we can ignore the terms constant in \mathcal{S} and lower bound the approximate supermodularity of J in terms of the approximate supermodularity of its components. Explicitly, if J_k in (B.54) is α_k -supermodular, then J is $\min(\alpha_k)$ -supermodular.

We can further reduce the problem by using (7.58) to get

$$J_k(\mathcal{S}) = \text{Tr} \left[\mathbf{H}_k \left(\mathbf{P}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \right] + \text{Tr} [\mathbf{\Pi}_{k-1} \mathbf{Q}_k]. \quad (\text{B.56})$$

where we used the circular shift property of the trace to get $\mathbf{H}_k = \mathbf{A}_k \mathbf{\Pi}_{k-1} \mathbf{A}_k^T$. Thus, applying Lemma 3 again, we can ignore the constant term in (B.56) and restrict to studying the α -supermodularity of

$$\bar{J}_k(\mathcal{S}) = \text{Tr} \left[\mathbf{H}_k \left(\mathbf{P}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \right]. \quad (\text{B.57})$$

To proceed, notice that $\mathbf{H}_k \succ 0$ since $\mathbf{\Pi}_k \succ 0$ and \mathbf{A}_k is full rank for all k . Hence, it has a unique positive definite square root $\mathbf{H}^{1/2} \succ 0$ such that $\mathbf{H} = \mathbf{H}^{1/2} \mathbf{H}^{1/2}$ [178]. Deploying the circular shift

property of the trace and the invertibility of $\mathbf{H}^{1/2}$, (B.57) can be written as

$$\bar{J}_k(\mathcal{S}) = \text{Tr} \left[\left(\tilde{\mathbf{P}}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k}^{-1} \tilde{\mathbf{b}}_{i,k} \tilde{\mathbf{b}}_{i,k}^T \right)^{-1} \right]. \quad (\text{B.58})$$

with $\tilde{\mathbf{P}}_{k+1}(\mathcal{S}) = \mathbf{H}_k^{1/2} \mathbf{P}_{k+1}(\mathcal{S}) \mathbf{H}_k^{1/2}$ and $\tilde{\mathbf{b}}_{i,k} = \mathbf{H}_k^{-1/2} \mathbf{b}_{i,k}$. The function \bar{J}_k in (B.58) has a form that allows us to leverage Theorem 6.

Comparing (B.58) and (7.12), we can bound the α_k for J_k as

$$\alpha_k \geq \min_{\mathcal{S} \subseteq \bar{\mathcal{V}}} \frac{\lambda_{\min} [\tilde{\mathbf{P}}_{k+1}^{-1}(\mathcal{S})]}{\lambda_{\max} [\tilde{\mathbf{P}}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{V}_k} r_{i,k}^{-1} \tilde{\mathbf{b}}_{i,k} \tilde{\mathbf{b}}_{i,k}^T]}. \quad (\text{B.59})$$

Nevertheless, the bound in (B.59) depends on the choice of \mathcal{S} . To obtain a closed-form expression, we use the following proposition:

Proposition 24. *For any $\mathcal{S} \subseteq \bar{\mathcal{V}}$, it holds that*

$$\tilde{\mathbf{P}}_k(\bar{\mathcal{V}}) \preceq \tilde{\mathbf{P}}_k(\mathcal{S}) \preceq \tilde{\mathbf{P}}_k(\emptyset), \text{ for } k = 1, \dots, N-1. \quad (\text{B.60})$$

Using (B.60) in (B.59) and the fact that matrix inversion is operator antitone yields the bound in (7.16). ■

All that remains is therefore to prove Proposition 24.

Proof of Proposition 24. Start by noticing that since $\mathbf{H}^{1/2}$ is full rank, it is enough to establish (24) directly for \mathbf{P}_k . Indeed, $\mathbf{A} \preceq \mathbf{B} \Leftrightarrow \mathbf{x}^T \mathbf{A} \mathbf{x} \leq \mathbf{x}^T \mathbf{B} \mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$ and for $\mathbb{M} \succ 0$, this is equivalent to taking $\mathbf{x} = \mathbb{M} \mathbf{y}$ for all $\mathbf{y} \in \mathbb{R}^n$.

We prove both inequalities by recursion. For the upper bound in (B.60), note from (7.58) that $\tilde{\mathbf{P}}_k$ can be increased by using no actuators at instant k . Formally, for any choice of $\mathcal{S} \subseteq \bar{\mathcal{V}}$,

$$\begin{aligned} \mathbf{P}_k(\mathcal{S}) &= \mathbf{Q}_k + \mathbf{A}_k^T \left(\mathbf{P}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \mathbf{A}_k \\ &\preceq \mathbf{Q}_k + \mathbf{A}_k^T \mathbf{P}_{k+1}(\mathcal{S}) \mathbf{A}_k = \mathbf{P}_k(\mathcal{S} \setminus \mathcal{V}_k). \end{aligned}$$

Additionally, if $\bar{\mathbf{P}}_{k+1} \succeq \mathbf{P}_{k+1}(\mathcal{S})$ for all $\mathcal{S} \subseteq \bar{\mathcal{V}}$, it holds that

$$\mathbf{P}_k(\mathcal{S}) \preceq \mathbf{Q}_k + \mathbf{A}_k^T \bar{\mathbf{P}}_{k+1} \mathbf{A}_k \triangleq \bar{\mathbf{P}}_k. \quad (\text{B.61})$$

Starting from $\mathbf{P}_N \preceq \mathbf{Q}_N \triangleq \bar{\mathbf{P}}_N$, we obtain that \mathbf{P}_k is upper bounded by taking $\mathcal{S} = \emptyset$, i.e., using no actuators.

The lower bound is obtained in a similar fashion by using all possible actuators. Explicitly,

$$\begin{aligned} \mathbf{P}_k(\mathcal{S}) &= \mathbf{Q}_k + \mathbf{A}_k^T \left(\mathbf{P}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{S} \cap \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \mathbf{A}_k \\ &\succeq \mathbf{Q}_k + \mathbf{A}_k^T \left(\mathbf{P}_{k+1}^{-1}(\mathcal{S}) + \sum_{i \in \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \mathbf{A}_k \\ &= \mathbf{P}_k(\mathcal{S} \cup \mathcal{V}_k). \end{aligned}$$

Moreover, if $\underline{\mathbf{P}}_{k+1} \preceq \mathbf{P}_{k+1}(\mathcal{S})$ for all $\mathcal{S} \subseteq \bar{\mathcal{V}}$, we have that

$$\mathbf{P}_k(\mathcal{S}) \succeq \mathbf{Q}_k + \mathbf{A}_k^T \left(\underline{\mathbf{P}}_{k+1}^{-1} + \sum_{i \in \mathcal{V}_k} r_{i,k}^{-1} \mathbf{b}_{i,k} \mathbf{b}_{i,k}^T \right)^{-1} \mathbf{A}_k \triangleq \underline{\mathbf{P}}_k. \quad (\text{B.62})$$

Starting from $\mathbf{P}_N \succeq \mathbf{Q}_N \triangleq \underline{\mathbf{P}}_N$ yields the lower bound in (B.60). ■

B.3.7 Proof of Proposition 17

Start by lifting the problem to make the proof more concise by defining the stacked quantities $\tilde{\mathbf{y}}_{\mathcal{D}} = [\mathbf{y}_e]_{e \in \mathcal{D}}$, an $n \times 1$ vector, $\tilde{\mathbf{A}}_{\mathcal{D}} = [\mathbf{A}_e]_{e \in \mathcal{D}}$, an $n \times p$ matrix, $\tilde{\mathbf{v}}_{\mathcal{D}} = [\mathbf{v}_e]_{e \in \mathcal{D}}$, an $n \times 1$ vector, and $\tilde{\mathbf{R}}_{\mathcal{D}} = \text{blkdiag}(\mathbf{R}_e)_{e \in \mathcal{D}}$, an $n \times n$ block diagonal matrix, with $n = \sum_{j \in \mathcal{D}} n_j$. Since the design is fixed, the dependence on \mathcal{D} is omitted throughout the proof for clarity.

Note that all affine estimators of \mathbf{z} can be written as $\hat{\mathbf{z}} = \mathbf{L}\tilde{\mathbf{y}} + \mathbf{b}$, so that the problem reduces to determining the optimal \mathbf{L}^* and \mathbf{b}^* . Using the model in (7.66), write $\mathbf{L}\tilde{\mathbf{y}} = \mathbf{L}(\tilde{\mathbf{A}}\boldsymbol{\theta} + \tilde{\mathbf{v}})$, so that

the error covariance matrix of the estimator has the form

$$\begin{aligned}
K &= \mathbb{E} \left[(\mathbf{H}\boldsymbol{\theta} - \mathbf{L}\tilde{\mathbf{A}}\boldsymbol{\theta} - \mathbf{L}\tilde{\mathbf{v}} - \mathbf{b})(\mathbf{H}\boldsymbol{\theta} - \mathbf{L}\tilde{\mathbf{A}}\boldsymbol{\theta} - \mathbf{L}\tilde{\mathbf{v}} - \mathbf{b})^T \mid \boldsymbol{\theta}, \tilde{\mathbf{R}} \right] \\
&= (\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})\mathbf{R}_\theta(\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})^T + \mathbf{L}\tilde{\mathbf{R}}\mathbf{L}^T \\
&\quad + \left[(\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})\bar{\boldsymbol{\theta}} - \mathbf{b} \right] \left[(\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})\bar{\boldsymbol{\theta}} - \mathbf{b} \right]^T,
\end{aligned}$$

where all terms linear in $\bar{\mathbf{v}}$ vanish since $\{\mathbf{v}_e, \boldsymbol{\theta}\}$ are independent for all $e \in \mathcal{E}$. It is ready that the last term is minimized by taking

$$\mathbf{b}^* = (\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})\bar{\boldsymbol{\theta}}.$$

Suffices now to minimize the sum of the first two terms.

