June 2019

Structured Sparsity Promoting Functions: Theory and Applications

Erin Tripp
Syracuse University

Abstract

Motivated by the minimax concave penalty based variable selection in high-dimensional linear regression, we introduce a simple scheme to construct structured semiconvex sparsity promoting functions from convex sparsity promoting functions and their Moreau envelopes. Properties of these functions are developed by leveraging their structure. In particular, we show that the behavior of the constructed function can be easily controlled by assumptions on the original convex function. We provide sparsity guarantees for the general family of functions via the proximity operator. Results related to the Fenchel Conjugate and Łojasiewicz exponent of these functions are also provided. We further study the behavior of the proximity operators of several special functions including indicator functions of closed convex sets, piecewise quadratic functions, and linear combinations of the two. To demonstrate these properties, several concrete examples are presented and existing instances are featured as special cases. We explore the effect of these functions on the penalized least squares problem and discuss several algorithms for solving this problem which rely on the particular structure of our functions. We then apply these methods to the total variation denoising problem from signal processing.

Structured Sparsity Promoting Functions: Theory and Applications

by

Erin E. Tripp

B.S., University of California, Santa Barbara, 2013

M.S., Syracuse University, 2017

Dissertation

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematics

Syracuse University

June 2019

Acknowledgements

I would like to express my profound gratitude to my advisor Professor Lixin Shen for his guidance, patience, and unending support. He has been a wonderful mentor and collaborator, and the training he has provided will carry me throughout my career.

I would like to thank Dr. Bruce Suter of the Air Force Research Laboratory for his mentorship. This work would not have been possible without him.

I would also like to thank my PhD Committee– Professor Uday Banerjee, Professor Biao Chen, Professor Leonid Kovalev, Professor Minghao Rostami, and Professor Andrew Vogel– for their helpful comments and insightful suggestions.

I am grateful to the faculty and staff of the Mathematics Department. In particular, I would like to acknowledge Professor Graham Leuschke and Mr. Jordan Correia for saving me from my many clerical mistakes.

Finally, I would like to thank my fellow graduate students, past and present, for their friendship and support. None of us are in this alone.

Contents

List of Figures

Chapter 1

Introduction

This dissertation concerns the study and application of sparsity promoting functions. Informally, a vector or matrix is sparse if it has few nonzero entries, and a function is sparsity promoting if it penalizes nonzero entries. Including such functions in an optimization problem, either as a constraint or as a penalty term, encourages sparse solutions. Interest in sparsity spans applications from image processing to machine learning and statistics for two primary reasons: sparsity describes structure, and sparse data is easier to manipulate and interpret. Natural signals and data are often sparse in an appropriate basis (see, e.g. [12, 35, 50]). For example, pixel values in an image are constant in blocks corresponding to objects or pieces of objects, so the image matrix is sparse in the basis defined by pixel differences. The feature selection problem in machine learning concerns identifying the most relevant components of the data and discarding the rest. Assuming there is such a sparse representation, it allows us to greatly decrease the dimension of the problem (e.g., [22]). We can also consider finding sparse representations as a method of compressing data, whether that be for computational efficiency or for security.

The natural mathematical measure of sparsity is the so-called "ℓ_0-norm", which simply returns the number of nonzero entries. That is, for $x \in \mathbb{R}^n$,

$$\|x\|_0 = \big|\{i : x_i \neq 0, i \in \{1, \ldots, n\}\}\big|.$$

This function is not truly a norm; it is nonconvex, discontinuous, and homogeneous of degree zero. Moreover, solving the ℓ_0-penalized least squares problem is combinatorial in nature and known to be NP hard.

To overcome these issues, regularization methods replacing the ℓ_0-penalty with the ℓ_1-penalty such as LASSO [51] and Dantzig selectors [12] have been proposed. This relaxation makes the problem numerically tractable and allows application of the many tools of convex optimization. However the convexity of the ℓ_1-norm introduces bias in the solution by heavily penalizing entries with large magnitude. To address this, nonconvex penalties including the ℓ_q-penalty with $0 < q < 1$ [23], the smoothly clipped absolute deviation penalty (SCAD) [22], the Continuous Exact ℓ_0 penalty (CEL0) [48], and the minimax concave penalty (MCP) [53] have been proposed to replace the ℓ_1-norm penalty.

We introduce a family of semiconvex sparsity promoting functions of which each is the difference of a convex sparsity promoting function with its Moreau envelope. We show that, as long as a convex function is a sparsity promoting function, so is this difference. This result makes the construction of nonconvex sparsity promoting functions effortless. Some interesting properties of such functions are: (i) they are always non-negative and semiconvex and (ii) they are a special type of difference of convex (DC) functions with one having a Lipschitz continuous gradient. Due to these properties, we refer to these functions as structured semiconvex sparsity promoting functions. These properties enable us to make use

of the fruitful results, for example, in DC programming [38], to develop efficient algorithms for the associated regularized optimization problems. What's more, these functions provide a bridge between convex and nonconvex sparsity promoting penalties. As a specific example, we recover the MCP from the difference of the ℓ_1-norm and its envelope. It has been shown (e.g. in [49]) that this closely approximates the ℓ_0-norm while preserving the continuity and subdifferentiability of ℓ_1.

The proximity operator, which was first introduced by Moreau in [36] as a generalization of the notion of projection onto a convex set, has been used extensively in nonlinear optimization (see, e.g., [4, 5, 16]). The desired features of the aforementioned regularization methods can be explained in terms of the proximity operators of the corresponding penalties. Therefore, to determine the effectiveness of our proposed functions, we must examine the behavior of their proximity operators. The proximity operator of the ℓ_0-norm is the hard thresholding operator, which annihilates all entries below a certain threshold and keeps all entries above the threshold. In fact, we see that hard thresholding rules are characteristic of penalties which are concave near the origin and constant elsewhere. More generally, we provide sparsity guarantees in terms of thresholding behavior for the entire family of structured semiconvex sparsity promoting functions, with further details for certain special functions.

Working with nonconvex functions provides greater model flexibility and accuracy, but there is no general theory of nonconvex optimization, a dilemma which mirrors the earlier transition from linear to nonlinear programming. The following quote attributed to Stanislaw Ulam describes the situation:

> Using a term like nonlinear science is like referring to the bulk of zoology as the study of non-elephant animals.

Identifying functions as nonlinear or nonconvex describes them only by the structure that

they lack and does not take advantage of any structure that they have. In the case of nonlinear programming, convexity was the key property that allowed us to move forward, and many years later convex analysis is considered a fundamental part of the study of optimization (see e.g. [27, 26, 6, 10]). In the same vein, we can take steps into the nonconvex world through functions with some nice structure. One approach is to consider generalizations of convexity like quasiconvexity and semiconvexity. There has been quite a lot of work done generalizing results from convex analysis to these classes (see, e.g., [25]). Another approach considers properties like subanalyticity and Kurdyka-Łojasiewicz inequalities (see, e.g., [8]). These are not mutually exclusive categories, and we will discuss both types of structure for our sparsity promoting functions.

Motivation

Our construction of semiconvex sparsity promoting functions was motivated mainly by the minimax concave penalty (MCP) based variable selection in high-dimensional linear regression [53]. Variable selection is fundamental in statistical analysis of high-dimensional data. It is also easily interpretable in terms of sparse signal recovery. We consider a linear regression model with n-dimensional response vector y, $n \times p$ model matrix X, p-dimensional regression vector γ, and n-dimensional error vector ϵ:

$$y = X\gamma + \epsilon.$$

The goal of variable selection is to recover the true underlying sparse model of the pattern $\{j : \gamma_j \neq 0\}$ and to estimate the non-zero regression coefficients γ_j, where γ_j is the j-th component of γ. For small p, subset selection methods can be used to find a good guess of

the pattern (see, e.g., [45]). However, subset selection becomes computationally infeasible for large p.

To overcome the computational difficulties of subset selection method, the method of penalized least squares is widely used in variable selection to produce meaningful interpretable models:

$$\min_{\gamma\in\mathbb{R}^p}\left[\frac{1}{2n}\|y-X\gamma\|^2+\sum_{j=1}^{p}\rho(|\gamma_j|,\lambda)\right], \tag{1.1}$$

where $\rho(\cdot,\lambda)$ is a penalty function indexed by $\lambda\geq 0$. The penalty function $\rho(t,\lambda)$, defined on $[0,+\infty)$, is assumed to be nondecreasing in t with $\rho(0,\lambda)=0$ and continuously differentiable for $t\in(0,+\infty)$. The formulation in (1.1) includes many popular variable selection methods. For example, the best subset selection amounts to using the ℓ_0 penalty $\rho(|t|,\lambda)=\frac{\lambda^2}{2}\mathbb{1}_{\{|t|\neq 0\}}$ while LASSO [51] and basis pursuit [15] use the ℓ_1 penalty $\rho(|t|,\lambda)=\lambda|t|$. Here $\mathbb{1}_{\{u\in E\}}$ denotes the characteristic function and $\mathbb{1}_{\{u\in E\}}$ equals 1 if $u\in E$ and 0 otherwise. The estimator (the hard thresholding operator) with the ℓ_0 penalty suffers from instability in model prediction while the estimator (the soft thresholding operator) with the ℓ_1 penalty suffers from the bias issue, severely interfering with variable selection for large p [22]. To remedy this issue, the SCAD penalty was introduced in [22]. The estimator with the SCAD penalty is continuous and leaves large components not excessively penalized. In [53], the MCP penalty was introduced and is defined as follows

$$\rho(|t|,\lambda)=\lambda\int_0^{|t|}\max\left\{0,1-\frac{x}{a\lambda}\right\}\,dx, \tag{1.2}$$

where the parameter $a>0$. This penalty function (see [53]) minimizes the maximum concavity

$$\kappa(\rho,\lambda):=\sup_{0<t_1<t_2<\infty}-\frac{\rho(t_2,\lambda)-\rho(t_1,\lambda)}{t_2-t_1}$$

subject to the unbiasedness $\frac{\partial}{\partial t}\rho(t,\lambda) = 0$ for all $t > a\lambda$ and selection features $\frac{\partial}{\partial t}\rho(0+,\lambda) = \lambda$. The number $\kappa(\rho,\lambda)$ is related to the computational complexity of regularization method for solving (1.1). The simulations in [22, 53] gave a strong statistical evidence that the estimators from the non-convex penalty functions SCAD and MCP are useful in variable selection. Recently, an application of MCP to signal processing was reported in [46].

Due to its success in applications, we take a closer look at MCP. The MCP function in (1.2) can be rewritten as

$$\rho(|t|,\lambda) = \lambda(|t| - \mathrm{env}_{a\lambda}\,|\cdot|(t)),$$

where $\mathrm{env}_{a\lambda}\,|\cdot|$ is the Moreau envelope of $|\cdot|$ with index $a\lambda$ (see next section). Clearly, the MCP penalty can be considered as a variation of the ℓ_1 penalty function, that is, the absolution function $|\cdot|$ is replaced by $|\cdot| - \mathrm{env}_{a\lambda}\,|\cdot|$. From this simple observation, we are drawn to consider a family of penalty functions defined by

$$f - \mathrm{env}_{\alpha}\, f$$

with f satisfying some proper properties and $\alpha > 0$.

Contributions

The theoretical contributions of this thesis include

- providing an easily verifiable definition of sparsity promotion;
- introducing a simple construction of nonconvex sparsity promoting functions;
- characterizing sparse thresholding behavior of proximity operator of the proposed spar-

sity promoting functions;

- providing information about the conjugate and dual problems of the proposed sparsity promoting functions; and
- studying the Łojasiewicz property and providing the Łojasiewicz exponent for the proposed sparsity promoting functions.