To do so, note that for $\mathbf{b} = \mathbf{b}^*$, we have $\mathbf{K} = (\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})\mathbf{R}_\theta(\mathbf{H} - \mathbf{L}\tilde{\mathbf{A}})^T$. Therefore, taking $\mathbf{L} = \mathbf{L}^* + (\mathbf{L} - \mathbf{L}^*)$ with

$$\mathbf{L}^* = \mathbf{H} \left(\mathbf{R}_\theta^{-1} + \tilde{\mathbf{A}}^T \tilde{\mathbf{R}}^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}^T \tilde{\mathbf{R}}^{-1},$$

and expanding gives

$$\begin{aligned}
\mathbf{K} &= \left(\mathbf{H} - \mathbf{L}^* \tilde{\mathbf{A}} \right) \mathbf{R}_\theta \left(\mathbf{H} - \mathbf{L}^* \tilde{\mathbf{A}} \right)^T + \mathbf{L}^* \tilde{\mathbf{R}} \mathbf{L}^{*T} \\
&\quad + (\mathbf{L} - \mathbf{L}^*) \left(\tilde{\mathbf{A}} \mathbf{R}_\theta \tilde{\mathbf{A}}^T + \tilde{\mathbf{R}} \right) (\mathbf{L} - \mathbf{L}^*)^T \\
&\quad + (\mathbf{L} - \mathbf{L}^*) \left[\mathbf{L}^* \tilde{\mathbf{R}} - \left(\mathbf{H} - \mathbf{L}^* \tilde{\mathbf{A}} \right) \mathbf{R}_\theta \tilde{\mathbf{A}}^T \right]^T \\
&\quad + \left[\mathbf{L}^* \tilde{\mathbf{R}} - \left(\mathbf{H} - \mathbf{L}^* \tilde{\mathbf{A}} \right) \mathbf{R}_\theta \tilde{\mathbf{A}}^T \right] (\mathbf{L} - \mathbf{L}^*)^T \\
&= \mathbf{K}^* + (\mathbf{L} - \mathbf{L}^*) \left(\tilde{\mathbf{A}} \mathbf{R}_\theta \tilde{\mathbf{A}}^T + \tilde{\mathbf{R}} \right) (\mathbf{L} - \mathbf{L}^*)^T,
\end{aligned} \tag{B.63}$$

where the two last terms vanish and $\mathbf{K}^* = (\mathbf{H} - \mathbf{L}^* \tilde{\mathbf{A}})\mathbf{R}_\theta(\mathbf{H} - \mathbf{L}^* \tilde{\mathbf{A}})^T + \mathbf{L}^* \tilde{\mathbf{R}} \mathbf{L}^{*T}$. Clearly, the minimum value of (B.63) is \mathbf{K}^* , attained for $\mathbf{L} = \mathbf{L}^*$. Finally, $\hat{\mathbf{z}} = \mathbf{L}^* \tilde{\mathbf{y}} + \mathbf{b}^*$ and \mathbf{K}^* can be unstacked and rearranged to yield (7.68) and (7.69). ■

B.3.8 Proof of Theorem 15

This proof relies on the fact that α depends only on rank-one updates of the covariance matrix. Therefore, we can use the matrix inversion lemma to obtain a closed-form expression for the increments required to evaluate (7.6). Spectral inequalities are then used to bound the increments ratio.

Explicitly, start by expressing the error covariance matrix from (7.69) as $\mathbf{K}(\mathcal{D}) = \mathbf{H}\mathbf{Y}(\mathcal{D})^{-1}\mathbf{H}^T$, with $\mathbf{Y}(\mathcal{D}) = \mathbf{R}_\theta^{-1} + \sum_{e \in \mathcal{D}} \mathbf{M}_e$ and define $g(\mathcal{D}) = \text{Tr}[\mathbf{K}(\mathcal{D})]$. Then, notice that α only depends on the incremental gains $\Delta_u(\mathcal{X}) = g(\mathcal{X}) - g(\mathcal{X} \cup \{u\})$. Indeed, we have from Definition 8 that

$$\alpha(a, b) = \min_{\substack{\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E}) \\ \mathcal{A} \subseteq \mathcal{B}, u \in \mathcal{E} \\ |\mathcal{A}|=a, |\mathcal{B}|=b}} \frac{\Delta_u(\mathcal{A})}{\Delta_u(\mathcal{B})}. \quad (\text{B.64})$$

Using the additivity of \mathbf{Y} gives $g(\mathcal{X} \cup \{u\}) = \text{Tr} \left[\mathbf{H} (\mathbf{Y}(\mathcal{X}) + \mathbf{M}_u)^{-1} \mathbf{H}^T \right]$, which suggests that the matrix inversion lemma can be used to obtain a simpler expression for Δ . However, although $\mathbf{Y}(\mathcal{X}) \succ 0$ due to $\mathbf{R}_\theta \succ 0$, the matrices \mathbf{M}_u need not be invertible. Thus, we use the inversion lemma version from [235] to get

$$g(\mathcal{X} \cup \{u\}) = \text{Tr} \left[\mathbf{H}\mathbf{Y}(\mathcal{X})^{-1}\mathbf{H}^T - \mathbf{H}\mathbf{Y}(\mathcal{X})^{-1}\mathbf{M}_u [\mathbf{Y}(\mathcal{X}) + \mathbf{M}_u]^{-1} \mathbf{H}^T \right].$$

Finally, the linearity of the trace operator implies

$$\Delta_u(\mathcal{X}) = \text{Tr} \left[\mathbf{H}\mathbf{Y}(\mathcal{X})^{-1}\mathbf{M}_u [\mathbf{Y}(\mathcal{X}) + \mathbf{M}_u]^{-1} \mathbf{H}^T \right]. \quad (\text{B.65})$$

Our goal is now to explicitly lower bound (B.64) by exploiting the expression in (B.65) and spectral bounds. We do so by using the following result:

Lemma 10. *For all $\mathcal{X} \subseteq \mathcal{P}(\mathcal{E})$ and $u \in \mathcal{E}$, it holds that for Δ as in (B.65)*

$$\begin{aligned} \lambda_{\min} [\mathbf{H}\mathbf{H}^T] \lambda_{\min} [\mathbf{Y}(\mathcal{X})^{-1}] \text{Tr} \left[\mathbf{M}_u [\mathbf{Y}(\mathcal{X}) + \mathbf{M}_u]^{-1} \right] &\leq \Delta_u(\mathcal{X}) \leq \\ &\leq \lambda_{\max} [\mathbf{H}\mathbf{H}^T] \lambda_{\max} [\mathbf{Y}(\mathcal{X})^{-1}] \text{Tr} \left[\mathbf{M}_u [\mathbf{Y}(\mathcal{X}) + \mathbf{M}_u]^{-1} \right]. \end{aligned} \quad (\text{B.66})$$

Before proving Lemma 10, let us see how it leads to the desired result. Using (B.66) we can bound (B.64) as in

$$\alpha(a, b) \geq \min_{\substack{\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E}) \\ \mathcal{A} \subseteq \mathcal{B}, u \in \mathcal{E} \\ |\mathcal{A}|=a, |\mathcal{B}|=b}} \frac{\lambda_{\min}(\mathbf{H}\mathbf{H}^T) \lambda_{\min}[\mathbf{Y}(\mathcal{A})^{-1}] \text{Tr}[\mathbf{M}_u[\mathbf{Y}(\mathcal{A}) + \mathbf{M}_u]^{-1}]}{\lambda_{\max}(\mathbf{H}\mathbf{H}^T) \lambda_{\max}[\mathbf{Y}(\mathcal{B})^{-1}] \text{Tr}[\mathbf{M}_u[\mathbf{Y}(\mathcal{B}) + \mathbf{M}_u]^{-1}]}.$$

Then, let $\kappa(\mathbf{X}) = \sigma_{\max}(\mathbf{X})/\sigma_{\min}(\mathbf{X})$ be the ℓ_2 -norm condition number with respect to inversion, where $\{\sigma_t(\mathbf{X})\}$ are the singular values of \mathbf{X} . Using the fact that $\lambda_t(\mathbf{H}^T \mathbf{H}) = \sigma_t^2(\mathbf{H})$ yields

$$\alpha(a, b) \geq \kappa(\mathbf{H})^{-2} \min_{\substack{\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E}) \\ \mathcal{A} \subseteq \mathcal{B}, u \in \mathcal{E} \\ |\mathcal{A}|=a, |\mathcal{B}|=b}} \frac{\lambda_{\min}[\mathbf{Y}(\mathcal{B})]}{\lambda_{\max}[\mathbf{Y}(\mathcal{A})]} \times \frac{\text{Tr}[\mathbf{M}_u[\mathbf{Y}(\mathcal{A}) + \mathbf{M}_u]^{-1}]}{\text{Tr}[\mathbf{M}_u[\mathbf{Y}(\mathcal{B}) + \mathbf{M}_u]^{-1}]} \quad (\text{B.67})$$

To proceed, recall from Lemma (4) that \mathbf{Y}^{-1} is a decreasing set function, so that $\mathcal{A} \subseteq \mathcal{B} \Rightarrow [\mathbf{Y}(\mathcal{A}) + \mathbf{M}_u]^{-1} \succeq [\mathbf{Y}(\mathcal{B}) + \mathbf{M}_u]^{-1}$. Since $\mathbf{M}_u \succeq 0$, the last term in (B.67) is lower bounded by one, giving

$$\alpha(a, b) \geq \kappa(\mathbf{H})^{-2} \min_{\substack{\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E}) \\ \mathcal{A} \subseteq \mathcal{B} \\ |\mathcal{A}|=a, |\mathcal{B}|=b}} \frac{\lambda_{\min}[\mathbf{Y}(\mathcal{B})]}{\lambda_{\max}[\mathbf{Y}(\mathcal{A})]}, \quad (\text{B.68})$$

which no longer depends on u , i.e., on which experiment is added to the design. We now remove the constraint $\mathcal{A} \subseteq \mathcal{B}$, which increases the feasible set and therefore reduces the value of the right-hand side of (B.68). By also using the fact that $\lambda_{\min}[\mathbf{Y}(\mathcal{X})] \geq \lambda_{\min}[\mathbf{Y}(\emptyset)] = \lambda_{\min}[\mathbf{R}_\theta^{-1}]$ for every $\mathcal{X} \in \mathcal{P}(\mathcal{E})$, we can eliminate the dependence on \mathcal{B} obtaining

$$\alpha(a, b) \geq \kappa(\mathbf{H})^{-2} \min_{\substack{|\mathcal{A}|=a \\ |\mathcal{B}|=b}} \frac{\lambda_{\min}[\mathbf{Y}(\mathcal{B})]}{\lambda_{\max}[\mathbf{Y}(\mathcal{A})]} \geq \frac{\kappa(\mathbf{H})^{-2} \lambda_{\min}[\mathbf{R}_\theta^{-1}]}{\max_{|\mathcal{A}|=a} \lambda_{\max}[\mathbf{Y}(\mathcal{A})]}. \quad (\text{B.69})$$

Finally, the lower bound in (7.70) is obtained using Weyl's inequality to get $\lambda_{\max}[\mathbf{Y}(\mathcal{A})] \leq \lambda_{\max}[\mathbf{R}_\theta^{-1}] + \sum_{e \in \mathcal{A}} \lambda_{\max}[\mathbf{M}_e]$ and letting $\sum_{e \in \mathcal{A}} \lambda_{\max}[\mathbf{M}_e] \leq a\ell_{\max}$.

Proof of Lemma 10. Start by defining the perturbed gain as

$$\Delta_\epsilon = \text{Tr}[\mathbf{H}\mathbf{Y}(\mathcal{X})^{-1} \bar{\mathbf{M}}_u (\mathbf{Y}(\mathcal{X}) + \bar{\mathbf{M}}_u)^{-1} \mathbf{H}^T],$$

for $\epsilon > 0$, where $\bar{\mathbf{M}}_u = \mathbf{M}_u + \epsilon \mathbf{I} \succ 0$. We omit the dependence on \mathcal{X} and u for clarity. Note that, $\Delta_\epsilon \rightarrow \Delta$ as $\epsilon \rightarrow 0$. Using the circular commutation property of the trace and the invertibility of $\mathbf{Y}(\mathcal{X})$ and $\bar{\mathbf{M}}_u$, we obtain

$$\Delta_\epsilon = \text{Tr}[(\mathbf{H}^T \mathbf{H}) \mathbf{Z}], \quad (\text{B.70})$$

where $\mathbf{Z} = \mathbf{Y}(\mathcal{X})^{-1} \left(\mathbf{Y}(\mathcal{X})^{-1} + \bar{\mathbf{M}}_i^{-1} \right)^{-1} \mathbf{Y}(\mathcal{X})^{-1}$. Notice that (B.70) is a product of two PSD matrices, so that we can use the bound from [236] to obtain

$$\lambda_{\min}(\mathbf{H}^T \mathbf{H}) \text{Tr}(\mathbf{Z}) \leq \Delta_\epsilon \leq \lambda_{\max}(\mathbf{H}^T \mathbf{H}) \text{Tr}(\mathbf{Z}). \quad (\text{B.71})$$

Let us proceed by bounding $\text{Tr}(\mathbf{Z})$. To do so, notice that $\mathbf{Y}(\mathcal{A})^{-1} \succ 0$, its square-root $\mathbf{Y}(\mathcal{A})^{-1/2}$ is well-defined and unique [178]. We can therefore use the circular commutation property of the trace to get

$$\text{Tr}(\mathbf{Z}) = \text{Tr} \left\{ \mathbf{Y}(\mathcal{X})^{-1} \left[\mathbf{Y}(\mathcal{X})^{-1/2} \left(\mathbf{Y}(\mathcal{X})^{-1} + \bar{\mathbf{M}}_i^{-1} \right)^{-1} \mathbf{Y}(\mathcal{X})^{-1/2} \right] \right\}. \quad (\text{B.72})$$

Since (B.72) depends again the product of PSD matrices, we can reapply the spectral bound from [236] and obtain

$$\begin{aligned} \lambda_{\min}[\mathbf{Y}(\mathcal{X})^{-1}] \text{Tr} \left[\mathbf{Y}(\mathcal{X})^{-1} \left(\mathbf{Y}(\mathcal{X})^{-1} + \bar{\mathbf{M}}_i^{-1} \right)^{-1} \right] &\leq \text{Tr}(\mathbf{Z}) \leq \\ &\leq \lambda_{\max}[\mathbf{Y}(\mathcal{X})^{-1}] \text{Tr} \left[\mathbf{Y}(\mathcal{X})^{-1} \left(\mathbf{Y}(\mathcal{X})^{-1} + \bar{\mathbf{M}}_i^{-1} \right)^{-1} \right]. \end{aligned} \quad (\text{B.73})$$

Reversing the manipulations used to get to (B.70) and combining (B.71) and (B.73) finally yields

$$\begin{aligned} \lambda_{\min}(\mathbf{H}^T \mathbf{H}) \lambda_{\min}[\mathbf{Y}(\mathcal{X})^{-1}] \text{Tr} \left[\bar{\mathbf{M}}_u [\mathbf{Y}(\mathcal{A}) + \bar{\mathbf{M}}_u]^{-1} \right] &\leq \Delta_\epsilon(\mathcal{X}) \leq \\ &\leq \lambda_{\max}(\mathbf{H}^T \mathbf{H}) \lambda_{\max}[\mathbf{Y}(\mathcal{X})^{-1}] \text{Tr} \left[\bar{\mathbf{M}}_u [\mathbf{Y}(\mathcal{X}) + \bar{\mathbf{M}}_u]^{-1} \right]. \end{aligned} \quad (\text{B.74})$$

The inequalities in (B.66) are obtained from (B.74) by continuity as $\epsilon \rightarrow 0$. ■

B.3.9 Proof of Theorem 16

This proof follows from a homotopy argument, i.e., we define a continuous map between the increments at \mathcal{A} and \mathcal{B} in Definition 9 and bound its derivative. The inequality in (7.71) follows from applying bounds on the spectrum of Hermitian matrices.

Let $\Delta_u(\mathcal{X}) = \lambda_{\max}[\mathbf{K}(\mathcal{X})] - \lambda_{\max}[\mathbf{K}(\mathcal{X} \cup \{u\})]$ be the gain of adding u to \mathcal{X} . Then, for $\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E})$, $\mathcal{A} \subseteq \mathcal{B}$, define the homotopy

$$h_{\mathcal{AB}}(t) = \lambda_{\max}[\mathbf{H}\mathbf{Z}(t)^{-1}\mathbf{H}^T] - \lambda_{\max}[\mathbf{H}(\mathbf{Z}(t) + \mathbf{M}_u)^{-1}\mathbf{H}^T] \quad (\text{B.75})$$

with $t \in [0, 1]$ and $\mathbf{Z}(t) = \mathbf{Y}(\mathcal{A}) + t[\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})]$. Note that $h_{\mathcal{AB}}(0) = \Delta_u(\mathcal{A})$ and $h_{\mathcal{AB}}(1) = \Delta_u(\mathcal{B})$. Thus, if $\dot{h}(t)$ is the derivative of h with respect to t , it is ready that $\Delta_u(\mathcal{B}) = \Delta_u(\mathcal{A}) + \int_0^1 \dot{h}_{\mathcal{AB}}(t)dt$. Using the definition of ϵ in (7.11) then yields

$$\epsilon(a, b) = \max_{\substack{\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E}) \\ \mathcal{A} \subseteq \mathcal{B}, u \in \mathcal{E} \\ |\mathcal{A}|=a, |\mathcal{B}|=b}} \int_0^1 \dot{h}_{\mathcal{AB}}(t)dt. \quad (\text{B.76})$$

In the sequel, we proceed by upper bounding \dot{h} , thus getting the bound in (7.71). We omit the dependence on \mathcal{A} and \mathcal{B} for conciseness. First, to find the derivative of (B.75), recall from matrix analysis that $\frac{d}{dt}\mathbf{X}(t)^{-1} = -\mathbf{X}(t)^{-1}\dot{\mathbf{X}}(t)\mathbf{X}(t)^{-1}$ and $\frac{d}{dt}\lambda_{\max}[\mathbf{X}(t)] = \mathbf{u}(t)^T \frac{d}{dt}\dot{\mathbf{X}}(t)\mathbf{u}(t)$, with $\mathbf{u}(t)$ the eigenvector relative to the maximum eigenvalue of $\mathbf{X}(t)$ [179]. Then,

$$\begin{aligned} \frac{d}{dt}\lambda_{\max}[\mathbf{H}\mathbf{Z}(t)^{-1}\mathbf{H}^T] &= \mathbf{u}(t)^T \left[\frac{d}{dt}\mathbf{H}\mathbf{Z}(t)^{-1}\mathbf{H}^T \right] \mathbf{u}(t) \\ &= -\tilde{\mathbf{u}}(t)^T \mathbf{Z}(t)^{-1}\dot{\mathbf{Z}}(t)\mathbf{Z}(t)^{-1}\mathbf{Z}^T\tilde{\mathbf{u}}(t) \end{aligned}$$

where $\tilde{\mathbf{u}}(t) = \mathbf{H}^T\mathbf{u}(t)$ and $\mathbf{u}(t)$ is the eigenvector for the maximum eigenvalue of $\mathbf{H}\mathbf{X}(t)^{-1}\mathbf{H}^T$. Thus,

$$\begin{aligned} \dot{h}(t) &= \tilde{\mathbf{w}}(t)^T [\mathbf{Z}(t) + \mathbf{M}_u]^{-1} [\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})] [\mathbf{Z}(t) + \mathbf{M}_u]^{-1} \tilde{\mathbf{w}}(t) \\ &\quad - \tilde{\mathbf{v}}(t)^T \mathbf{Z}(t)^{-1} [\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})] \mathbf{Z}(t)^{-1} \tilde{\mathbf{v}}(t), \end{aligned} \quad (\text{B.77})$$

where $\tilde{\mathbf{v}}(t) = \mathbf{H}^T\mathbf{v}(t)$, $\tilde{\mathbf{w}}(t) = \mathbf{H}^T\mathbf{w}(t)$, and $\mathbf{v}(t)$ and $\mathbf{w}(t)$ are the eigenvectors relative to the max-

imum eigenvalues of $\mathbf{H}\mathbf{Z}(t)^{-1}\mathbf{H}^T$ and $\mathbf{H}[\mathbf{Z}(t) + \mathbf{M}_u]^{-1}\mathbf{H}^T$ respectively. To upper bound (B.77), start by noticing that since $\mathbf{Y}(\mathcal{A}) \preceq \mathbf{Y}(\mathcal{B})$, the second term in (B.77) is negative. Then, using the Rayleigh's inequality yields

$$\dot{h}(t) \leq \lambda_{\max} \left[(\mathbf{Z}(t) + \mathbf{M}_u)^{-1} (\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})) (\mathbf{Z}(t) + \mathbf{M}_u)^{-1} \right] \|\tilde{\mathbf{w}}(t)\|_2^2. \quad (\text{B.78})$$

We now find a bound for (B.78) that does not depend on t , so that we can apply (B.76). First, given that $\mathbf{w}(t)$ is a unit-norm vector, $\|\tilde{\mathbf{w}}(t)\|_2^2 = \mathbf{w}(t)^T \mathbf{H} \mathbf{H}^T \mathbf{w}(t) \leq \lambda_{\max}(\mathbf{H} \mathbf{H}^T) = \sigma_{\max}^2(\mathbf{H})$, where $\sigma_{\max}(\mathbf{H})$ is the maximum singular value of \mathbf{H} . Then, note that $\mathbf{Z}(t) \preceq \mathbf{Z}(0) = \mathbf{Y}(\mathcal{A})$. Thus, using the fact that for $\mathbf{A}, \mathbf{B} \succeq 0$ it holds that $\lambda_{\max}(\mathbf{A} \mathbf{B} \mathbf{A}) = \sigma_{\max}^2(\mathbf{A} \mathbf{B}^{1/2}) \leq \sigma_{\max}^2(\mathbf{A}) \sigma_{\max}^2(\mathbf{B}^{1/2}) = \lambda_{\max}^2(\mathbf{A}) \lambda_{\max}(\mathbf{B})$ yields

$$\dot{h}(t) \leq \sigma_{\max}(\mathbf{H})^2 \lambda_{\min}[\mathbf{Y}(\mathcal{A}) + \mathbf{M}_u]^{-2} \lambda_{\max}[\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})],$$

Thus, from (B.76),

$$\epsilon(a, b) \leq \max_{\substack{\mathcal{A}, \mathcal{B} \in \mathcal{P}(\mathcal{E}) \\ \mathcal{A} \subseteq \mathcal{B}, u \in \mathcal{E} \\ |\mathcal{A}|=a, |\mathcal{B}|=b}} \frac{\sigma_{\max}(\mathbf{H})^2 \lambda_{\max}[\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})]}{\lambda_{\min}[\mathbf{Y}(\mathcal{A}) + \mathbf{M}_u]^2}, \quad (\text{B.79})$$

The inequality in (7.71) is obtained using $\lambda_{\max}[\mathbf{Y}(\mathcal{B}) - \mathbf{Y}(\mathcal{A})] \leq \sum_{e \in \mathcal{B} \setminus \mathcal{A}} \lambda_{\max}(\mathbf{M}_e) \leq (b - a) \ell_{\max}$ for $|\mathcal{A}| = a$ and $|\mathcal{B}| = b$ and $\lambda_{\min}[\mathbf{Y}(\mathcal{A}) + \mathbf{M}_u] \geq \lambda_{\min}[\mathbf{R}_{\theta}^{-1}]$. \blacksquare