The contributions of this thesis in applications include

- studying special classes of sparsity promoting functions and providing examples;
- demonstrating algorithmic performance;
- demonstrating applicability the proposed sparsity promoting functions to denoising signals and images corrupted by noise.

Content

Chapter 2 introduces notation and reviews the essential tools from convex analysis, nonsmooth optimization, and real analytic geometry. Where there is no standard notation or name in the literature we choose one that is at least widely recognized and make a note of other possible conventions. Sections 2.1-2.3 reviews the basic objects of convex analysis. Section 2.4 provides a deeper examination of two fundamental concepts: the Moreau envelope and the proximity operator. We make an effort to give insight and intuition about these operators as they figure into every aspect of our research. Section 2.5 briefly reviews sets and functions definable on o-minimal structures.

Chapter 3 contains the theoretical aspects of our work. We define sparsity promoting functions, introduce a construction of nonconvex sparsity promoting functions from convex

sparsity promoting functions, and study the properties of both. Section 3.1 contains this definition and a characterization of convex sparsity promoting functions. Section 3.2 deals with our construction of families of nonconvex sparsity promoting functions and details what can be considered their core properties: sparsity promotion and semiconvexity. Many of these results are included in our paper of the same title [47]. Section 3.3 explores further properties of both convex and nonconvex sparsity promoting functions. Specifically, we look at the Kurdyka-Łojasiewicz property and the Fenchel conjugate.

In Chapter 4, we refine our results for certain classes of functions. Section 4.2 deals with quadratic functions, Section 4.1 with indicator functions, and Section 4.3 with sums of the two. We close the chapter with several examples drawn from a variety of applications in Section 4.4.

We highlight the benefits of our construction by studying its algorithmic performance in Chapter 5. In particular, we consider a penalized least squares problem and apply three widely used algorithms: Primal-Dual Splitting, Difference of Convex, and Alternating Directions Method of Multipliers. Each algorithm corresponds to a different view of the model through the lens of our proposed penalty function. Convergence analysis is provided for each and improved when possible.

In Chapter 6, we apply the model and algorithms to the problem of signal and image denoising. Numerical results are provided for each algorithm, as well as a comparison to the standard Rudin-Osher-Fatemi total variation denoising model for each case. We also offer insight into parameter choices for tuning the model.

Finally, we conclude with a summary of our results and describe their context in current optimization research. We discuss remaining questions as well as future directions and possible applications.

Chapter 2

Preliminaries

All functions and sets are defined on $\mathbb{R}^n$ equipped with the inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. We denote the boundary, interior, and closure of a set A by $\mathrm{bd}(A), \mathrm{int}(A)$, and $\overline{A}$ respectively. The relative interior $\mathrm{ri}(A)$ is the interior of A when viewed as a subset of the affine space that it spans.

We consider functions from $\mathbb{R}^n$ to the extended real line $f : \mathbb{R}^n \to (-\infty, +\infty]$. The domain of f is

$$\mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

We say that f is proper if $\mathrm{dom}(f) \neq \emptyset$. The graph of f is $\mathrm{gr}(f) = \{(x, y) \in \mathrm{dom}(f) \times \mathbb{R} : f(x) = y\}$ and the epigraph $\mathrm{epi}(f) = \{(x, \xi) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \xi\}$. Recall that f is lower semicontinuous if $\liminf_{y \to x} f(y) \geq f(x)$ for every x. We denote

$$\Gamma(\mathbb{R}^n) = \{ \text{ proper, lower semicontinuous functions } f : \mathbb{R}^n \to (-\infty, +\infty]\}.$$

If $f : \mathbb{R}^n \to (-\infty, +\infty]$ is Lipschitz continuous with constant L, i.e. $|f(x) - f(y)| \leq L\|x - y\|$

for all $x, y \in \mathbb{R}^n$, we say f is L-Lipschitz.

The results included in this chapter are classical and can be found in essentially any text on convex analysis (see, e.g. [5, 27, 6]).

2.1 Convexity and Semiconvexity

We briefly review the definitions, provide some examples, and cover some of the essential properties which make convex sets and functions so useful for optimization.

A set $C \subseteq \mathbb{R}^n$ is convex if for all $x, y \in C$ and any $\lambda \in [0, 1]$, the point $\lambda x + (1 - \lambda)y \in C$ as well. That is, C contains the line segment joining any two of its points. Some familiar examples are intervals in $\mathbb{R}$ and balls $B_r(x) = \{y : \|x - y\| < r\}$ in $\mathbb{R}^n$. For nonempty convex sets C, the relative interior can be written as follows [44]:

$$\mathrm{ri}(C) = \{z \in C : \forall x \in C \ \exists \lambda > 1 \, ((1 - \lambda)x + \lambda z \in C)\}.$$

In particular, if $0 \in C$, then for any $x \in \mathrm{ri}(C)$, there exists $\lambda > 1$ such that $\lambda x \in C$. Recall that if C is closed as well as convex then for any $x \in \mathbb{R}^n$, the projection of x onto C, denoted $\Pi_C(x)$, exists and is unique.

A function f is convex if for any $x, y \in \mathrm{dom}(f)$ and any $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

If the inequality above is strict for all $x \neq y$, then we say f is strictly convex. Equivalently, a function is convex if its epigraph is a convex set. We denote by $\Gamma_0(\mathbb{R}^n)$ the set of proper, lower semicontinuous, convex functions $f : \mathbb{R}^n \to (-\infty, +\infty]$.

For $\sigma > 0$, f is σ-strongly convex if $f - \frac{\sigma}{2}\|\cdot\|^2$ is convex. Similarly, for $\rho > 0$, f is ρ-semiconvex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex. Of course, a strongly convex function is also convex, and a convex function is also semiconvex. Semiconvex functions can be further generalized to prox-regular [43] and primal lower nice functions [42]. For a more thorough study of semiconvex functions with applications to variational analysis see [13].

For example, any norm $\|\cdot\|$ is convex, and the quadratic function $\frac{1}{2}\|\cdot\|^2$ is 1-strongly convex. Both examples are proper and continuous. Any function which is convex is trivially semiconvex, and if f is convex, the function $f - \frac{\rho}{2}\|\cdot\|^2$ is ρ-semiconvex. A particularly useful example is the indicator function of a set C defined as

$$\iota_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{if } x \notin C. \end{cases}$$

The function ι_C is convex (as a function) if and only if C is convex (as a set). This is also sometimes called the characteristic function of C.

2.2 Derivatives and Subdifferentials

Perhaps the most fundamental theorem in optimization is also the most familiar; Fermat's rule states that if $\bar{x}$ is a maximizer or minimizer of a differentiable function f, then $\bar{x}$ is a critical point of f. The field can be largely summarized as tools and methods for finding critical points and verifying that they are optimal. To provide context, we review some important results for differentiable functions. We then introduce a generalized derivative called the Fréchet subdifferential which naturally extends these results to nonsmooth functions.

Recall that a function f is differentiable at x if there exists a bounded linear operator

$B : \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{u \to x} \frac{|f(u) - f(x) - B(u - x)|}{\|u - x\|} = 0.$$

If this operator exists, we denote it by $\nabla f(x)$. If $f \in \Gamma_0(\mathbb{R}^n)$ is differentiable with gradient ∇f, then it satisfies the following inequality for all $x, y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

In other words, the graph of f is supported by its tangent hyperplanes. An immediate consequence is that any critical point $\bar{x}$, i.e. any point such that $\nabla f(\bar{x}) = 0$, must be a global minimizer. Thus for convex functions, finding a global minimum is equivalent to finding a critical point. We note that while the minimum value is unique, convex functions may have more than one minimizer. In fact, they may have a continuum of minimizers. If f is strongly convex, however, the minimizer is unique. We denote by $\operatorname{Argmin} f$ the set of minimizers of f.

The Fréchet subdifferential of a function f at x is

$$\partial f(x) \coloneqq \Big\{ \eta \in \mathbb{R}^n : \liminf_{u \to x} \frac{f(u) - f(x) - \langle \eta, u - x \rangle}{\|u - x\|} \geq 0 \Big\}.$$

An element $\eta \in \partial f(x)$ is called a subgradient of f at x. If $\partial f(x) \neq \emptyset$ we say that f is Fréchet subdifferentiable at x. Viewed as a set-valued operator from $\mathbb{R}^n$ to itself, we define the domain $\operatorname{dom}(\partial f)$ as the set of all points x at which $\partial f(x)$ is nonempty. Note that $\operatorname{dom}(\partial f) \subseteq \operatorname{dom}(f)$.

The subdifferential generalizes the usual Fréchet derivative: if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. The subdifferential obeys a generalization of Fermat's rule: if f

attains a local minimum at x, then $0 \in \partial f(x)$ and the point x is called a (generalized) critical point of f. The set of critical points of f is denoted $\operatorname{crit} f$.

For example, $f(x) = |x|$ is differentiable for all nonzero x and subdifferentiable at the origin with $\partial \| \cdot \|_1(0) = \overline{B_1}(0)$. The function $q(x) = \frac{1}{2}\|x\|^2$ is differentiable everywhere, thus for all $x \in \mathbb{R}^n$, we have $\partial q(x) = \{x\}$. The subdifferential of the indicator function ι_C at a point $x \in C$ is the normal cone of C at x:

$$N_C(x) = \{\eta \in \mathbb{R}^n : \langle \eta, y - x \rangle \leq 0, \forall y \in C\}.$$

For $x \notin C$, $\partial \iota_C(x) = \emptyset$.

We briefly review the calculus of subdifferentials. For any $\alpha > 0$, $\partial(\alpha f)(x) = \alpha \partial f(x)$. For any functions $f_1, f_2 : \mathbb{R}^n \to (-\infty, \infty]$ which are Fréchet subdifferentiable, the sum $f_1 + f_2$ is Fréchet subdifferentiable with subdifferential $\partial(f_1 + f_2)(x) \subseteq \partial f_1(x) + \partial f_2(x)$. If f_1 is differentiable, then $\partial(f_1 + f_2)(x) = \nabla f_1(x) + \partial f_2(x)$.

If $f \in \Gamma_0(\mathbb{R}^n)$, then $\partial f(x) = \{\eta \in \mathbb{R}^n : f(u) - f(x) - \langle \eta, u - x \rangle \geq 0, u \in \mathbb{R}^n\}$. In this case, the subdifferential operator ∂f is monotone, i.e. for any $u, v \in \mathbb{R}^n$

$$\langle \eta - \xi, u - v \rangle \geq 0$$

for every $\eta \in \partial f(u)$ and $\xi \in \partial f(v)$. For differentiable functions of a single variable, this is the familiar property that the derivative of a convex function is increasing. The gradient inequality for convex functions then becomes a subgradient inequality: for all $x, y \in \mathbb{R}^n$ and any $\eta \in \partial f(x)$

$$f(y) \geq f(x) + \langle \eta, y - x \rangle.$$

Similarly, f is σ-strongly convex if and only if

$$f(y) \geq f(x) + \langle \eta, y - x \rangle + \frac{\sigma}{2}\|y - x\|^2,$$

and f is ρ-semiconvex if and only if

$$f(y) \geq f(x) + \langle \eta, y - x \rangle - \frac{\rho}{2}\|y - x\|^2.$$

From these expressions, we can derive corresponding monotonicity properties.

2.3 The Fenchel Conjugate

Duality plays a role in many areas of mathematics, allowing us to approach problems from a new perspective. In convex analysis, this duality is given by conjugation. The Fenchel conjugate of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is defined by

$$f^*(u) = \sup_{x \in \mathbb{R}^n} \langle x, u \rangle - f(x).$$

If we view f as a collection of points $(x, f(x))$, then the conjugate f^* describes the epigraph of f through its supporting hyperplanes. This is closely related to the classical Legendre transform, though here we allow nonconvex f, and it is often referred to as simply the convex conjugate. For example, if $f = \|\cdot\|$, then $f^* = \iota_{B_1(0)}$, and if $f = \frac{1}{2}\|\cdot\|^2$, then $f^* = \frac{1}{2}\|\cdot\|^2$. In fact, $f = f^*$ if and only if $f = \frac{1}{2}\|\cdot\|^2$.

An immediate consequence of the definition is the Fenchel-Young Inequality:

$$f(x) + f^*(u) \geq \langle x, u \rangle. \tag{2.1}$$

If $f \in \Gamma_0(\mathbb{R}^n)$, the subdifferentials of f^* and f are related by

$$u \in \partial f(x) \iff x \in \partial f^*(u).$$

That is, if f is differentiable, we see that ∇f^* inverts ∇f. We note here that equality is achieved in (2.1) if and only if $u \in \partial f(x)$ or equivalently $x \in \partial f(u)$.

We note that if $f \in \Gamma_0(\mathbb{R}^n)$, then $f^* \in \Gamma_0(\mathbb{R}^n)$ as well. If $f \geq f(0) = 0$, then $f^* \geq f^*(0) = 0$ as well. If we define the biconjugate $f^{**}(u) = (f^*)^*(u)$, then we generally have that $f^{**} \leq f$. The Fenchel-Moreau Theorem states that if $f : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, then $f^{**} = f$ if and only if f is lower semicontinuous and convex.

If we consider the primal problem

$$\min\{f_1(x) + f_2(x) : x \in \mathbb{R}^n\}$$

where $f_1, f_2 \in \Gamma_0(\mathbb{R}^n)$, then the Fenchel dual problem is given by

$$\min\{f_1^*(y) + f_2^*(-y) : y \in \mathbb{R}^n\}.$$

Similarly, if we consider the primal problem

$$\min\{f_1(x) + f_2(Bx) : x \in \mathbb{R}^n\}$$

where $B \in \mathbb{R}^{m \times n}$, then the Fenchel Rockafellar dual problem is

$$\min\{f_1^*(B^\top y) + f_2^*(-y) : y \in \mathbb{R}^m\}.$$

The sum of the optimal values of a primal-dual pair is called the duality gap. While we will not discuss it here, the duality theorems due to Fenchel and others provide conditions under which the duality gap is zero (see, e.g., [5]).

2.4 The Moreau Envelope

Due to the fundamental roles the Moreau Envelope and the proximity operator play in this work, we devote this section to reviewing their definitions and properties, as well as providing some discussion and history. For a more thorough discussion centered around proximal point algorithms, we recommend [37].

First defined by J. J. Moreau in [36] , the Moreau envelope of $f \in \Gamma(\mathbb{R}^n)$ with parameter $\alpha > 0$ is defined by the infimal convolution of f with $\|\cdot\|^2$:

$$\operatorname{env}_\alpha f(x) \coloneqq \inf\left\{ f(u) + \frac{1}{2\alpha}\|u - x\|^2 : u \in \operatorname{dom}(f) \right\}.$$

We hereafter refer to the $\operatorname{env}_\alpha f$ as simply the envelope of f. The closely related proximity operator of f with parameter $\alpha > 0$ is defined by

$$\operatorname{prox}_{\alpha f}(x) \coloneqq \operatorname{Argmin}\left\{ f(u) + \frac{1}{2\alpha}\|u - x\|^2 : u \in \operatorname{dom}(f) \right\}.$$

This is also sometimes referred to as the proximal operator or proximal mapping of f. While these definitions do not require f to be convex, they are much more nicely behaved when it is. We therefore restrict our attention to $f \in \Gamma_0(\mathbb{R}^n)$ for the remainder of this section.

These two definitions have roots in monotone operator theory. The proximity operator is the resolvent of the subdifferential of f: $\operatorname{prox}_{\alpha f}(x) = (I + \alpha \partial f)^{-1}$. The envelope is sometimes

called the Moreau-Yosida envelope or Moreau-Yosida regularization of f based on the fact that $\nabla \operatorname{env}_\alpha f$ is the Yosida approximation to ∂f with parameter α.

The proximity operator was originally defined as a generalization of projection onto convex sets. In fact, if $f = \iota_C$ for some closed convex set C, then $\operatorname{prox}_{\alpha f}(x) = \Pi_C(x)$ for all $\alpha > 0$. In general, we view $\operatorname{prox}_{\alpha f}(x)$ as projecting x to a lower sublevel set of the function. Like projection, the proximity operator is firmly nonexpansive: for any $x, y \in \mathbb{R}^n$,

$$\|p - q\|^2 \leq \langle p - q, x - y \rangle$$

for all $p \in \operatorname{prox}_{\alpha f}(x)$ and $q \in \operatorname{prox}_{\alpha f}(y)$. It is easy to see that for $\bar{x} \in \operatorname{Argmin} f$, $\operatorname{prox}_{\alpha f}(\bar{x}) = f(\bar{x})$. In fact, the fixed points of the proximity operator are precisely the minimizers of f.

Another interpretation is that $\operatorname{prox}_{\alpha f}$ computes an implicit gradient descent step. As above, $p = \operatorname{prox}_{\alpha f}(x)$ if and only if $0 \in \partial f(p) + \frac{1}{\alpha}(p - x)$. That is,

$$p \in x - \alpha \partial f(p).$$

As an example, the proximity operator of the absolute value function is the soft-thresholding operator from signal processing: $\operatorname{prox}_{\alpha|\cdot|}(x) = \max\{0,\ \operatorname{sgn}(x)(|x| - \alpha)\}$. Note that for any nonzero x, $f(x) = |x|$ is differentiable and $f'(x) = \operatorname{sgn}(x)$. If $\operatorname{prox}_{\alpha f}(x) \neq 0$ as well, then it is precisely a gradient descent step with step size α: $p = x - \alpha f'(p) = x - \alpha f'(x)$.

It follows from the definition that $\operatorname{env}_\alpha f(x) \leq f(x)$ for every x and $\operatorname{env}_\alpha f$ approaches f as $\alpha \to 0$. For $f \in \Gamma_0(\mathbb{R}^n)$, the proximity operator is single-valued, and we see that

$$\operatorname{env}_\alpha f(x) = f(\operatorname{prox}_{\alpha f}(x)) + \frac{1}{2\alpha} \| \operatorname{prox}_{\alpha f}(x) - x \|^2.$$

By our previous discussion, we see that $\operatorname{Argmin} \operatorname{env}_\alpha f = \operatorname{Argmin} f$. A less obvious fact is that the gradient of $\operatorname{env}_\alpha f$ is $\frac{1}{\alpha}$-Lipschitz and is given by

$$\nabla \operatorname{env}_\alpha f = \frac{1}{\alpha}(\mathrm{Id} - \operatorname{prox}_{\alpha f}).$$

Thus, we view $\operatorname{env}_\alpha f$ as a smoothed approximation of f which preserves its minimizers.

We finish this section with some useful properties connecting the envelope and the conjugate. The conjugate of the envelope is given by $(\operatorname{env}_\alpha f)^* = f^* + \frac{\alpha}{2}\|\cdot\|^2$. The Moreau Identity is

$$\frac{1}{2\alpha}\|x\|^2 = \operatorname{env}_\alpha f(x) + \operatorname{env}_{\alpha^{-1}} f^*(x/\alpha).$$

Differentiating gives us

$$x = \operatorname{prox}_{\alpha f}(x) + \alpha \operatorname{prox}_{\alpha^{-1} f^*}(x/\alpha).$$

This is often also referred to as the Moreau Identity. Note that this allows us to write $\nabla \operatorname{env}_\alpha f(x) = \operatorname{prox}_{\alpha^{-1} f^*}(x/\alpha)$.

2.5 Semialgebraic and Subanalytic Functions

We now include some definitions from real analytic geometry which are becoming more prevalent in optimization research. These results are drawn largely from [18] and [7]. A set $A \subseteq \mathbb{R}^n$ is called semialgebraic if it can be defined by a Boolean combination of polynomial equalities and inequalities. That is, we can write

$$A = \bigcap_{j=1}^{m} \bigcup_{i=1}^{n} \{x \in \mathbb{R}^n : f_i(x) = 0,\ g_{ij}(x) > 0\}.$$

If f_i and g_{ij} are replaced by real analytic functions for all i, j above, the set A is called semianalytic. A set A is subanalytic if every $x \in A$ admits a neighborhood V such that $V \cap A$ is the projection of a bounded semianalytic subset in $\mathbb{R}^{n+1}$.

It is easy to verify that each definition is a strict generalization of the one before, giving us the following relationship:

$$\text{Subanalytic} \supset \text{Semianalytic} \supset \text{Semialgebraic}.$$

These are all special cases of sets definable on o-minimal structures (see [18] for an overview), which provides us with a useful way to determine whether a given set belongs to any of these classes. In the following discussion, we use the term definable as a catch-all, but the reader may substitute in semialgebraic, semianalytic, or subanalytic.

A first-order formula of the language of the o-minimal structure can be constructed by the following rules [18].

1. If $P(x_1, \ldots, x_n)$ is a polynomial, then $P(x_1, \ldots, x_n) = 0$ and $P(x_1, \ldots, x_n) > 0$ are first-order formulas.

2. If $A \subseteq \mathbb{R}^n$ is definable, then $x \in A$ is a first-order formula.

3. If Φ and Ψ are first-order formulas, then "Φ and Ψ", "Φ or Ψ", "not Φ, and $\Phi \Rightarrow \Psi$ are all first-order formulas.

4. If $\Phi(y, x)$ is a first-order formula and A is a definable subset, then $\exists x \in A \Phi(y, x)$ and $\forall x \in A \Phi(y, x)$ are first-order formulas.

If $\Phi(x)$ is a first order formula, the set $A = \{x \in \mathbb{R}^n : \Phi(x)\}$ is definable. For example, any algebraic set is definable.

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is definable if $\text{graph}(f) = \{(x, y) \in \mathbb{R}^{n+1} : y = f(x)\}$ is definable. We note that set of definable functions $f : \mathbb{R}^n \to \mathbb{R}$ form a $\mathbb{R}$-algebra. This allows us to create new definable functions from old ones through algebraic operations.

Following our previous discussion, the standard way to show that a function belongs to one of these categories is to show that its graph is definable in the appropriate language. For example, if A is a definable set, then $\text{dist}_A(x) = \inf\{\|x - a\| : a \in A\}$ is definable. Its graph is given by all $(x, r) \in \mathbb{R}^{n+1}$ satisfying the following first order formula:

$$r \geq 0 \text{ and } \forall a \in A(r^2 \leq \|x - a\|^2) \text{ and } \forall \epsilon \in \mathbb{R}, \epsilon > 0 \Rightarrow \exists a \in A(t^2 + \epsilon > \|x - a\|^2).$$

By the same argument, if A is a semialgebraic or subanalytic set, then dist_A is a semialgebraic or subanalytic function respectively.

Chapter 3

Sparsity Promoting Functions

In this chapter, we define the class of sparsity promoting functions and introduce a simple method for constructing new structured sparsity promoting functions from convex sparsity promoting functions. A comprehensive study of their thresholding behavior is given in Section 3.2. We explore other properties which may be of interest in Section 3.3. In particular, we show that our functions satisfy the nonsmooth Łojasiewicz inequality near the origin and that the Fenchel conjugate can be partially or wholly computed based on knowledge of f. Results which appear in our paper [47] include this citation.