B.4 Chapter 8: Risk constraints

B.4.1 Proof of Lemma 6

We start with the objective of problem (P-RISK), for which it is obviously true that

$$\mathbb{E}\{\|\mathbf{x} - \phi(\mathbf{y})\|^2\} \equiv \mathbb{E}\{\mathbb{E}\{\|\mathbf{x}\|_2^2 - 2\mathbf{x}^T \phi(\mathbf{y}) + \|\phi(\mathbf{y})\|_2^2 | \mathbf{y}\}\}, \quad (\text{B.80})$$

since the expectation of $\|\mathbf{x} - \phi(\mathbf{y})\|^2$ always exists. Additionally, by invoking Cauchy-Schwarz twice, we observe that

$$\begin{aligned}\mathbb{E}\{\|\mathbf{x}^T \phi(\mathbf{y})\|\} &\leq \mathbb{E}\{\|\mathbf{x}\|_2 \|\phi(\mathbf{y})\|_2\} \\ &\equiv \mathbb{E}\{\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\} \|\phi(\mathbf{y})\|_2\} \\ &\leq \|\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\}\|_{L_2} \|\|\phi(\mathbf{y})\|_2\|_{L_2},\end{aligned}\tag{B.81}$$

where $\|\|\phi(\mathbf{y})\|_2\|_{L_2} < \infty \iff \phi(\mathbf{y}) \in L_2$ by assumption, and Jensen implies that

$$\begin{aligned}\|\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\}\|_{L_2} &\leq \|\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\}\|_{L_3} \\ &\leq (\mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2^3 | \mathbf{y}\})^{2 \cdot 1/2}\})^{1/3} \\ &\leq (\mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2^3 | \mathbf{y}\})^2\})^{1/(2 \cdot 3)} \\ &\leq \|\mathbb{E}\{\|\mathbf{x}\|_2^3 | \mathbf{y}\}\|_{L_2}^{1/3} < \infty,\end{aligned}\tag{B.82}$$

as well. Then $\mathbb{E}\{\mathbf{x}^T \phi(\mathbf{y})\}$ is finite, and it follows that

$$\mathbb{E}\{\|\mathbf{x} - \phi(\mathbf{y})\|^2\} \equiv \mathbb{E}\{\|\phi(\mathbf{y})\|_2^2 - 2(\mathbb{E}[\mathbf{x} | \mathbf{y}])^T \phi(\mathbf{y}) + \mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\}\},\tag{B.83}$$

as in the objective of (PXIV).

The constraint of (P-RISK) may be equivalently reexpressed in a similar fashion, although the procedure is slightly more involved. Specifically, by definition of $\text{var}(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y})$, we may initially expand as

$$\begin{aligned}&(\|\mathbf{x} - \phi(\mathbf{y})\|^2 - \mathbb{E}\{\|\mathbf{x} - \phi(\mathbf{y})\|^2 | \mathbf{y}\})^2 \\ &\equiv (\|\mathbf{x}\|_2^2 - \mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\})^2 + 4\phi(\mathbf{y})^T (\mathbf{x} - \mathbb{E}[\mathbf{x} | \mathbf{y}]) (\mathbf{x} - \mathbb{E}[\mathbf{x} | \mathbf{y}])^T \phi(\mathbf{y}) \\ &\quad - 4\|\mathbf{x}\|_2^2 \mathbf{x}^T \phi(\mathbf{y}) + 4\|\mathbf{x}\|_2^2 (\mathbb{E}[\mathbf{x} | \mathbf{y}])^T \phi(\mathbf{y}) \\ &\quad + 4\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} \mathbf{x}^T \phi(\mathbf{y}) - 4\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} (\mathbb{E}[\mathbf{x} | \mathbf{y}])^T \phi(\mathbf{y}),\end{aligned}\tag{B.84}$$

where the first two terms of the right-hand side of (B.84) are nonnegative. Consequently, it suffices

to concentrate on the respective last four dot product terms.

Using the same argument as in (B.81), in order to show that all these four terms have finite expectations, it suffices to ensure that

$$\|\mathbb{E}\{\|\mathbf{x}\|_2^2 \mathbf{x} | \mathbf{y}\}\|_{L_2} \equiv \|\mathbb{E}\{\|\mathbf{x}\|_2^3 | \mathbf{y}\}\|_{L_2} < \infty, \quad (\text{B.85})$$

which is of course automatically true by Assumption 7, but also that

$$\begin{aligned} \|\mathbb{E}\{\|\mathbf{x}\|_2^2 \mathbb{E}[\mathbf{x} | \mathbf{y}]\|_{L_2}\|_{L_2} &\equiv \|\mathbb{E}\{\|\mathbf{x}\|_2^2 \mathbb{E}[\mathbf{x} | \mathbf{y}]\|_{L_2}\|_{L_2} \\ &\equiv \|\mathbb{E}[\mathbf{x} | \mathbf{y}]\|_{L_2} \|\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\}\|_{L_2} \\ &< \infty, \end{aligned} \quad (\text{B.86})$$

$$\begin{aligned} \|\mathbb{E}\{\|\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} \mathbf{x}\|_{L_2}\|_{L_2} &\equiv \|\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\} \mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\}\|_{L_2} \\ &< \infty \quad \text{and} \end{aligned} \quad (\text{B.87})$$

$$\|\mathbb{E}\{\|\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} \mathbb{E}[\mathbf{x} | \mathbf{y}]\|_{L_2}\|_{L_2} \equiv \|\mathbb{E}[\mathbf{x} | \mathbf{y}]\|_{L_2} \|\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\}\|_{L_2} \quad (\text{B.88})$$

$$< \infty. \quad (\text{B.89})$$

Observe, though, that all three latter quantities are upper bounded by the quantity $\|\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\} \mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\}\|_{L_2}$, for which we may write (by Jensen)

$$\begin{aligned} \|\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\} \mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\}\|_{L_2}^2 &\equiv \mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2 | \mathbf{y}\})^2 (\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\})^2\} \\ &\leq \mathbb{E}\{\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} (\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\})^2\} \\ &\equiv \mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\})^3\} \\ &\equiv \mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\})^{2 \cdot 3/2}\} \\ &\leq \mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2^3 | \mathbf{y}\})^2\} < \infty, \end{aligned} \quad (\text{B.90})$$

where the last line follows again by Assumption 7.

Given the discussion above, we may now take conditional expectations on (B.84), to obtain the expression (note that all operations involving conditional expectations are technically allowed under

our assumptions)

$$\begin{aligned}
& \text{var} \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \right) \\
& \equiv \mathbb{E} \left\{ \left(\|\mathbf{x} - \phi(\mathbf{y})\|^2 - \mathbb{E} \{ \|\mathbf{x} - \phi(\mathbf{y})\|^2 \mid \mathbf{y} \} \right)^2 \mid \mathbf{y} \right\} \\
& \equiv \mathbb{V}_{\mathbf{y}} \{ \|\mathbf{x}\|_2^2 \} + 4\phi(\mathbf{y})^T \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \phi(\mathbf{y}) \\
& \quad - 4 \left(\mathbb{E} \{ \|\mathbf{x}\|_2^2 \mathbf{x} \mid \mathbf{y} \} - \mathbb{E} \{ \|\mathbf{x}\|_2^2 \mid \mathbf{y} \} \mathbb{E} [\mathbf{x} \mid \mathbf{y}] \right)^T \phi(\mathbf{y}).
\end{aligned} \tag{B.91}$$

Taking expectations on both sides of (B.91) and rearranging terms gives the desired expression for the constraint of the QCQP (PXIV). ■

B.4.2 Proof of Theorem 18

Let us first evaluate the dual function d in (8.5). Define the extended real-valued, random function as

$$\begin{aligned}
r(\mathbf{x}, \mathbf{y}) & \triangleq \frac{1}{2} \mathbf{x}^T (\mathbf{I} + 2\mu \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \mathbf{x} \\
& \quad - \left(\mathbb{E} [\mathbf{x} \mid \mathbf{y}] + \mu \left(\mathbb{E} \{ \|\mathbf{x}\|_2^2 \mathbf{x} \mid \mathbf{y} \} - \mathbb{E} \{ \|\mathbf{x}\|_2^2 \mid \mathbf{y} \} \mathbb{E} [\mathbf{x} \mid \mathbf{y}] \right) \right)^T \mathbf{x}.
\end{aligned} \tag{B.92}$$

Observe that the quadratic term $\mathbf{x}^T (\mathbf{I} + 2\mu \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \mathbf{x}$ is finite $\mathbf{p}_{\mathbf{y}}$ -almost everywhere; indeed, for every $\mathbf{x} \in \mathbb{R}^d$, it is true that

$$\begin{aligned}
0 & \leq \mathbf{x}^T \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \mathbf{x} \leq \|\mathbf{x}\|_2^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \\
& \leq \|\mathbf{x}\|_2^2 \text{tr}(\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \\
& \equiv \|\mathbf{x}\|_2^2 \mathbb{E} \{ \|\mathbf{x} - \mathbb{E} [\mathbf{x} \mid \mathbf{y}]\|_2^2 \mid \mathbf{y} \} \\
& \equiv \|\mathbf{x}\|_2^2 \left(\mathbb{E} \{ \|\mathbf{x}\|_2^2 \mid \mathbf{y} \} - \|\mathbb{E} [\mathbf{x} \mid \mathbf{y}]\|_2^2 \right) \\
& \leq \|\mathbf{x}\|_2^2 \mathbb{E} \{ \|\mathbf{x}\|_2^2 \mid \mathbf{y} \},
\end{aligned} \tag{B.93}$$

where

$$\begin{aligned}
0 \leq (\mathbb{E}\{\|\mathbf{x}\|_2^2\})^3 &\leq (\mathbb{E}\{\|\mathbf{x}\|_2^3\})^2 \\
&\leq \mathbb{E}\{(\mathbb{E}\{\|\mathbf{x}\|_2^3|\mathbf{y}\})^2\} \\
< \infty &\implies \mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\} < \infty, \quad \mathbf{p}_{\mathbf{y}} - a.e.
\end{aligned} \tag{B.94}$$

Therefore, due to our assumptions, the function $r(\cdot, \mathbf{y})$ is trivially continuous and finite on \mathbb{R}^d up to sets of $\mathbf{p}_{\mathbf{y}}$ -measure zero, those being independent of each choice of $\mathbf{x} \in \mathbb{R}^d$, on which $r(\cdot, \mathbf{y})$ may be arbitrarily defined. Consequently, $r(\cdot, \mathbf{y})$ has a real-valued version, and thus may be taken as Carathéodory on $\mathbb{R}^d \times \Omega$ [227, p. 421]. Equivalently, r may also be taken as Carathéodory on $\mathbb{R}^d \times \mathbb{R}^k$, jointly measurable relative to the Borel σ -algebra $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^k)$.