3.1 Definition

Roughly speaking, a function is sparsity promoting if it penalizes nonzero entries. The most natural sparsity promoting function is the ℓ_0-norm, which simply counts the number of nonzero entries, though this is often relaxed to the ℓ_1-norm. The construction of sparsity promoting penalty functions is an active area of research which spans many applications. Some of the proposed functions include the convex elastic net [54], the nonconvex SCAD

[22], and of course the semiconvex MCP [53]. What all of these functions have in common is that they send zero to zero and are, in some sense, sharp at the origin. Based on this observation, we propose the following definition.

Definition 3.1.1 ([47]). *Let $f \in \Gamma(\mathbb{R}^n)$. Then f is said to be a sparsity promoting function provided that (i) $f(0) = 0$ and f achieves its global minimum at the origin; and (ii) the set $\partial f(0)$ contains at least one nonzero element.*

Item (i) ensures that any sparsity already present is preserved, and Item (ii) captures the notion of "sharpness" at the origin. While we are hardly the first to make note of these properties, to the best of our knowledge, this formalization is novel. Our definition has the added benefits of being easy to verify, and, as we will see, immediately implies many other properties.

It is perhaps more intuitive to describe sparsity promoting behavior in terms of the proximity operator. The subdifferential and proximity operator of a function $f \in \Gamma_0(\mathbb{R}^n)$ satisfy the following relationship (see, e.g., [5]): for any $\alpha > 0$

$$x \in \alpha \partial f(y) \iff y = \operatorname{prox}_{\alpha f}(x + y). \tag{3.1}$$

From this relationship, we get the following characterization of convex sparsity promoting functions.

Lemma 3.1.1 ([47]). *Let $f \in \Gamma_0(\mathbb{R}^n)$ be a sparsity promoting function and let $\alpha > 0$. Then the following statements hold.*

(i) If $x \in \partial \alpha f(0)$, then $\operatorname{prox}_{\alpha f}(x) = 0$.

(ii) For all $x \in \operatorname{dom}(f)$, $\| \operatorname{prox}_{\alpha f}(x)\| \leq \|x\|$.

Proof. (i): This is a direction consequence of (3.1).

(ii): By Item (i), $0 \in \alpha\partial f(0)$ implies $\operatorname{prox}_{\alpha f}(0) = 0$. Because $\operatorname{prox}_{\alpha f}$ is a nonexpansive operator, we have $\|\operatorname{prox}_{\alpha f}(x)\| = \|\operatorname{prox}_{\alpha f}(x) - \operatorname{prox}_{\alpha f}(0)\| \leq \|x\|$ for all $x \in \operatorname{dom}(f)$. □

The proximity operator of a convex sparsity promoting function therefore shrinks all entries towards zero and sends all entries below a certain threshold to zero. This behavior was described by Tibshirani for the ℓ_1 penalty: Least Absolute Shrinkage and Selection Operator (LASSO).

3.2 Structured Sparsity Promoting Functions

We now introduce a simple construction of new sparsity promoting penalties. For any $f \in \Gamma_0(\mathbb{R}^n)$ and any positive number $\alpha > 0$, we define

$$f_\alpha(x) \coloneqq f(x) - \operatorname{env}_\alpha f(x). \tag{$\mathcal{F}_\alpha$}$$

Recall that when f is the absolute value function, f_α is the MCP function. Several other examples are provided in Chapter 4.

Sparsity promotion depends entirely on the behavior of a function and its subdifferential at the origin. Since the Moreau envelope of any function in $\Gamma_0(\mathbb{R}^n)$ is differentiable (see Section 2.4), the subdifferentials of f_α and f are related as follows (see [44]):

$$\partial f_\alpha(x) = \partial f(x) - \nabla \operatorname{env}_\alpha f(x). \tag{3.2}$$

Due to this inherent relationship between ∂f_α and ∂f we see immediately that f_α must be sparsity promoting if f is.

Theorem 3.2.1 ([47]). *Let $f \in \Gamma_0(\mathbb{R}^n)$ be a sparsity promoting function. For any $\alpha > 0$, the function f_α defined by $(\mathcal{F}_\alpha)$ is a sparsity promoting function. Moreover, $\partial f_\alpha(0) = \partial f(0)$.*

Proof. As a direct consequence of the definition of the Moreau envelope, $0 \leq \operatorname{env}_\alpha f(x) \leq f(x)$ for all $x \in \mathbb{R}^n$, hence $f_\alpha(x) \geq 0$ for all $x \in \operatorname{dom}(f)$. Since f is a sparsity promoting function, we have $f_\alpha(0) = f(0) - \operatorname{env}_\alpha f(0) = 0$. Therefore $\min_{x\in\mathbb{R}^n} f_\alpha(x) = f_\alpha(0) = 0$. On the other hand, from (3.2) and the relation $\nabla \operatorname{env}_\alpha f(x) = \frac{1}{\alpha}(x - \operatorname{prox}_{\alpha f}(x))$, we get $\partial f_\alpha(0) = \partial f(0)$, which contains at least one nonzero element by assumption. Therefore, f_α is sparsity promoting. □

Remark 3.2.1. We note that f_α does not approximate the function f but does inherit properties from it. Because sparsity promotion is a property centered around behavior at the origin, it only provides information about f_α near the origin. However, given global information about f, we are able to determine global properties of f_α. For example, if f is L-Lipschitz, it is straightforward to show that $0 \leq f_\alpha(x) \leq L^2\alpha$ for all $x \in \mathbb{R}^n$.

As an immediate consequence, we see that the sparsity promoting property is preserved under reflection. This fact can be used to expedite proofs for functions which are symmetric in some sense (see Chapter 4).

Lemma 3.2.1. *Let $f \in \Gamma_0(\mathbb{R}^n)$ be a sparsity promoting function, and let $g : x \mapsto f(-x)$. Then both g and g_α are sparsity promoting. Moreover, $g_\alpha = f_\alpha(-\cdot)$ and $\partial g_\alpha(0) = -\partial f(0)$.*

Proof. Since $g(0) = f(0) = \min_{x\in\mathbb{R}^n} f(x) = \min_{x\in\mathbb{R}^n} g(x)$ and $\partial g(0) = -\partial f(0)$, so g is sparsity promoting. By Theorem 3.2.1, g_α is sparsity promoting and $\partial g_\alpha(0) = -\partial f(0)$. By the definition of the Moreau Envelope, $\operatorname{env}_\alpha g(x) = \operatorname{env}_\alpha f(-x)$, which gives us $g_\alpha(x) = f_\alpha(-x)$. □

Theorem 3.2.1 tells us not only that f_α is sparsity promoting but that it preserves the structure of f around the origin. As we will see, the inherent relationship between f and f_α allows us to impose structure on f_α through assumptions on f. The first of these is that the convexity of f controls the nonconvexity of f_α. We remind the reader of two definitions. For $\sigma > 0$ a function $g \in \Gamma(\mathbb{R}^n)$ is σ-strongly convex if and only if the function $g - \frac{\sigma}{2}\|\cdot\|^2$ is convex. For $\rho > 0$, a function $g \in \Gamma(\mathbb{R}^n)$ is ρ-semiconvex if $g + \frac{\rho}{2}\|\cdot\|^2$ is convex.

Proposition 3.2.1 ([47]). *Let $f \in \Gamma_0(\mathbb{R}^n)$. Then f_α, defined by $(\mathcal{F}_\alpha)$, is $\frac{1}{\alpha}$-semiconvex. If f is μ-strongly convex, then f_α is $(\mu - \frac{1}{\alpha})$-strongly convex if $\mu > \frac{1}{\alpha}$, convex if $\mu = \alpha$, and $(\frac{1}{\alpha} - \mu)$-semiconvex if $\mu < \frac{1}{\alpha}$.*

Proof. Write

$$f_\alpha = f - \operatorname{env}_\alpha f = f + (-\operatorname{env}_\alpha f + \frac{1}{2\alpha}\|\cdot\|^2) - \frac{1}{2\alpha}\|\cdot\|^2.$$

For all $x \in \mathbb{R}^n$ we have that

$$f_\alpha(x) = f(x) + \big(f + \frac{1}{2\alpha}\|\cdot\|^2\big)^*(\alpha^{-1}x) - \frac{1}{2\alpha}\|x\|^2,$$

which implies that f_α is $\frac{1}{\alpha}$-semiconvex (See Section 2.4).

In addition, if f is μ-strongly convex, then there exists a convex function g such that $f = g + \frac{\mu}{2}\|\cdot\|^2$. Replacing $f(x)$ in the previous equation, we get

$$f_\alpha(x) = g(x) + \big(f + \frac{1}{2\alpha}\|\cdot\|^2\big)^*(\alpha^{-1}x) + \frac{1}{2}(\mu - \frac{1}{\alpha})\|x\|^2.$$

The result follows. □

As an easy corollary, we can specify the convexity (or semiconvexity) of the sum of f_α and a quadratic term.

Corollary 3.2.1. *Let $f \in \Gamma_0(\mathbb{R}^n)$, and let f_α be defined by $(\mathcal{F}_\alpha)$. For any given $x \in \mathbb{R}^n$ and positive parameters α and β, we define*

$$F(u) = f_\alpha(u) + \frac{1}{2\beta}\|u - x\|^2, \tag{3.3}$$

where $u \in \mathbb{R}^n$. Then, F is $(\beta^{-1} - \alpha^{-1})$-strongly convex if $\beta < \alpha$, convex if $\beta = \alpha$, and $(\alpha^{-1} - \beta^{-1})$-semiconvex if $\beta > \alpha$. If, in addition, f is μ-strongly convex, then F is $(\mu - \alpha^{-1} + \beta)$-strongly convex, if $\mu > \alpha^{-1} - \beta^{-1}$, convex if $\mu = \alpha^{-1} - \beta^{-1}$, and $(\alpha^{-1} - \beta^{-1} - \mu)$-semiconvex if $\mu < \alpha^{-1} - \beta^{-1}$.

The next two results extend these properties to compositions with linear operators.

Proposition 3.2.2. *If $f \in \Gamma_0(\mathbb{R}^n)$ and $D \in \mathbb{R}^{n\times m}$ such that $\operatorname{im} D \cap \operatorname{dom}(f) \neq \emptyset$, then $f_\alpha \circ D$ is $\frac{\|D\|^2}{\alpha}$-semiconvex.*

Proof. We first show that $\nabla(\operatorname{env}_\alpha f \circ D)$ is $\frac{\|D\|^2}{\alpha}$-Lipschitz:

$$\begin{aligned}\|\nabla(\operatorname{env}_\alpha f \circ D)(x) - \nabla(\operatorname{env}_\alpha f \circ D)(y)\| &= \|D^\top(\nabla \operatorname{env}_\alpha f(Dx) - \nabla \operatorname{env}_\alpha f(Dy))\| \\ &\leq \frac{1}{\alpha}\|D^\top D(x - y)\| \leq \frac{\|D\|^2}{\alpha}\|x - y\|,\end{aligned}$$

where the second line follows from the fact that $\nabla \operatorname{env}_\alpha f$ is $\frac{1}{\alpha}$-Lipschitz. Now by the monotonicity of $\partial(f \circ D)$,

$$\langle \partial(f_\alpha \circ D)(x) - \partial(f_\alpha \circ D)(y), x - y\rangle \geq -\frac{\|D\|^2}{\alpha}\|x - y\|^2$$

□

Corollary 3.2.2. *Given $x \in \mathbb{R}^n$, then the function $f_\alpha(D\cdot) + \frac{1}{2\lambda}\|\cdot - x\|^2$ is strictly convex if $\lambda < \frac{\alpha}{\|D\|^2}$, convex if $\lambda = \frac{\alpha}{\|D\|^2}$, and semiconvex if $\lambda > \frac{\alpha}{\|D\|^2}$.*

With these first results, we are able to generalize our characterization Lemma 3.1.1 to the functions f_α. Roughly, we see that $\text{prox}_{\beta f_\alpha}$ sends entries in $x \in \min\{\alpha, \beta\} \cdot \partial f(0)$ to zero. Theorem 3.2.2 refines this result and specifies the thresholding behavior of the proximity operator. Recall that for convex functions $\text{prox}_{\alpha f}$ forces all entries towards zero, which is actually undesirable in applications. Item (i) below tells us that $\text{prox}_{\beta f_\alpha}(x)$ will be relatively close to x, thus reducing the bias of solutions.

Lemma 3.2.2 ([47])**.** *Let $f \in \Gamma_0(\mathbb{R}^n)$ be sparsity promoting and f_α as defined in $(\mathcal{F}_\alpha)$. For any $\alpha, \beta > 0$, the following statements hold.*

(i) For any $x \in \text{dom}(f)$, $\text{prox}_{\beta f_\alpha}(x) \subseteq \overline{B_{\|x\|}}(x)$.

(ii) If $x \in \min\{\alpha, \beta\} \cdot \partial f(0)$, then $0 \in \text{prox}_{\beta f_\alpha}(x)$.

Proof. For a fixed $x \in \mathbb{R}^n$, define F as in (3.3), so that $\text{prox}_{\beta f_\alpha}(x) = \text{argmin}_{u \in \mathbb{R}^n} F(u)$.

(i): Since $F(0) = \frac{1}{2\beta}\|x\|^2$ and $0 \in \overline{B_{\|x\|}}(x)$, to show $\text{prox}_{\beta f_\alpha}(x) \subseteq \overline{B_{\|x\|}}(x)$, we only need to show that for all $u \in \mathbb{R}^n \backslash \overline{B_{\|x\|}}(x)$, $F(u) > F(0)$. Actually, if $u \in \mathbb{R}^n \backslash \overline{B_{\|x\|}}(x)$, then $\|u - x\|^2 > \|x\|^2$. Since f_α is non-negative, it follows from (3.3) that $F(u) > \frac{1}{2\beta}\|x\|^2 = F(0)$. Thus the conclusion of Item (i) holds.

(ii): To prove Item (ii), from Item (i) and $F(0) = \frac{1}{2\beta}\|x\|^2$, it suffices to show $F(u) \geq \frac{1}{2\beta}\|x\|^2$ for all $u \in \overline{B_{\|x\|}}(x)$. From the assumption of $x \in \min\{\alpha, \beta\} \cdot \partial f(0)$, we have that for all $u \in \mathbb{R}^n$, $f(u) \geq \frac{1}{\min\{\alpha,\beta\}}\langle x, u\rangle$. Since $f(0) = 0$, we have $\text{env}_\alpha f(u) \leq \frac{1}{2\alpha}\|u\|^2$ for all $u \in \mathbb{R}^n$. Hence

$$f_\alpha(u) \geq \frac{1}{\min\{\alpha, \beta\}}\langle x, u\rangle - \frac{1}{2\alpha}\|u\|^2.$$

Therefore,

$$F(u) \geq \frac{1}{\min\{\alpha,\beta\}}\langle x,u\rangle - \frac{1}{2\alpha}\|u\|^2 + \frac{1}{2\beta}\|u-x\|^2$$
$$= \begin{cases} \left(\frac{1}{2\beta}-\frac{1}{2\alpha}\right)\|u\|^2 + \frac{1}{2\beta}\|x\|^2, & \text{if } \beta \leq \alpha, \\ \left(\frac{1}{2\alpha}-\frac{1}{2\beta}\right)(\|x\|^2 - \|u-x\|^2) + \frac{1}{2\beta}\|x\|^2, & \text{if } \alpha < \beta. \end{cases}$$

So, $F(u) \geq \frac{1}{2\beta}\|x\|^2 = F(0)$ holds for all $u \in \overline{B_{\|x\|}}(x)$. This completes the proof of the lemma. □

Remark 3.2.2. From item (i) of Lemma 3.2.2 we see for $x \in \mathbb{R}$, $\operatorname{sgn}(x) = \operatorname{sgn}(p)$ if $p \in \operatorname{prox}_{\beta f_\alpha}(x)$ and both x and p are simultaneously nonzero. We note that this is also true for $\operatorname{prox}_{\alpha f}(x)$.

Before we can prove our main thresholding result, we need the following technical lemma. Recall that for convex sets containing the origin, $x \in A$ implies that for some $\lambda > 1$, $\lambda x \in A$ (see 2.1). While the statement of the lemma may appear strange at first glance, the conditions therein arise naturally when computing the proximity operator.

Lemma 3.2.3 ([47])**.** *Let $f \in \Gamma_0(\mathbb{R}^n)$ be a sparsity promoting function and $w \in \operatorname{dom}(\partial f)$. If $w \in \partial f(0)$ and there exists a nonzero $\xi \in \operatorname{ri}(\partial f(0)) \cap \partial f(w)$, then $w = 0$.*

Proof. Assume that $w \neq 0$. First, since $w \in \partial f(0)$ and $f(0) = 0$, we have $f(w) \geq \|w\|^2 > 0$.

Second, since $\xi \in \partial f(0)) \cap \partial f(w)$, then $\xi \in \partial f(0)$ implies $f(0) + f^*(\xi) = \langle 0, \xi\rangle$ while $\xi \in \partial f(w)$ implies $f(w) + f^*(\xi) = \langle \xi, w\rangle$. Hence,

$$f(w) = \langle \xi, w\rangle. \tag{3.4}$$

By the monotonicity of ∂f, for any $\eta \in \partial f(0)$, $\langle \xi - \eta, w \rangle \geq 0$. Together with (3.4) we get

$$f(w) \geq \langle \eta, w \rangle. \tag{3.5}$$

Finally, since $\xi \in \mathrm{ri}(\partial f(0))$ and $\partial f(0)$ is convex, there exists $\lambda > 1$ such that $\lambda\xi \in \partial f(0)$. By (3.4) and (3.5), we get

$$f(w) \geq \langle \lambda\xi, w \rangle = \lambda f(w),$$

which implies $f(w) \leq 0$. This is a contradiction, so $w = 0$. □

The following theorem provides a more exact guarantee of thresholding behavior. Because, in general, the function f_α may be quite different from f, we can only describe the proximity operator for x sufficiently close to the origin.

Theorem 3.2.2 ([47]). *Let $f \in \Gamma_0(\mathbb{R}^n)$ be a sparsity promoting function. For any $x \in \mathrm{dom}(f)$, the following statements hold:*

(i) *If $\beta < \alpha$, then $\mathrm{prox}_{\beta f_\alpha}(x) = 0$ for $x \in \beta\partial f(0)$;*

(ii) *If $\beta = \alpha$, then $\mathrm{prox}_{\beta f_\alpha}(x) = 0$ for $x \in \mathrm{ri}(\alpha\partial f(0))$;*

(iii) *If $\beta > \alpha$, then $\mathrm{prox}_{\beta f_\alpha}(x) = 0$ for $x \in \alpha\partial f(0)$.*

Proof. Given $x \in \mathbb{R}^n$, define F as in (3.3).

(i) We first consider the situation $\beta < \alpha$. From Corollary 3.2.1, we know that F is $\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$-strongly convex and therefore has a unique minimizer. By Lemma 3.2.3, $x \in \beta\partial f(0)$ implies that $0 = \mathrm{argmin}_{u \in \mathbb{R}^n} F(u)$. Together these imply that $\mathrm{prox}_{\beta f_\alpha}(x) = 0$.

(ii) Next we consider $\alpha = \beta$. From Corollary 3.2.1, $F(u)$ is convex but not strongly, and the minimizer may no longer be unique. By Lemma 3.2.3, $0 \in \operatorname{prox}_{\beta f_\alpha}(x)$ for $x \in \alpha\partial f(0)$.

Now suppose $x \in \operatorname{ri}(\alpha\partial f(0))$ and let w^* be an element of $\operatorname{prox}_{\beta f_\alpha}(x)$. To show that $w^* = 0$, by identifying αf, x, and w^*, respectively, as f, ξ, and w in Lemma 3.2.2, it suffices to show that $x \in \partial(\alpha f)(w^*)$ and $w^* \in \partial(\alpha f)(0)$. By Fermat's rule, $w^* \in \operatorname{prox}_{\beta f_\alpha}(x)$ implies that $0 \in \partial f_\alpha(w^*) + \frac{1}{\beta}(w^* - x)$. As we saw earlier that $\partial f_\alpha(w^*) = \partial f(w^*) - \nabla \operatorname{env}_\alpha f(w^*)$ and $\nabla \operatorname{env}_\alpha f(w^*) = \frac{1}{\alpha}(w^* - \operatorname{prox}_{\alpha f}(w^*))$, this can be rewritten as

$$\frac{1}{\beta}x + \left(\frac{1}{\alpha} - \frac{1}{\beta}\right) w^* - \frac{1}{\alpha}\operatorname{prox}_{\alpha f}(w^*) \in \partial f(w^*). \tag{3.6}$$

From (3.6), we get $x - \operatorname{prox}_{\alpha f}(w^*) \in \partial(\alpha f)(w^*)$. Therefore, the conditions $x \in \partial(\alpha f)(w^*)$ and $w^* \in \partial(\alpha f)(0)$ hold if and only if $\operatorname{prox}_{\alpha f}(w^*) = 0$.

Since $x \in \partial(\alpha f)(0)$, by the monotonicity of ∂f we have

$$\langle x - \operatorname{prox}_{\alpha f}(w^*) - x, w^* \rangle \geq 0.$$

That is, $\langle \operatorname{prox}_{\alpha f}(w^*), w^* \rangle \leq 0$. But due to the nonexpansiveness of $\operatorname{prox}_{\alpha f}$ and the fact that $\operatorname{prox}_{\alpha f}(0) = 0$,

$$\langle \operatorname{prox}_{\alpha f}(w^*), w^* \rangle \geq \| \operatorname{prox}_{\alpha f}(w^*) \|^2.$$

This implies that $\operatorname{prox}_{\alpha f}(w^*) = 0$. Thus by Lemma 3.2.3, $w^* = 0$.

(iii) Finally, we consider the situation of $\beta > \alpha$. In this case, we assume that $0 \neq x \in \alpha\partial f(0)$. From Lemma 3.2.3, we know that $0 \in \operatorname{prox}_{\beta f_\alpha}(x)$. We further show that the point 0 is the only element in $\operatorname{prox}_{\beta f_\alpha}(x)$.