Under the above considerations, and given that Assumptions 7 and 8 are in effect, we may drop additive terms which do not depend on the decision $\phi(\mathbf{y})$ in (8.5), resulting in the equivalent problem

$$d = \min_{\phi \in \mathcal{H}} \mathbb{E}[r(\phi(\mathbf{y}), \mathbf{y})] \tag{PXX}$$

where expectation may be conveniently taken directly over the Borel probability space $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), \mathbf{p}_{\mathbf{y}})$. Note that (PXX) is uniformly lower bounded over L_2 , through the definition of the Lagrangian L , and also that, trivially, there is at least one choice of $\phi(\mathbf{y}) \in L_2$ such that $\mathbb{E}\{r(\phi(\mathbf{y}), \mathbf{y})\} < \infty$, say $\mathbb{E}\{r(\mathbf{0}, \mathbf{y})\} \equiv 0$, for $\phi \equiv \mathbf{0}$.

Problem (PXX) may now be solved in closed form via application of the *Interchangeability Principle* [227, Theorem 7.92], which is a fundamental result in variational optimization. To avoid unnecessary generalities, we state it here for completeness adapted to our setting, as follows.

Theorem 22 (Interchangeability Principle [227]). *Let $f : \mathbb{R}^d \times \mathbb{R}^k \rightarrow \mathbb{R}$ be Carathéodory, and fix $p \in [1, \infty]$. It is true that*

$$\inf_{\phi \in L_p} \mathbb{E}\{f(\phi(\mathbf{y}), \mathbf{y})\} \equiv \mathbb{E}\{\inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}, \mathbf{y})\}, \tag{B.95}$$

provided that the left-hand side of (B.95) is less than $+\infty$. If, additionally, either of the sides of

(B.95) is not $-\infty$, it is also true that

$$\phi^*(\mathbf{y}) \in \arg \min_{\phi \in L_p} \mathbb{E}\{f(\phi(\mathbf{y}), \mathbf{y})\} \quad (\text{B.96})$$

$$\iff \phi^*(\mathbf{y}) \in \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}, \mathbf{y}), \text{ for } \mathbf{p}_{\mathbf{y}}\text{-almost all } \mathbf{y}, \text{ and } \phi^* \in L_p. \quad (\text{B.97})$$

Let us apply Theorem 22 to the variational problem (PXX), for $p \equiv 2$. Then, the (PXX) may be exchanged by the pointwise (over constants) quadratic problem

$$\inf_{\mathbf{x} \in \mathbb{R}^d} r(\mathbf{x}, \mathbf{y}), \quad (\text{B.98})$$

whose unique solution is, for every $\mu \in \mathbb{R}_+$ and for every value of $\mathbf{y} \in \mathbb{R}^k$,

$$\phi^*(\mathbf{y}) = (\mathbf{I} + 2\mu\mathbf{\Sigma}_{\mathbf{x}|\mathbf{y}})^{-1}(\mathbb{E}[\mathbf{x}|\mathbf{y}] + \mu(\mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\} - \mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\}\mathbb{E}[\mathbf{x}|\mathbf{y}])), \quad (\text{B.99})$$

which is precisely the expression claimed in Theorem 18, for a generic μ . In order to show that (B.99) is a solution of problem (PXX) and, in turn, (8.5), we also have to verify that $\phi^* \in L_2$. We may write, by Cauchy-Schwarz (note that $\|(\mathbf{I} + 2\mu\mathbf{\Sigma}_{\mathbf{x}|\mathbf{y}})^{-1}\|_2 \leq 1$), the triangle inequality, and Jensen,

$$\begin{aligned} \|\phi^*(\mathbf{y})(\mu)\|_2 &\leq \|\mathbb{E}[\mathbf{x}|\mathbf{y}] + \mu(\mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\} - \mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\}\mathbb{E}[\mathbf{x}|\mathbf{y}])\|_2 \\ &\leq \|\mathbb{E}[\mathbf{x}|\mathbf{y}]\|_2 + \mu\|\mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\} - \mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\}\mathbb{E}[\mathbf{x}|\mathbf{y}]\|_2 \\ &\leq \mathbb{E}\{\|\mathbf{x}\|_2|\mathbf{y}\} + \mu\mathbb{E}\{\|\mathbf{x}\|_2^3|\mathbf{y}\} + \mu\mathbb{E}\{\|\mathbf{x}\|_2^2|\mathbf{y}\}\mathbb{E}\{\|\mathbf{x}\|_2|\mathbf{y}\}, \end{aligned} \quad (\text{B.100})$$

and we are done, since we have already shown that all three terms in the right-hand side of (B.100) are in L_2 .

The final step in the proof of Theorem 18 is to exploit strong duality of the QCQP (PXIV) by invoking Theorem 17. Indeed, it follows that the optimal value of the primal problem (PXIV)

coincides with that of the dual problem (PXV), which may be expressed as

$$\begin{aligned}
& \sup_{\mu \geq 0} d(\mu) \\
& \equiv \sup_{\mu \geq 0} \inf_{\phi \in \mathcal{H}} L(\phi(\mathbf{y}), \mu) \\
& \equiv \sup_{\mu \geq 0} \left\{ \frac{1}{2} \mathbb{E}\{\|\mathbf{x}\|_2^2\} + \frac{1}{4} \mu \mathbb{E}\{\mathbb{V}_{\mathbf{y}}\{\|\mathbf{x}\|_2^2\}\} \right. \\
& \quad + \mathbb{E}\left\{ \frac{1}{2} \|\phi^*(\mathbf{y})(\mu)\|_2^2 + \mu \phi^*(\mathbf{y})(\mu)^T \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \phi^*(\mathbf{y})(\mu) \right. \\
& \quad \left. - (\mathbb{E}[\mathbf{x}|\mathbf{y}])^T \phi^*(\mathbf{y})(\mu) - \mu (\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} - \mathbb{E}\{\|\mathbf{x}\|_2^2\} \mathbb{E}[\mathbf{x} | \mathbf{y}])^T \phi^*(\mathbf{y})(\mu) \right\} \\
& \quad \left. - \frac{\mu \varepsilon}{4} \right\} \\
& \equiv \frac{1}{2} \mathbb{E}\{\|\mathbf{x}\|_2^2\} + \sup_{\mu \geq 0} \left\{ \frac{1}{4} \mu \mathbb{E}\{\mathbb{V}_{\mathbf{y}}\{\|\mathbf{x}\|_2^2\}\} \right. \\
& \quad + \mathbb{E}\left\{ \frac{1}{2} \phi^*(\mathbf{y})(\mu)^T (\mathbf{I} + 2\mu \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \phi^*(\mathbf{y})(\mu) \right. \\
& \quad \left. - (\mathbb{E}[\mathbf{x}|\mathbf{y}] + \mu (\mathbb{E}\{\|\mathbf{x}\|_2^2 | \mathbf{y}\} - \mathbb{E}\{\|\mathbf{x}\|_2^2\} \mathbb{E}[\mathbf{x} | \mathbf{y}]))^T \phi^*(\mathbf{y})(\mu) \right\} - \frac{\mu \varepsilon}{4} \Big\} \\
& \equiv \frac{1}{2} \mathbb{E}\{\|\mathbf{x}\|_2^2\} + \sup_{\mu \geq 0} \left\{ \frac{1}{4} \mu \mathbb{E}\{\mathbb{V}_{\mathbf{y}}\{\|\mathbf{x}\|_2^2\}\} \right. \\
& \quad + \mathbb{E}\left\{ \frac{1}{2} \phi^*(\mathbf{y})(\mu)^T (\mathbf{I} + 2\mu \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \phi^*(\mathbf{y})(\mu) - \phi^*(\mathbf{y})(\mu)^T (\mathbf{I} + 2\mu \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \phi^*(\mathbf{y})(\mu) \right\} - \frac{\mu \varepsilon}{4} \Big\} \\
& \equiv \frac{1}{2} \mathbb{E}\{\|\mathbf{x}\|_2^2\} + \frac{1}{4} \sup_{\mu \geq 0} \left\{ \mu \mathbb{E}\{\mathbb{V}_{\mathbf{y}}\{\|\mathbf{x}\|_2^2\}\} \right. \\
& \quad \left. - 2\mathbb{E}\{\phi^*(\mathbf{y})^T(\mu)(\mathbf{I} + 2\mu \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}) \phi^*(\mathbf{y})(\mu)\} - \mu \varepsilon \right\}. \tag{B.101}
\end{aligned}$$

Finally, let $\mu^* \geq 0$ be a maximizer of d over \mathbb{R}_+ such that $D_{\text{RISK}}^* \equiv P_{\text{RISK}}^* < \infty$, and suppose that $\tilde{\mathbf{x}}_*$ is primal optimal for (PXIV). By strong duality, it is true that

$$\tilde{\mathbf{x}}_* \equiv \tilde{\mathbf{x}}_*(\mu^*) \in \arg \min_{\phi \in \mathcal{H}} L(\phi(\mathbf{y}), \mu^*). \tag{B.102}$$

By uniqueness of $\phi^*(\mathbf{y})$ in (B.99) (pointwise in \mathbf{y}), all members of the possibly infinite set of optimal solutions of (8.5) (for $\mu \equiv \mu^*$) in (B.102) differ at most on sets of measure zero, and result in exactly the same values for both the objective and constraints of (PXIV). Therefore, all such optimal solutions to (8.5) are also optimal for (PXIV) and, in particular, $\phi^*(\mathbf{y})$ is one of them. \blacksquare

Bibliography

- [1] A. Datta, M. C. Tschantz, and A. Datta, “Automated experiments on ad privacy settings,” *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.
- [2] M. Kay, C. Matuszek, and S. A. Munson, “Unequal representation and gender stereotypes in image search results for occupations,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 3819–3828. [Online]. Available: <https://doi.org/10.1145/2702123.2702520>
- [3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias,” *ProPublica*, 2016. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [4] P. Bright, “Tay, the neo-Nazi millennial chatbot, gets autopsied,” *Ars Technica*, 2016. [Online]. Available: <https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/>
- [5] National Transportation Safety Board, “HWY18MH010: Preliminary report,” National Transportation Safety Board, Tech. Rep., 2018. [Online]. Available: <https://www.nts.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf>
- [6] J. Destin, “Amazon scraps secret ai recruiting tool that showed bias against women,” *Reuters*, 2018. [Online]. Available: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

- [7] R. Ge, J. D. Lee, and T. Ma, “Learning one-hidden-layer neural networks with landscape design,” *arXiv preprint arXiv:1711.00501*, 2017.
- [8] H. Lin and S. Jegelka, “Resnet with one-neuron hidden layers is a universal approximator,” in *Advances in Neural Information Processing Systems*, 2018, pp. 6169–6178.
- [9] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, “Theoretical insights into the optimization landscape of over-parameterized shallow neural networks,” *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, 2018.
- [10] S. S. Du, X. Zhai, B. Póczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” *arXiv preprint arXiv:1810.02054*, 2018.
- [11] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *CoRR*, 2014.
- [12] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, “A convex framework for fair regression,” in *Fairness, Accountability, and Transparency in Machine Learning*, 2017.
- [13] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. Van den Broeck, “A semantic loss function for deep learning with symbolic knowledge,” in *International Conference on Machine Learning*, 2018.
- [14] S. Zhao, J. Song, and S. Ermon, “The information autoencoding family: A Lagrangian perspective on latent variable generative models,” in *Conference on Uncertainty in Artificial Intelligence*, 2018.
- [15] A. Sinha, H. Namkoong, and J. Duchi, “Certifying some distributional robustness with principled adversarial training,” in *International Conference on Learning Representations*, 2018.
- [16] J. Chen and L. Deng, “A primal-dual method for training recurrent neural networks constrained by the echo-state property,” in *International Conference on Learning Representations*, 2014.