Recall from the proof of Lemma 3.2.3 that when $\beta > \alpha$,

$$F(u) \geq \left(\frac{1}{2\alpha} - \frac{1}{2\beta}\right)(\|x\|^2 - \|u-x\|^2) + \frac{1}{2\beta}\|x\|^2 \geq \frac{1}{2\beta}\|x\|^2.$$

Actually, if $w^* \in \operatorname{prox}_{\beta f_\alpha}(x)$, then w^* must be on the boundary of $\overline{B_{\|x\|}}(x)$ and $F(w^*) = f_\alpha(w^*) + \frac{1}{2\beta}\|w^* - x\|^2 = \frac{1}{2\beta}\|x\|^2$. Thus, $f_\alpha(w^*) = 0$, that is, $f(w^*) = \operatorname{env}_\alpha f(w^*)$. We also know that $f(w^*) \geq \frac{1}{\alpha}\langle x, w^*\rangle$ and $\operatorname{env}_\alpha f(w^*) \leq \frac{1}{2\alpha}\|w^*\|^2$. Therefore, because $2\langle x, w^*\rangle = \|w^*\|^2$, we get

$$\operatorname{env}_\alpha f(w^*) = \frac{1}{2\alpha}\|w^*\|^2,$$

which implies that $0 = \operatorname{prox}_{\alpha f}(w^*)$. On the other hand, the identity $f(w^*) = \operatorname{env}_\alpha f(w^*)$ indicates $w^* = \operatorname{prox}_{\alpha f}(w^*)$. Therefore, $w^* = 0$. This completes the proof. □

Remark 3.2.3. Item (iii) of the theorem is not tight. In fact in every example, when $\beta > \alpha$, $\operatorname{prox}_{\beta f_\alpha}(x) = 0$ for all x in a set strictly larger than $\alpha\partial f(0)$. However, the exact form of this set depends entirely on the function in question.

3.3 Further Properties

We have seen that the convexity or semiconvexity of a sparsity promoting function provides information about the convexity of a particular problem model as well as the thresholding behavior of the proximity operator. We now turn our attention to other structural properties of practical interest to optimizers. First, we describe the conjugate behavior of sparsity promoting functions and show when the Fenchel dual problem will be differentiable. Then we discuss the relationship between sparsity promotion and other notions of sharpness and determine the Łojasiewicz exponent of our functions f_α.

3.3.1 Conjugation Results

Without further ado, we consider conjugates of sparsity promoting functions. For further discussion of the conjugate and its role in optimization, we refer to Section 2.3. Recall that for any $f : \mathbb{R}^n \to (-\infty, +\infty]$, the Fenchel conjugate is defined by

$$f^*(u) = \sup\{\langle x, u\rangle - f(x) : x \in \mathbb{R}^n\}.$$

An immediate consequence of this definition is that if $g : \mathbb{R}^n \to (-\infty, +\infty]$ such that $g \leq f$, then $f^* \leq g^*$. Another consequence is the Fenchel Young inequality: for all $x \in \operatorname{dom}(f)$ and $x^* \in \operatorname{dom}(f^*)$,

$$f(x) + f^*(x^*) \geq \langle x^*, x\rangle. \tag{3.7}$$

It follows that if $f \in \Gamma(\mathbb{R}^n)$ is sparsity promoting, then for all x, $f^*(x) \geq f^*(0) = 0$. Recall also that for $f \in \Gamma_0(\mathbb{R}^n)$, $x^* \in \partial f(x)$ if and only if $x \in \partial f^*(x^*)$, and equality is achieved in (3.7) if and only if $x^* \in \partial f(x)$.

Theorem 3.3.1. *Suppose $f \in \Gamma_0(\mathbb{R}^n)$ is sparsity promoting. Then*

(i) $x^* \in \partial f(0)$ if and only if $f^*(x^*) = 0$, and

(ii) f^* is sparsity promoting if and only if there exists nonzero $\bar{x} \in \operatorname{argmin} f$.

Proof. (i) Set $x = 0$ in (3.7). The result follows immediately from the fact that $f(0) = 0$.

(ii) Because $f(0) = 0$ is the global minimum of f, we have $f^*(0) = 0$ as well. From the convexity of f, we have

$$\bar{x} \in \operatorname{argmin} f \iff 0 \in \partial f(\bar{x}) \iff \bar{x} \in \partial f^*(0).$$

Thus, f^* is sparsity promoting if and only if there is a nonzero element $\bar{x} \in \operatorname{argmin} f$. $\square$

While we cannot directly extend this result to f_α, we once again see that f_α preserves the behavior of f near the origin. First, we collect some relevant properties.

Lemma 3.3.1 ([5, Chapter 13])**.** *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ and let $\alpha > 0$. Then the following hold.*

(i) $(\operatorname{env}_\alpha f)^* = f^* + \frac{\alpha}{2}\|\cdot\|^2$.

(ii) $(\alpha f)^* = \alpha f^*(\cdot/\alpha)$.

(iii) For all $v, y \in \mathbb{R}^n$, $(f(\cdot - y) + \langle \cdot, v\rangle_\alpha)^* = f^*(\cdot - v) + \langle y, \cdot\rangle - \langle y, v\rangle - \alpha$.

(iv) Let $q = \frac{1}{2}\|\cdot\|^2$. If f is proper, then $(\alpha f - q)^* = \alpha(\alpha q - f^*)^* - q$.

Lemma 3.3.2. *Suppose $f \in \Gamma_0(\mathbb{R}^n)$ is sparsity promoting, and f_α is defined by $(\mathcal{F}_\alpha)$. If we consider the restriction of f_α to $\partial f(0)$, $\tilde{f}_\alpha \coloneqq f_\alpha + \iota_{\partial f(0)}$, then $(\tilde{f}_\alpha)^*(x) = 0$ for all $x \in \partial f(0)$.*

Proof. Theorem 3.2.2 implies that for $x \in \alpha\partial f(0)$, $\operatorname{env}_\alpha f(x) = \frac{1}{2\alpha}\|x\|^2$. Therefore $\tilde{f}_\alpha = f - \frac{1}{2\alpha}\|\cdot\|^2$ and $\alpha\tilde{f}_\alpha = \alpha f - \frac{1}{2}\|\cdot\|^2$. By Lemma 3.3.1 (ii) and (iv), we have that

$$\alpha f^*(x/\alpha) = (\frac{1}{2}\|\cdot\|^2 - \alpha f^*(\cdot/\alpha))^*(x) - \frac{1}{2}\|x\|^2,$$

for $x \in \alpha\partial f(0)$. Now by Theorem 3.3.1, we see that $(\tilde{f}_\alpha)^*(x) = 0$ for all $x \in \partial f(0)$. $\square$

Remark 3.3.1. Note that $f_\alpha^*(x) \geq \tilde{f}_\alpha^*(x)$ for all $x \in \mathbb{R}^n$. In fact, if the supremum is achieved outside of $\partial f(0)$, we see that $f_\alpha^*(x)$ may be infinite. Proposition 3.3.1 is an example of this.

The following theorem shows how the conjugates of f_α, its parent function f, and the convex function $f + \frac{1}{2\alpha}\|\cdot\|^2$ relate to each other. We also extend Item (ii) of Theorem 3.3.1.

Theorem 3.3.2. *Suppose $f \in \Gamma_0(\mathbb{R}^n)$ is sparsity promoting, and f_α is defined by $(\mathcal{F}_\alpha)$. Then the following statements hold.*

(i) For all $x \in \mathrm{dom}(f)$,

$$(f_\alpha + \frac{1}{2\alpha}\|\cdot\|^2)^*(x) \leq f^*(x) \leq f_\alpha^*(x). \tag{3.8}$$

(ii) If f^* is sparsity promoting, then f_α^* is as well. Equivalently, if $f(\bar{x}) = 0$ for some $\bar{x} \neq 0$, then f_α^* is sparsity promoting.

Proof. (i) Because $\mathrm{env}_\alpha f(x) \leq \frac{1}{2\alpha}\|x\|^2$ for all $x \in \mathbb{R}^n$ and by the definition of f_α, we have $f_\alpha(x) \leq f(x) \leq (f_\alpha + \frac{1}{2\alpha}\|\cdot\|^2)(x)$. Conjugation reverses the inequalities.

(ii) We have already seen that $f_\alpha^*(x) \geq f_\alpha^*(0) = 0$ for all x. It remains to show that there is a nonzero element $x^* \in \partial f_\alpha^*(0)$. By item (i), we see that $\eta \in \partial f^*(0)$ implies $\eta \in \partial f_\alpha^*(0)$:

$$\liminf_{y\to 0} \frac{f_\alpha^*(y) - \langle \eta, y\rangle}{\|y\|} \geq \liminf_{y\to 0} \frac{f^*(y) - \langle \eta, y\rangle}{\|y\|} \geq 0.$$

Therefore, f_α^* is sparsity promoting if f^* is. □

Proposition 3.3.1. *Suppose $f \in \Gamma_0(\mathbb{R}^n)$ with $\mathrm{dom}(f) = \mathbb{R}^n$ and f_α is defined as in $(\mathcal{F}_\alpha)$. If f is L-Lipschitz, then*

$$f_\alpha^*(x) = \begin{cases} 0, & \text{if } x = 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.9}$$

Proof. If f is L-Lipschitz, then $0 \leq f(x) - \mathrm{env}_\alpha f(x) \leq L^2\alpha$ for all x, so it follows that $f_\alpha^* \geq (L^2\alpha)^*$. By direct computation, we see that $(L^2\alpha)^*$ is $-L^2\alpha$ if $x = 0$ and $+\infty$ otherwise. □

For example, the absolute value function $f(x) = |x|$ is 1-Lipschitz. Recall that the function $f_\alpha(x) = |x| - \frac{1}{2\alpha}x^2$ for $|x| \leq \alpha$ and $\frac{\alpha}{2}$ otherwise. We can compute the conjugate

f_α^* directly by taking the maximum of the $s_1(x) = \sup\{ux - |u| + \frac{1}{2\alpha}u^2 : |u| \leq \alpha\}$ and $s_2(x) = \sup\{ux - \frac{\alpha}{2} : |u| > \alpha\}$. If $x \neq 0$, then $s_2(x) = +\infty$, so we see that $f_\alpha^*(x) = +\infty$. If $x = 0$, we see that $-|u| + \frac{1}{2\alpha}u^2 < 0$ for all $|u| \leq \alpha$, so $s_1(0) = 0$. Clearly, $s_2(0) = -\frac{\alpha}{2}$. Therefore $f_\alpha^*(0) = 0$.

Finally, we provide conjugation results for the f-penalized least squares model. As we will see when considering the difference of convex model (Section 5.2), this provides smoothness guarantees for the corresponding dual problem.

Proposition 3.3.2. *Let $D \in \mathbb{R}^{n\times m}$ and define $G(x) = f(Dx) + \frac{1}{2\lambda}\|x - z\|^2$. Then the following hold.*

(i) $G^*(x) = \mathrm{env}_\lambda(f \circ D)^*(x - \frac{1}{\lambda}z) - \frac{1}{2\lambda}\|z\|^2$, and

(ii) G^* is differentiable with derivative $\nabla G^*(x) = \mathrm{prox}_{\lambda^{-1}(f\circ D)}(\lambda x - z)$.

Proof. (i) We expand G as follows

$$G(x) = f(Dx) + \frac{1}{2\lambda}\|x\|^2 - \frac{1}{\lambda}\langle x, z\rangle + \frac{1}{2\lambda}\|z\|^2.$$

Applying Lemma 3.3.1 (i) and (iii),

$$\begin{aligned} G^*(x) &= (f(D\,\cdot) + \frac{1}{2\lambda}\|\cdot\|^2)^*(x - \frac{1}{\lambda}z) - \frac{1}{2\lambda}\|z\|^2 \\ &= \mathrm{env}_\lambda(f \circ D)^*(x - \frac{1}{\lambda}z) - \frac{1}{2\lambda}\|z\|^2. \end{aligned}$$

(ii) We simply differentiate the result from part (i) and apply the Moreau Identity.

□

3.3.2 Sharpness and the Łojasiewicz Inequality

Most optimization methods are based on some variation of gradient descent, but even convexity is not enough to guarantee fast convergence. If the objective is too "flat", gradient curves may have infinite length [9]. The Łojasiewicz inequality for real analytic functions $f : \mathbb{R}^n \to \mathbb{R}$ states that if $\bar{x} \in \operatorname{crit} f$, then there exists $\theta \in [0, 1)$ such that

$$\frac{|f - f(\bar{x})|^\theta}{\|\nabla f\|} \tag{3.10}$$

is bounded around $\bar{x}$. The boundedness of this function tells us that f approaches $f(\bar{x})$ faster than $\|\nabla f\|$ approaches zero, or, in other words, f is sufficiently steep near $\bar{x}$. For such functions, every bounded gradient trajectory converges to a critical point [34], and this allows us to estimate convergence rates for many common descent methods [1].

To extend this to nonsmooth functions, we must generalize the norm of the gradient. For $f \in \Gamma(\mathbb{R}^n)$, we define the nonsmooth slope of f at x as

$$m_f(x) \coloneqq \inf\{\|\eta\| : \eta \in \partial f(x)\}.$$

Then (3.10) becomes

$$\frac{|f - f(\bar{x})|^\theta}{m_f} \tag{3.11}$$

and we say f satisfies the nonsmooth Łojasiewicz inequality at $\bar{x}$ if this function remains bounded near $\bar{x}$. Naturally, if a function is "sharp" enough around the minimizer, (3.11) will be bounded. This is made rigorous by the following definition.

Definition 3.3.1 ([8])**.** *A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is sharp on $S \subseteq \mathbb{R}^n$ if there exists $c > 0$ such that for all $x \in S$, $m_f(x) \geq c$.*

The "sharpness" of sparsity promoting functions is measured by the existence of a nonzero element $x^* \in \partial f(0)$. We show that, at least for convex functions, this is a slightly weaker property than being sharp around the origin. We first recall an important continuity property of the subdifferential of a convex function.

Proposition 3.3.3 ([44, Theorem 24.4])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. If a sequence $\{x_k\}$ converges to $x \in \mathbb{R}^n$ and $d_k \in \partial f(x_k)$ for all k then the sequence $\{d_k\}$ is bounded and each of its limit points is a subgradient of f at x.*

Theorem 3.3.3. *Suppose $f \in \Gamma_0(\mathbb{R}^n)$ and let S be a convex set containing the origin such that $\operatorname{dom} \partial f \cap S \neq \emptyset$. If f achieves a global minimum of zero at zero and f is sharp on $S\backslash\{0\}$, then f is sparsity promoting.*

Proof. Take $x_k \to 0$ and let $d_k \in \partial f(x_k)$ for each k. Then by above, each of the limit points of d_k is a subgradient of f at 0. Since $\|d_k\| \geq c > 0$ for each k, any limit point d^* also has $\|d^*\| \geq c$. Therefore there is a nonzero element in $\partial f(0)$. □

While sharpness is enough to imply a nonzero subgradient, the converse is not true in general. Essentially, the definition of sparsity promotion ensures that the function f is relatively steep in the direction $x^* \in \partial f(0)$, but there may still be other directions in which f is flat. However, we can say that f is sharp on line segments starting at the origin.

Proposition 3.3.4. *Assume $f \in \Gamma_0(\mathbb{R}^n)$ is sparsity promoting. For any $x \in \partial f(0)$, let $\Lambda = \{\lambda x : \lambda \geq 0\}$ be the positive ray extending through x. Then*

$$m_f(x) \geq c(x) > 0,$$

where $c(x) = \max\{\|x^*\| : x^* \in \partial f(0) \cap \Lambda\}$. That is, the nonsmooth slope at x is bounded away from zero by a constant which depends on x.

Proof. Fix a nonzero $x \in \partial f(0)$. Since $\partial f(0)$ is a compact set, $\Lambda \cap \partial f(0)$ is also compact – it is simply a line segment in $\partial f(0)$. The continuous function $\|\cdot\|$ is maximized on this set, say at $x^* \neq 0$. By the monotonicity of ∂f, for any $\eta \in \partial f(x)$, $\langle \eta - x^*, x\rangle \geq 0$. Of course, then $\langle \eta, x\rangle \geq \langle x^*, x\rangle$. By the Cauchy Schwarz inequality and the fact that x and x^* are colinear, we see that $\|\eta\| \geq \|x^*\|$. Since η was any element of $\partial f(x)$, this must also hold for the infimum. □

Things are less complicated on the real line, where there are only two possible directions of increase. In this case, our definition does imply sharp near the origin.

Lemma 3.3.3. *If $f \in \Gamma_0(\mathbb{R})$, sparsity promoting implies sharp on $\partial f(0)\backslash\{0\}$.*

Proof. Since $\partial f(0)$ is compact and convex, it must be a closed interval containing origin, say $[-\lambda_1, \lambda_2]$, where at least one of the nonnegative numbers λ_1, λ_2 is nonzero. Then for any nonzero $x \in \partial f(0)$, either $x \in [-\lambda_1, 0)$ or $x \in (0, \lambda_2]$. By the previous lemma, all positive x have $m_f(x) \geq \lambda_2$ and all negative x have $m_f(x) \geq \lambda_1$. The result follows. □

Functions which are not sharp in the sense of Definition 3.3.1 may still satisfy the Łojasiewicz inequality. For instance, $\mathcal{C}^1$ functions which are defined on an o-minimal structure, and in particular subanalytic functions, satisfy (3.10) (see [29]). Recent work extends this to nonsmooth subanalytic functions [8] and characterizes subgradient trajectories of semiconvex functions [9]. See Section 2.5 for details about subanalytic functions.

Theorem 3.3.4 ([8, Theorem 3.3])**.** *Let $f \in \Gamma_0(\mathbb{R}^n)$ be subanalytic with* crit $f \neq \emptyset$*. For any bounded set K, there exists $\theta \in [0, 1)$ such that the function*

$$\frac{|f - \min f|^\theta}{m_f}$$

is bounded on K.

In order to extend this result to our semiconvex functions f_α, we first show that the subanalyticity of f_α depends on that of f. We note that the fact that $\text{env}_\alpha f$ is subanalytic if f is subanalytic is known (see, e.g. [8]), but we include the proof here as it was independently derived.

Proposition 3.3.5. *If f is subanalytic, then $\text{env}_\alpha f$ and f_α are subanalytic as well.*

Proof. Recall that real subanalytic functions define an $\mathbb{R}$-algebra (see 2.5), so if f and $\text{env}_\alpha f$ are subanalytic, then f_α must be as well. It suffices to show that f subanalytic implies $\text{env}_\alpha f$ subanalytic.

Since $\text{env}_{\alpha f}(x) = \inf\{f(u) + \frac{1}{2\alpha}\|u - x\|^2 : u \in \mathbb{R}^n\}$, we can write the graph of the envelope as follows:

$$\text{gr}(\text{env}_{\alpha f}) = \{(x,t) \in \mathbb{R}^{n+1} : t \geq 0 \text{ and } \forall y \in \mathbb{R}^n \ \big(t \leq f(y) + \frac{1}{2\alpha}\|y - x\|^2 \\ \text{and } \forall \epsilon \in \mathbb{R} \ (\epsilon > 0 \Rightarrow t + \epsilon > f(y) + \frac{1}{2\alpha}\|y - x\|^2)\big)\}. \quad (3.12)$$

We rewrite (3.12) using the fact that $\text{gr}(f)$ is subanalytic:

$$\text{gr}(\text{env}_{\alpha f}) = \Big\{(x,t) \in \mathbb{R}^{n+1} : t \geq 0 \text{ and } \forall (y,s) \in \text{gr}(f) \ \big(t \leq s + \frac{1}{2\alpha}\|y - x\|^2 \\ \text{and } \forall \epsilon \in \mathbb{R} \ (\epsilon > 0 \Rightarrow t + \epsilon > s + \frac{1}{2\alpha}\|y - x\|^2)\big)\Big\}. \quad (3.13)$$

Thus $\text{env}_\alpha f$ is subanalytic, and because subanalytic functions form an $\mathbb{R}$-algebra, f_α is subanalytic as well (see Section 2.5). The result follows. □

We show that f_α satisfies the Łojasiewicz property and determine the exponent θ. While we do not go into further detail here, knowledge of this exponent can be used to determine

convergence rates of subgradient methods [33].

Theorem 3.3.5. *If $f \in \Gamma_0(\mathbb{R}^n)$ is sparsity promoting and subanalytic, then for all $\alpha > \frac{1}{2}$ the function*

$$\frac{(f_\alpha)^{1/2}}{m_f}$$

is bounded on $\alpha\partial f(0)\backslash\{0\}$.

Proof. We first note that for $x \in \alpha\partial f(0)$, $\operatorname{env}_\alpha f(x) = \frac{1}{2\alpha}\|x\|^2$ and $f(x) \geq \|x\|^2$. Therefore $f_\alpha(x) \geq (1 - \frac{1}{2\alpha})\|x\|^2 \geq \gamma \operatorname{env}_\alpha f(x)$, for some constant $\gamma > 0$. We now essentially follow the proof of Theorem 3.3.4. By the above, we get $\|x\| \leq (2\alpha/\gamma)^{1/2}(f_\alpha(x))^{1/2}$. Using the semiconvexity of $f_\alpha(x)$, we see that

$$0 \geq f_\alpha(x) - \langle x^*, x\rangle - \frac{1}{2\alpha}\|x\|^2 \Longrightarrow f_\alpha(x) \leq \|x^*\|\|x\| + \frac{1}{2\alpha}\|x\|^2,$$

for any $x^* \in \partial f_\alpha(x)$. Applying the bound on $\|x\|$, we get

$$f_\alpha(x) \leq \|x^*\|(2\alpha/\gamma)^{1/2}(f_\alpha(x))^{1/2}(1 + \frac{1}{2\alpha}\|x\|) \leq C\|x^*\|(f_\alpha(x))^{1/2}.$$

□

Chapter 4

Some Special Functions

In this chapter, we take a closer look at some special functions of particular interest in applications: indicator functions, piecewise quadratic functions, and sums of the two. The additional structure assumed here allows us to exactly determine f_α and its proximity operator on the entire space. In particular, we show that these functions determine thresholding rules similar to the ℓ_0 and ℓ_1 norms. We collect these results in Table 1.

4.1 Indicator Functions

Indicator functions are widely used to incorporate a constraint set into the objective function of a minimization problem by restricting the domain of the objective to the set C. Recall that the indicator function of a set C is defined by

$$\iota_C(x) \coloneqq \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.1}$$

The function ι_C is convex if and only if C is a convex set (see, e.g. [5]). Throughout we assume that C is a closed convex set with boundary $\mathrm{bd}(C)$.

We show in this section that indicator functions are not only fixed by the mapping $f \mapsto f_\alpha$, but they are in fact the only functions that are fixed. As a first result, we determine when ι_C will be sparsity promoting.