- [17] S. N. Ravi, T. Dinh, V. S. Lokhande, and V. Singh, “Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence,” in *AAAI Conference on Artificial Intelligence*, 2019, pp. 4772–4779.
- [18] D. Bertsekas, *Convex Optimization Theory*. Athena Scientific, 2009.
- [19] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000.
- [20] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2004.
- [21] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.
- [22] T. Cohen and M. Welling, “Group equivariant convolutional networks,” in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 2016, pp. 2990–2999.
- [23] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, “Rotation equivariant vector field networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5058–5067.
- [24] J. F. Henriques and A. Vedaldi, “Warped convolutions: Efficient invariance to spatial transformations,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 1461–1469.
- [25] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3856–3866.

- [26] M. Weiler, F. A. Hamprecht, and M. Storath, “Learning steerable filters for rotation equivariant CNNs,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.
- [27] L. Ruiz, F. Gama, A. G. Marques, and A. Ribeiro, “Invariance-preserving localized activation functions for graph neural networks,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 127–141, 2020.
- [28] V. Silva, L. Chamon, and A. Ribeiro, “Model predictive selection: A receding horizon scheme for actuator selection,” in *American Control Conference*, 2019, pp. 347–353.
- [29] L. Chamon, S. Paternain, and A. Ribeiro, “Learning gaussian processes with bayesian posterior optimization,” in *Asilomar*, 2019.
- [30] M. Eisen, C. Zhang, L. F. Chamon, D. D. Lee, and A. Ribeiro, “Learning optimal resource allocations in wireless systems,” *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2775–2790, 2019.
- [31] S. Paternain, M. Calvo-Fullana, L. Chamon, and A. Ribeiro, “Safe policies for reinforcement learning via primal-dual methods,” *IEEE Trans. on Autom. Control (under review)*, 2019, <https://arxiv.org/abs/1911.09101>.
- [32] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, “Constrained reinforcement learning has zero duality gap,” in *Advances in Neural Information Processing Systems*, 2019, pp. 7553–7563.
- [33] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, “Learning safe policies via primal–dual methods,” in *IEEE Conference on Decision and Control*, 2019.
- [34] L. Chamon, S. Paternain, and A. Ribeiro, “Counterfactual programming for optimal control,” in *Learning for Dynamics & Control (L4DC)*, 2020.
- [35] L. Chamon, A. Amice, S. Paternain, and A. Ribeiro, “Resilient control: Compromising to adapt,” in *IEEE Control and Decision Conference (CDC)*, 2020, <https://arxiv.org/abs/2004.03726>.

- [36] L. Chamon and A. Ribeiro, “Probably approximately correct constrained learning,” *Advances in Neural Information Processing Systems (under review)*, 2020, <https://arxiv.org/abs/2006.05487>.
- [37] L. Chamon, A. Amice, and A. Ribeiro, “Approximately supermodular scheduling subject to matroid constraints,” *IEEE Trans. on Autom. Control (under review)*, 2020, <https://arxiv.org/abs/2003.08841>.
- [38] L. Chamon, Y. Eldar, and A. Ribeiro, “Functional nonlinear sparse models,” *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2449–2463, 2020.
- [39] M. Peifer, L. Chamon, S. Paternain, and A. Ribeiro, “Sparse multiresolution representations with adaptive kernels,” *IEEE Trans. on Signal Process.*, vol. 68[1], pp. 2031–2044, 2020.
- [40] L. Chamon, G. J. Pappas, and A. Ribeiro, “Approximate supermodularity of Kalman filter sensor selection,” *IEEE Trans. on Autom. Control.*, 2020.
- [41] L. F. O. Chamon and A. Ribeiro, “Greedy sampling of graph signals,” *IEEE Trans. Signal Process.*, vol. 66[1], pp. 34–47, 2018.
- [42] L. Chamon and A. Ribeiro, “Approximate supermodularity bounds for experimental design,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5403–5412.
- [43] A. Tsiamis, D. Kalogerias, L. Chamon, A. Ribeiro, and G. Pappas, “Risk-constrained linear-quadratic regulators,” in *IEEE Control and Decision Conference (CDC)*, 2020, <https://arxiv.org/abs/2004.04685>.
- [44] L. Chamon, A. Amice, and A. Ribeiro, “Matroid-constrained approximately supermodular optimization for near-optimal actuator scheduling,” in *IEEE Control and Decision Conference (CDC)*, 2019, pp. 3391–3398.
- [45] L. Chamon, G. Pappas, and A. Ribeiro, “The mean square error in Kalman filtering sensor selection is approximately supermodular,” in *Conf. on Decision and Contr.*, 2017, pp. 343–350.

- [46] L. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro, “The empirical duality gap of constrained statistical learning problems,” in *International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2613–2616.
- [47] D. Kalogerias, L. Chamon, G. J. Pappas, and A. Ribeiro, “Better safe than sorry: Risk-aware nonlinear Bayesian estimation,” in *IEEE International Conference in Acoustic, Speech, and Signal Processing (ICASSP)*, 2020.
- [48] L. Chamon, Y. C. Eldar, and A. Ribeiro, “Sparse recovery over nonlinear dictionaries,” in *ICASSP*, 2019.
- [49] M. Peifer, L. Chamon, S. Paternain, and A. Ribeiro, “Sparse learning of parsimonious reproducing kernel Hilbert space models,” in *IEEE International Conference in Acoustic, Speech, and Signal Processing (ICASSP)*, 2019, pp. 3292–3296.
- [50] M. Eisen, C. Zhang, L. Chamon, D. D. Lee, and A. Ribeiro, “Dual domain learning of optimal resource allocations in wireless systems,” in *IEEE International Conference in Acoustic, Speech, and Signal Processing (ICASSP)*, 2019, pp. 4729–4733.
- [51] M. Peifer, L. Chamon, S. Paternain, and A. Ribeiro, “Locally adaptive kernel estimation using sparse functional programming,” in *Asilomar Conference on Signals, Systems and Computers*, 2018, pp. 2022–2026.
- [52] M. Eisen, C. Zhang, L. Chamon, D. D. Lee, and A. Ribeiro, “Online deep learning in wireless communication systems,” in *Asilomar Conference on Signals, Systems and Computers*, 2018, pp. 1289–1293.
- [53] L. Chamon, Y. C. Eldar, and A. Ribeiro, “Strong duality of sparse functional optimization,” in *ICASSP*, 2018, pp. 4739–4743.
- [54] L. Chamon and A. Ribeiro, “Universal bounds for the sampling of graph signals,” in *Int. Conf. on Acoust., Speech and Signal Process.*, 2016.
- [55] —, “Near-optimality of greedy set selection in the sampling of graph signals,” in *Global Conf. on Signal and Inform. Process.*, 2016, pp. 1265–1269.

- [56] T. Kailath, A. Sayed, and B. Hassibi, *Linear estimation*. Prentice-Hall, 2000.
- [57] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander, “Satisfying real-world goals with dataset constraints,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2415–2423.
- [58] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, “A reductions approach to fair classification,” in *International Conference on Machine Learning*, 2018, pp. 60–69.
- [59] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, “Empirical risk minimization under fairness constraints,” in *Advances in Neural Information Processing Systems*, 2018, pp. 2791–2801.
- [60] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, “Preventing fairness gerrymandering: Auditing and learning for subgroup fairness,” in *International Conference on Machine Learning*, 2018, pp. 2564–2572.
- [61] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, “Fairness constraints: A flexible approach for fair classification,” *Journal of Machine Learning Research*, vol. 20, no. 75, pp. 1–42, 2019.
- [62] A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan, “Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals,” *Journal of Machine Learning Research*, vol. 20, no. 172, pp. 1–59, 2019.
- [63] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations*, 2018.
- [64] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *International Conference on Machine Learning*, 2019, pp. 7472–7482.
- [65] J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

- [66] J. Achiam, D. Held, A. Tamar, and P. Abbeel, “Constrained policy optimization,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 22–31.
- [67] G. Wahba, *Spline models for observational data*. Siam, 1990, vol. 59.
- [68] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- [69] Y. C. Eldar and G. Kutyniok, Eds., *Compressed Sensing: Theory and Applications*. Cambridge, 2012.
- [70] S. Mallat, “Group invariant scattering,” *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.
- [71] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.
- [72] M. Ehrgott, *Multicriteria Optimization*. Springer, 2005.
- [73] K. Miettinen, *Nonlinear Multiobjective Optimization*. Springer, 1998.
- [74] A. Messac, A. Ismail-Yahaya, and C. Mattson, “The normalized normal constraint method for generating the Pareto frontier,” *Structural and Multidisciplinary Optimization*, vol. 25[2], pp. 86–98, 2003.
- [75] D. Mueller-Gritschneider, H. Graeb, and U. Schlichtmann, “A successive approach to compute the bounded Pareto front of practical multiobjective optimization problems,” *SIAM Journal on Optimization*, vol. 20[2], pp. 915–934, 2009.
- [76] T. Schaul, D. Borsa, J. Modayil, and R. Pascanu, “Ray interference: a source of plateaus in deep reinforcement learning,” *arXiv preprint arXiv:1904.11455*, 2019.
- [77] D. Bertsekas, *Convex optimization algorithms*. Athena Scientific, 2015.
- [78] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, “Learning with a strong adversary,” 2015.

- [79] U. Shaham, Y. Yamada, and S. Negahban, “Understanding adversarial training: Increasing local stability of supervised models through robust optimization,” *Neurocomputing*, vol. 307, pp. 195–204, 2018.
- [80] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.
- [81] L. G. Valiant, “A theory of the learnable,” *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.
- [82] D. Haussler, “Decision theoretic generalizations of the PAC model for neural net and other learning applications,” *Information and Computation*, vol. 100, no. 1, pp. 78–150, 1992.
- [83] A. Garg and D. Roth, “Learning coherent concepts,” in *Algorithmic Learning Theory*, 2001, pp. 135–150.
- [84] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, *et al.*, “A closer look at memorization in deep networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 233–242.
- [85] A. Brutzkus and A. Globerson, “Globally optimal gradient descent for a convnet with gaussian inputs,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 605–614.
- [86] J. Neyman and E. S. Pearson, “IX. On the problem of the most efficient tests of statistical hypotheses,” *Philosophical Transactions of the Royal Society of London*, vol. 231, no. 694–706, pp. 289–337, 1933.
- [87] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice-Hall, 2005.
- [88] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, “Global solutions of variational models with convex regularization,” *SIAM Journal on Imaging Sciences*, vol. 3[4], pp. 1122–1145, 2010.

- [89] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, “Recovery of sparse translation-invariant signals with continuous basis pursuit,” *IEEE Trans. Signal Process.*, vol. 59[10], pp. 4735–4744, 2011.
- [90] O. Bar-Ilan and Y. C. Eldar, “Sub-Nyquist radar via Doppler focusing,” *IEEE Trans. Signal Process.*, vol. 62[7], pp. 1796–1811, 2014.
- [91] D. Ma, V. Gulani, N. Seiberlich, K. Liu, J. Sunshine, J. Duerk, and M. Griswold, “Magnetic resonance fingerprinting,” *Nature*, vol. 495[7440], pp. 187–192, 2013.
- [92] J. Bazerque, G. Mateos, and G. Giannakis, “Group-lasso on splines for spectrum cartography,” *IEEE Trans. Signal Process.*, vol. 59[10], pp. 4648–4663, 2011.
- [93] Y. Xie, J. Ho, and B. Vemuri, “On a nonlinear generalization of sparse coding and dictionary learning,” in *ICML*, 2013, pp. III–1480–III–1488.
- [94] M. Unser, “Sampling—50 years after Shannon,” *Proc. IEEE*, vol. 88[4], pp. 569–587, 2000.
- [95] M. Vetterli, P. Marziliano, and T. Blu, “Sampling signals with finite rate of innovation,” *IEEE Trans. Signal Process.*, vol. 50[6], pp. 1417–1428, 2002.
- [96] M. Mishali, Y. C. Eldar, and A. J. Elron, “Xampling: Signal acquisition and processing in union of subspaces,” *IEEE Trans. Signal Process.*, vol. 59[10], pp. 4719–4734, 2011.
- [97] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, 2015.
- [98] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birhäuser, 2013.
- [99] B. K. Natarajan, “Sparse approximate solutions to linear systems,” *SIAM J. Computing*, vol. 24[2], pp. 227–234, 1995.
- [100] A. Bandeira, E. Dobriban, D. Mixon, and W. Sawin, “Certifying the restricted isometry property is hard,” *IEEE Trans. Inf. Theory*, vol. 59[6], pp. 3448–3450, 2013.
- [101] A. Tillmann and M. Pfetsch, “The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 60[2], pp. 1248–1259, 2014.

- [102] A. Natarajan and Y. Wu, “Computational complexity of certifying restricted isometry property,” in *Approximation, Randomization, and Combinatorial Optimization Algorithms and Techniques*, 2014, pp. 371–380.
- [103] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, “Sensitivity to basis mismatch in compressed sensing,” *IEEE Trans. Signal Process.*, vol. 59[5], pp. 2182–2195, 2011.
- [104] B. Adcock and A. C. Hansen, “Generalized sampling and infinite-dimensional compressed sensing,” *Foundations of Computational Mathematics*, vol. 16[5], pp. 1263–1323, 2016.
- [105] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, “Breaking the coherence barrier: A new theory for compressed sensing,” *Forum of Mathematics, Sigma*, vol. 5, 2017.
- [106] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, “Compressed sensing off the grid,” *IEEE Trans. Inf. Theory*, vol. 59[11], pp. 7465–7490, 2013.
- [107] B. Bhaskar, G. Tang, and B. Recht, “Atomic norm denoising with applications to line spectral estimation,” *IEEE Trans. Signal Process.*, vol. 61[23], pp. 5987–5999, 2013.
- [108] E. J. Candès and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” *Communications on Pure and Applied Mathematics*, vol. 67[6], pp. 906–956, 2014.
- [109] M. Cho, K. Mishra, J. Cai, and W. Xu, “Block iterative reweighted algorithms for super-resolution of spectrally sparse signals,” *IEEE Signal Process. Lett.*, vol. 22[12], pp. 2319–2313, 2015.
- [110] Z. Yang and L. Xie, “On gridless sparse methods for line spectral estimation from complete and incomplete data,” *IEEE Trans. Signal Process.*, vol. 63[12], pp. 3139–3153, 2015.
- [111] G. Puy, M. E. Davies, and R. Gribonval, “Recipes for stable linear embeddings from hilbert spaces to \mathbb{R}^m ,” *IEEE Trans. Inf. Theory*, vol. 63[4], pp. 2171–2187, 2017.
- [112] S. Bruno, S. Ahmed, A. Shapiro, and A. Street, “Risk neutral and risk-averse approaches to multistage renewable investment planning under uncertainty,” *European Journal of Operational Research*, vol. 250, no. 3, pp. 979–989, 2016.

- [113] S. Moazeni, W. B. Powell, and A. H. Hajimiragha, “Mean-conditional value-at-risk optimal energy storage operation in the presence of transaction costs,” *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1222–1232, 2015.
- [114] H. Markowitz, “Portfolio selection,” *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.
- [115] H. Föllmer and A. Schied, “Convex measures of risk and trading constraints,” *Finance and Stochastics*, vol. 6, no. 4, pp. 429–447, 2002.
- [116] D. Shang, V. Kuzmenko, and S. Uryasev, “Cash flow matching with risks controlled by buffered probability of exceedance and conditional value-at-risk,” *Annals of Operations Research*, vol. 260, no. 1-2, pp. 501–514, 2018.
- [117] S.-K. Kim, R. Thakker, and A.-a. Agha-mohammadi, “Bi-directional value learning for risk-aware planning under uncertainty,” *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2493–2500, 2019.
- [118] A. A. Pereira, J. Binney, G. A. Hollinger, and G. S. Sukhatme, “Risk-aware path planning for autonomous underwater vehicles using predictive ocean models,” *Journal of Field Robotics*, vol. 30, no. 5, pp. 741–762, 2013.
- [119] A. S. Bedi, A. Koppel, and K. Rajawat, “Nonparametric compositional stochastic optimization,” *Arxiv*, 2019.
- [120] W.-J. Ma, C. Oh, Y. Liu, D. Dentcheva, and M. M. Zavlanos, “Risk-averse access point selection in wireless communication networks,” *IEEE Transactions on Control of Network Systems*, vol. 5870, no. c, pp. 1–1, 2018.
- [121] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*, 2nd ed. Society for Industrial and Applied Mathematics, 2009.
- [122] D. S. Kalogerias and W. B. Powell, “Recursive optimization of convex risk measures: Mean-semideviation models,” *Extended Preprint, Arxiv*, 2018.
- [123] R. T. Rockafellar and S. Uryasev, “Optimization of conditional value-at-risk,” *Journal of Risk*, vol. 2, pp. 21–41, 1997.

- [124] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” *arXiv preprint arXiv:1611.03530*, 2016.
- [125] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2010.
- [126] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.
- [127] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [128] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.
- [129] C. A. Micchelli and M. Pontil, “Learning the kernel function via regularization,” *Journal of machine learning research*, vol. 6, no. Jul, pp. 1099–1125, 2005.
- [130] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
- [131] R. Rockafellar and R. Wets, *Variational Analysis*. Springer, 1998.
- [132] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*. Stanford University Press, 1958.
- [133] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.
- [134] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [135] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980v9*, 2017.
- [136] ProPublica, “COMPAS dataset analysis,” 2016, <https://github.com/propublica/compas-analysis/>.

- [137] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014.
- [138] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [139] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [140] G. Kimeldorf and G. Wahba, “A correspondence between bayesian estimation on stochastic processes and smoothing by splines,” *The Annals of Mathematical Statistics*, vol. 41[2], pp. 495–502, 1970.
- [141] B. Schölkopf, R. Herbrich, and A. J. Smola, “A generalized representer theorem,” in *International conference on computational learning theory*. Springer, 2001, pp. 416–426.
- [142] V. Duval and G. Peyré, “Sparse regularization on thin grids I: the Lasso,” *Inverse Problems*, vol. 33[5], p. 055008, 2017.
- [143] —, “Sparse spikes super-resolution on thin grids II: the continuous basis pursuit,” *Inverse Problems*, vol. 33[9], p. 095008, 2017.
- [144] A. Beck and Y. Eldar, “Sparsity constrained nonlinear optimization: Optimality conditions and algorithms,” *SIAM Journal on Optimization*, vol. 23[3], pp. 1480–1509, 2013.
- [145] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang, “Sparse nonlinear regression: Parameter estimation under nonconvexity,” in *ICML*, 2016, pp. 2472–2481.
- [146] A. Shapiro, “On duality theory of convex semi-infinite programming,” *Optimization*, vol. 54[6], pp. 535–543, 2006.
- [147] E. Hendrix and B. G.-Tóth, *Introduction to Nonlinear and Global Optimization*. Springer, 2010.
- [148] J. Diestel and J. J. Uhl, Jr., *Vector measures*. AMS, 1977.

- [149] A. Ruszczyński and W. Syski, “On convergence of the stochastic subgradient method with on-line stepsize rules,” *Journal of Mathematical Analysis and Applications*, vol. 114[2], pp. 512–527, 1986.
- [150] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” 2016, arXiv:1606.04838.
- [151] W. Rudin, *Functional Analysis*. McGraw-Hill, 1991.
- [152] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
- [153] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2018.
- [154] M. da Costa and W. Dai, “Sampling patterns for off-the-grid spectral estimation,” in *Asilomar*, 2017, pp. 318–322.
- [155] N. Rao, P. Shah, and S. Wright, “Forward–backward greedy algorithms for atomic norm regularization,” *IEEE Trans. Signal Process.*, vol. 63[21], pp. 5798–5811, 2015.
- [156] J. Ramsay and B. Silverman, *Functional Data Analysis*. Springer, 2005.
- [157] Y. Chen, C. Caramanis, and S. Mannor, “Robust sparse regression under adversarial corruption,” in *ICML*, 2013, pp. 774–782.
- [158] J. Feng, H. Xu, S. Mannor, and S. Yan, “Robust logistic regression and classification,” in *Advances in Neural Information Processing Systems*, 2014, pp. 253–261.
- [159] Y. Plan and R. Vershynin, “Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach,” *IEEE Trans. Inf. Theory*, vol. 59[1], pp. 482–494, 2013.
- [160] J. Tibshirani and C. Manning, “Robust logistic regression using shift parameters,” in *Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 124–129.
- [161] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, “UEA & UCR time series classification repository: ECG200 dataset,” <http://www.timeseriesclassification.com/description.php?Dataset=ECG200>.

- [162] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng, and H. Stanley, “Physiobank, physiotoolkit, and physionet,” *Circulation*, vol. 101[23], pp. e215–e220, 2000, <http://physionet.org/physiobank/database/svdb/>.
- [163] L. Ye, S. Roy, and S. Sundaram, “On the complexity and approximability of optimal sensor selection for Kalman filtering,” 2017, arXiv:1711.01920v1.
- [164] A. Olshevsky, “On (non)supermodularity of average control energy,” *IEEE Trans. Contr. Netw. Syst.*, vol. 5[3], pp. 1177–1181, 2018.
- [165] V. Tzoumas, M. Rahimian, G. Pappas, and A. Jadbabaie, “Minimal actuator placement with bounds on control effort,” *IEEE Trans. Contr. Netw. Syst.*, vol. 3[1], pp. 67–78, 2016.
- [166] A. Krause and D. Golovin, “Submodular function maximization,” in *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014.
- [167] J. Ranieri, A. Chebira, and M. Vetterli, “Near-optimal sensor placement for linear inverse problems,” *IEEE Trans. Signal Process.*, vol. 62[5], pp. 1135–1146, 2014.
- [168] F. Bach, “Learning with submodular functions: A convex optimization perspective,” *Foundations and Trends in Machine Learning*, vol. 6[2-3], pp. 145–373, 2013.
- [169] G. Nemhauser, L. Wolsey, and M. Fisher, “An analysis of approximations for maximizing submodular set functions—I,” *Mathematical Programming*, vol. 14[1], pp. 265–294, 1978.
- [170] T. Horel and Y. Singer, “Maximization of approximately submodular functions,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3045–3053.
- [171] B. Lehmann, D. Lehmann, and N. Nisan, “Combinatorial auctions with decreasing marginal utilities,” *Games and Economic Behavior*, vol. 55[2], pp. 270–296, 2006.
- [172] M. Sviridenko, J. Vondrák, and J. Ward, “Optimal approximation for submodular and supermodular optimization with bounded curvature,” in *SIAM Symposium on Discrete Algorithms*, 2014, pp. 1134–1148.
- [173] A. Bian, J. Buhmann, A. Krause, and S. Tschischek, “Guarantees for greedy maximization of non-submodular functions with applications,” in *Int. Conf. on Mach. Learning*, 2017.

- [174] A. Das and D. Kempe, “Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection,” in *Int. Conf. on Mach. Learning*, 2011.
- [175] O. Karaca and M. Kamgarpour, “Exploiting weak supermodularity for coalition-proof mechanisms,” in *Conf. on Decision and Contr.*, 2018, pp. 1118–1123.
- [176] T. Summers and M. Kamgarpour, “Performance guarantees for greedy maximization of non-submodular set functions in systems and control,” in *European Contr. Conf.*, 2019, pp. 2796–2801.
- [177] A. Krause and V. Cevher, “Submodular dictionary selection for sparse representation,” in *Int. Conf. on Mach. Learning*, 2010.
- [178] R. Horn and C. Johnson, *Matrix analysis*. Cambridge University Press, 2013.
- [179] R. Bhatia, *Matrix analysis*. Springer, 1997.
- [180] L. Lovász, “Submodular functions and convexity,” in *Mathematical Programming: The State of the Art*. Springer-Verlag, 1982.
- [181] A. Schrijver, *Combinatorial Optimization*. Springer, 2003.
- [182] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, “An analysis of approximations for maximizing submodular set functions—ii,” in *Polyhedral combinatorics*. Springer, 1978, pp. 73–87.
- [183] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” *IEEE Signal Process. Mag.*, vol. 30[3], pp. 83–98, 2013.
- [184] A. Sandryhaila and J. Moura, “Discrete signal processing on graphs,” *IEEE Trans. Signal Process.*, vol. 61[7], pp. 1644–1656, 2013.
- [185] S. Chen, R. Varma, A. Singh, and J. Kovačević, “Signal recovery on graphs: Fundamental limits of sampling strategies,” 2016, arXiv:1512.05405v2.
- [186] A. Anis, A. Gadde, and A. Ortega, “Efficient sampling set selection for bandlimited graph signals using graph spectral proxies,” *IEEE Trans. Signal Process.*, vol. 64[14], pp. 3775–3789, 2016.

- [187] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, “Discrete signal processing on graphs: Sampling theory,” *IEEE Trans. Signal Process.*, vol. 63[24], pp. 6510–6523, 2015.
- [188] H. Shomorony and A. Avestimehr, “Sampling large data on graphs,” in *Global Conf. on Signal and Inform. Process.*, 2014, pp. 933–936.
- [189] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, “Signals on graphs: Uncertainty principle and sampling,” 2016, arXiv:1507.08822v3.
- [190] T. Adali and P. Schreier, “Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation,” *IEEE Signal Process. Mag.*, vol. 31[5], pp. 112–128, 2014.
- [191] B. Girault, “Stationary graph signals using an isometric graph translation,” in *European Signal Process. Conf.*, 2015, pp. 1516–1520.
- [192] A. Marques, S. Segarra, G. Leus, and A. Ribeiro, “Stationary graph processes and spectral estimation,” 2016, arXiv:1603.04667v1.
- [193] N. Perraudin and P. Vandergheynst, “Stationary signal processing on graphs,” 2016, arXiv:1601.02522v3.
- [194] P. Di Lorenzo, S. Barbarossa, P. Banelli, and S. Sardellitti, “Adaptive least mean squares estimation of graph signals,” *IEEE Trans. Signal Inf. Process. over Netw.*, vol. 2[4], pp. 555–568, 2016.
- [195] X. Wang, P. Liu, and Y. Gu, “Local-set-based graph signal reconstruction,” *IEEE Trans. Signal Process.*, vol. 63[9], pp. 2432–2444, 2015.
- [196] A. Krause, A. Singh, and C. Guestrin, “Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies,” *J. Mach. Learning Research*, vol. 9, pp. 235–284, 2008.
- [197] G. Sagnol, “Approximation of a maximum-submodular-coverage problem involving spectral functions, with application to experimental designs,” *Discrete Appl. Math.*, vol. 161[1-2], pp. 258–276, 2013.

- [198] D. Thanou, D. Shuman, and P. Frossard, “Learning parametric dictionaries for signals on graphs,” *IEEE Trans. Signal Process.*, vol. 62[15], pp. 3849–3862, 2014.
- [199] S. Joshi and S. Boyd, “Sensor selection via convex optimization,” *IEEE Trans. Signal Process.*, vol. 57[2], pp. 451–462, 2009.
- [200] S. Chepuri and G. Leus, “Subsampling for graph power spectrum estimation,” in *Sensor Array and Multichannel Signal Process. Workshop*, 2016.
- [201] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286[5439], pp. 509–512, 1999.
- [202] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10[5], pp. 1299–1319, 1998.
- [203] J. Arenas-García, K. Petersen, G. Camps-Valls, and L. Hansen, “Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods,” *IEEE Signal Process. Mag.*, vol. 30[4], pp. 16–29, 2013.
- [204] C. Bishop, *Pattern recognition and machine learning*. Springer, 2007.
- [205] M. Tipping, “Sparse kernel principal component analysis,” in *NIPS*, 2000.
- [206] Y. Washizawa, “Subset kernel principal component analysis,” in *Int. Workshop on Mach. Learning for Signal Process.*, 2009.
- [207] D. Woodruff, “Sketching as a tool for numerical linear algebra,” *Foundations and Trends in Theoretical Computer Science*, vol. 10[1-2], pp. 1–157, 2014.
- [208] D. Feldman, M. Schmidt, and C. Sohler, “Turning big data into tiny data: Constant-size coresets for K-means, PCA and projective clustering,” in *ACM-SIAM Symp. on Discrete Algorithms*, 2013, pp. 1434–1453.
- [209] A. L. Cambridge, “The ORL database of faces,” 1994, <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>.

- [210] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Netw.*, vol. 13[2], pp. 415–425, 2002.
- [211] H. Zhang, R. Ayoub, and S. Sundaram, “Sensor selection for Kalman filtering of linear dynamical systems: Complexity, limitations and greedy algorithms,” *Automatica*, vol. 78, pp. 202–210, 2017.
- [212] F. Muller, F. Seyler, and J.-L. Guyot, “Utilisation d’imagerie radar (ROS) JERS-1 pour l’obtention de réseaux de drainage. Exemple du Rio Negro (Amazonie),” in *Hydrological and Geochemical Processes in Large Scale River Basins*, 1999, <http://www.ore-hybam.org/index.php/eng/Data/Cartography/Amazon-basin-hydrography>.
- [213] K. Yu, J. Bi, and V. Tresp, “Active learning via transductive experimental design,” in *International Conference on Machine Learning*, 2006, pp. 1081–1088.
- [214] F. Pukelsheim, *Optimal Design of Experiments*. SIAM, 2006.
- [215] T. Summers, F. Cortesi, and J. Lygeros, “On submodularity and controllability in complex dynamical networks,” *IEEE Trans. Contr. Netw. Syst.*, vol. 3[1], pp. 91–101, 2016.
- [216] A. Das and D. Kempe, “Algorithms for subset selection in linear regression,” in *ACM Symp. on Theory of Comput.*, 2008, pp. 45–54.
- [217] M. Benzi, “Preconditioning techniques for large linear systems: A survey,” *Journal of Computational Physics*, vol. 182[2], pp. 418–477, 2002.
- [218] R. Braatz and M. Morari, “Minimizing the Euclidian condition number,” *SIAM Journal of Control and Optimization*, vol. 32[6], pp. 1763–1768, 1994.
- [219] Y. Wang, A. Yu, and A. Singh, “On computationally tractable selection of experiments in regression models,” 2017, arXiv:1601.02068v5.
- [220] D. E. Corporation, “EachMovie dataset,” <http://www.gatsby.ucl.ac.uk/~chuwei/data/EachMovie/>.
- [221] A. R. Cardoso and H. Xu, “Risk-averse stochastic convex bandit,” in *International Conference on Artificial Intelligence and Statistics*, vol. 89, 2019, pp. 39–47.

- [222] W. Huang and W. B. Haskell, “Risk-aware q-learning for markov decision processes,” in *2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017*, vol. 2018-Janua. IEEE, 2018, pp. 4928–4933.
- [223] D. R. Jiang and W. B. Powell, “Risk-averse approximate dynamic programming with quantile-based risk measures,” *Mathematics of Operations Research*, p. moor.2017.0872, 2017.
- [224] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, “Sequential decision making with coherent risk,” *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3323–3338, 2017.
- [225] C. Vitt, D. Dentcheva, and H. Xiong, “Risk-averse classification,” *Arxiv*, 2018.
- [226] L. Zhou and P. Tokekar, “An approximation algorithm for risk-averse submodular optimization,” *Arxiv*, 2018.
- [227] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.
- [228] P. Whittle, “Risk-sensitive linear/quadratic/gaussian control,” *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981.
- [229] J. L. Speyer and W. H. Chung, *Stochastic Processes, Estimation, and Control*. Siam, 2008, vol. 17.
- [230] J. B. Moore, R. J. Elliott, and S. Dey, “Sensitive generalization of minimum variance estimation and control,” in *IFAC Symposium on Nonlinear Control Systems Design*, 1995, pp. 423–428.
- [231] S. Dey and J. B. Moore, “Risk-sensitive filtering and smoothing via reference probability methods,” *IEEE Trans. Autom. Control*, vol. 42, no. 11, pp. 1587–1591, 1997.
- [232] —, “Finite-dimensional risk-sensitive filters and smoothers for discrete-time nonlinear systems,” *IEEE Trans. Autom. Control*, vol. 44[6], pp. 1234–1239, 1999.
- [233] D. G. Luenberger, *Optimization by Vector Space Methods*. Wiley, 1968.
- [234] A. Shapiro, “Semi-infinite programming, duality, discretization and optimality conditions,” *Optimization*, vol. 58, no. 2, pp. 133–161, 2009.

- [235] H. Henderson and S. Searle, “On deriving the inverse of a sum of matrices,” *SIAM Review*, vol. 23[1], pp. 53–60, 1981.
- [236] B. Wang and F. Zhang, “Some inequalities for the eigenvalues of the product of positive semidefinite Hermitian matrices,” *Linear Algebra and its Applications*, vol. 160, pp. 113–118, 1992.
- [237] A. Marshall, I. Olkin, and B. Arnold, *Inequalities: Theory of Majorization and Its Applications*. Springer, 2009.
- [238] D. Bertsekas, *Dynamic Programming and Optimal Control – Vol. I*. Athena Scientific, 2017.