Lemma 4.1.1. *The indicator function ι_C is sparsity promoting if and only if $0 \in \mathrm{bd}(C)$ and $\{0\} \subsetneq C$.*

Proof. As long as $0 \in C$, $\iota_C(0) = 0$, but to be sparsity promoting, there must be a nonzero element in $\partial\iota_C(0)$. Recall that for any x, $\partial\iota_C(x)$ is the normal cone to C at x. That is,

$$\partial\iota_C(x) = N_C(x) \coloneqq \begin{cases} \{u : \sup\langle C - x, u\rangle \leq 0\}, & \text{if } x \in C, \\ \emptyset, & \text{otherwise.} \end{cases}$$

Note that for $x \in C$, the normal cone is nonempty because$\{0\} \subseteq N_C(x)$. We further recall the following result from [5]:

$$x \in \mathrm{int}(C) \iff N_C(x) = \{0\}.$$

If $0 \in \mathrm{bd}(C)$, it follows that $N_C(x)$ is nonempty and contains a nonzero element. Conversely, if we assume $N_C(0)$ is nonempty, we must have $0 \in C$. If we further assume that $N_C(0)$ contains a nonzero element, then $0 \notin \mathrm{int}(C)$. So we see that $0 \in \mathrm{bd}(C)$ is equivalent to the sparsity promoting definition given in Section 3.1. □

It is well known (see, e.g. [5]) that $\mathrm{prox}_{\alpha\iota_C}(x) = P_C(x)$ and that $p = P_C(x)$ if and only if $x - p \in N_C(p)$. Here $P_C(x)$ is the unique operator such that $\|x - P_C(x)\|$ is the distance

from x to C. In terms of the proximity operator, this becomes $0 = \operatorname{prox}_{\alpha\iota_C}(x)$ if and only if $x \in N_C(0)$. Moreover $\operatorname{env}_\alpha \iota_C(x) = \frac{1}{2\alpha}\|P_C(x) - x\|^2$ and

$$(\iota_C)_\alpha(x) \coloneqq \iota_C(x) - \operatorname{env}_\alpha \iota_C(x) = \iota_C(x). \tag{$\mathcal{I}_\alpha$}$$

This immediately implies that $\operatorname{prox}_{\beta(\iota_C)_\alpha}(x) = P_C(x)$ as well. The converse of the above is also true.

Theorem 4.1.1. *Let $f \in \Gamma_0(\mathbb{R}^n)$ be sparsity promoting. If $f = f_\alpha$ as defined by $(\mathcal{F}_\alpha)$, then $f = \iota_{\operatorname{dom}(f)}$.*

Proof. Notice that $\operatorname{dom}(\operatorname{env}_\alpha f) = \mathbb{R}^n$ so $\operatorname{dom}(f_\alpha) = \operatorname{dom}(f)$. Hence $f = f_\alpha$ implies that $\operatorname{env}_\alpha f(x) = 0$ for all $x \in \operatorname{dom}(f)$. Because f is sparsity promoting, $f(x) \geq 0$ for all x. Hence, $0 = \operatorname{env}_\alpha f(x) = \min_{u\in\mathbb{R}^n}\{f(u) + \frac{1}{2\alpha}\|u - x\|^2\}$ for all $x \in \operatorname{dom}(f)$ implies that $f(x) = 0$ for all $x \in \operatorname{dom}(f)$. □

Remark 4.1.1. The proposition is true more generally if $f \in \Gamma_0(\mathbb{R}^n)$ is simply nonnegative.

Proposition 4.1.1. *Let $f \in \Gamma_0(\mathbb{R}^n)$ be a sparsity promoting function. Suppose that $C \subseteq \mathbb{R}^n$ is a closed convex set such that $\{0\} \subsetneq \partial f(0) \cap C$. Then the sum $\tilde{f} \coloneqq f + \iota_C$ is sparsity promoting and $\tilde{f}_\alpha = f_\alpha + \iota_C$.*

Proof. Since f is sparsity promoting, $\min_{x\in\mathbb{R}^n} f(x) = f(0) = 0$. Because $\{0\} \subsetneq \partial f(0) \cap C$, we know that $\tilde{f}(0) = \min_{x\in C} f(x) = 0$. That is, $\tilde{f}$ achieves its minimum at the origin. We can further say that $\partial\tilde{f}(0) = \partial f(0) + \partial\iota_C(0) = \partial f(0) + N_C(0)$. If $0 \in \operatorname{ri} C$, then $N_C(0) = 0$ and $\partial\tilde{f}(0) = \partial f(0)$. If $0 \in \operatorname{bd} C$, we know that $\{0\} \subset N_C(0)$, so $\partial f(0) \subseteq \partial\tilde{f}(0)$. In either case $\partial\tilde{f}(0)$ must contain a nonzero element. Therefore, it is sparsity promoting.

By Lemma 3.1.1, $\text{prox}_{\alpha f}(x) \in C$ if $x \in C$. This indicates that for $x \in C$,

$$\text{env}_\alpha f(x) = \min_{u\in\mathbb{R}} \left\{ f(u) + \frac{1}{2\alpha}\|u - x\|^2 \right\} = \min_{u\in C} \left\{ f(u) + \frac{1}{2\alpha}\|u - x\|^2 \right\} = \text{env}_\alpha \tilde{f}(x).$$

It follows that $\tilde{f}_\alpha = f_\alpha + \iota_C$. This completes the proof of the result. □

4.2 Piecewise Quadratic Functions

Piecewise quadratic functions include a variety of important examples: absolute value, rectified linear unit (ReLU), and elastic net. We generalize the proximity-related properties of these functions and provide a framework for generating customized penalty functions. For simplicity and based on the separability of these examples, we only consider functions of a single variable.

The piecewise quadratic functions we consider here have the following form

$$f(x) = \begin{cases} \frac{1}{2}a_1x^2 + b_1x, & \text{if } x \leq 0; \\ \frac{1}{2}a_2x^2 + b_2x, & \text{if } x \geq 0, \end{cases} \tag{$\mathcal{Q}$}$$

where the coefficients a_1, a_2, b_1, and b_2 are real numbers. This is a special case of the functions considered in [40] and [39].

The characterization of sparsity promoting functions having a form given ($\mathcal{Q}$) is established in the following lemma.

Lemma 4.2.1. *Let f be a piecewise quadratic function defined by ($\mathcal{Q}$). Then f is sparsity*

promoting if and only if

$$a_1 \geq 0, \quad a_2 \geq 0, \quad b_1 \leq 0 \leq b_2, \quad \textit{and} \quad b_2 - b_1 > 0. \tag{4.2}$$

Proof. "⇒": Since f is sparsity promoting, then the assumption that f attains its minimum at 0 implies that $a_1 \geq 0$, $a_2 \geq 0$, $b_1 \leq 0$, and $b_2 \geq 0$. One can directly verify that $\partial f(0) = [b_1, b_2]$. This must contain at least one nonzero element, hence, $b_2 - b_1 > 0$.

"⇐": One can see that f is nonincreasing on $(-\infty, 0]$ from $a_1 \geq 0$ and $b_1 \leq 0$ and that f is nondecreasing on $[0, \infty)$ from $a_2 \geq 0$ and $b_2 \geq 0$. So f achieves its global minimum at 0. The condition $b_2 - b_1 > 0$ implies that the set $\partial f(0) = [b_1, b_2]$ has nonzero elements. Therefore, f is a sparsity promoting function. □

Remark 4.2.1. As a by-product of the above lemma, if f given by $(\mathcal{Q})$ is a sparsity promoting function, then f must be convex, hence $f \in \Gamma_0(\mathbb{R})$.

In the rest of this section, we assume that the coefficients in $(\mathcal{Q})$ satisfy the conditions listed in (4.2). The proximity operator and Moreau envelope of f with index α at $x \in \mathbb{R}$ are

$$\operatorname{prox}_{\alpha f}(x) = \begin{cases} \min\left\{0, \frac{1}{\alpha a_1 + 1}(x - \alpha b_1)\right\}, & \text{if } x \leq 0; \\ \max\left\{0, \frac{1}{\alpha a_2 + 1}(x - \alpha b_2)\right\}, & \text{if } x \geq 0; \end{cases}$$

and

$$\operatorname{env}_\alpha f(x) = \begin{cases} \frac{1}{\alpha a_1 + 1}(f(x) - \frac{\alpha b_1^2}{2}), & \text{if } x \leq \alpha b_1; \\ \frac{1}{2\alpha}x^2, & \text{if } \alpha b_1 \leq x \leq \alpha b_2; \\ \frac{1}{\alpha a_2 + 1}(f(x) - \frac{\alpha b_2^2}{2}), & \text{if } x \geq \alpha b_2. \end{cases}$$

respectively. From the above two equations, we get

$$f_\alpha(x) = \begin{cases} \frac{\alpha a_1}{\alpha a_1 + 1} f(x) + \frac{\alpha b_1^2}{2(\alpha a_1 + 1)}, & \text{if } x \leq \alpha b_1; \\ f(x) - \frac{1}{2\alpha} x^2, & \text{if } \alpha b_1 \leq x \leq \alpha b_2; \\ \frac{\alpha a_2}{\alpha a_2 + 1} f(x) + \frac{\alpha b_2^2}{2(\alpha a_2 + 1)}, & \text{if } x \geq \alpha b_2, \end{cases} \tag{$\mathcal{Q}_\alpha$}$$

which is a piecewise quadratic polynomial with possible breakpoints at αb_1, 0, and αb_2. We know this f_α is sparsity promoting by Theorem 3.2.1. Some other properties of this function which follow immediately from $(\mathcal{Q}_\alpha)$ are collected in the following lemma.

Lemma 4.2.2. *Let $f \in \Gamma_0(\mathbb{R})$ be a sparsity promoting function defined by $(\mathcal{Q})$. Then the following hold:*

(i) f_α is nonincreasing on $(-\infty, 0]$ and is nondecreasing on $[0, \infty)$;

(ii) f_α on $(-\infty, \alpha b_1]$ is convex and is a degree 2 polynomial if $a_1 > 0$ or constant if $a_1 = 0$;

(iii) f_α on $[\alpha b_2, \infty)$ is convex and is a degree 2 polynomial if $a_2 > 0$ or a constant if $a_2 = 0$;

(iv) f_α on $[\alpha b_1, \alpha b_2]$ is convex if $\min\{a_1, a_2\} \geq \frac{1}{\alpha}$.

Just as the sparsity promoting property corresponds to certain behavior in the proximity operator near the origin, this result in Lemma 4.2.2 guarantees special properties of the proximity operator away from the origin. To illustrate, we return to $f(x) = |x|$. This satisfies $(\mathcal{Q})$ with $a_1 = a_2 = 0$, $b_1 = -1$, and $b_2 = 1$. We saw in Section 3.2 that $f_\alpha(x) = \min\{|x| - \frac{1}{2\alpha}x^2, \frac{\alpha}{2}\}$. Because this function is constant away from the origin, $\operatorname{prox}_{\beta f_\alpha}(x)$ must be the identity for large values of x. For example, if $\beta > \alpha$, $\operatorname{prox}_{\beta f_\alpha}(x) = x$ when $|x| \sqrt{\alpha\beta}$. Some other details can be found in Example 1 of Section 4.4.

In the rest of this subsection, we will give a general discussion on the proximity operator $\text{prox}_{\beta f_\alpha}$ for f_α defined by $(\mathcal{Q}_\alpha)$. We assume that $x \geq 0$ for a moment. By Lemma 3.2.3, we know that $\text{prox}_{\beta f_\alpha}(x) \subseteq [0, \infty)$, therefore by the definition of the proximity operator,

$$\text{prox}_{\beta f_\alpha}(x) = \text{Argmin}\{E(x,u) : x \in [0,+\infty)\} \coloneqq \text{Argmin}\{f_\alpha(u) + \frac{1}{2\beta}(u-x)^2 : u \in [0,+\infty)\}.$$

In view of $(\mathcal{Q}_\alpha)$, the objective function $E(x,u)$ with $(x,u) \in [0,\infty) \times [0,\infty)$ is

$$E(x,u) = \begin{cases} E_1(x,u), & \text{if } u \in [0, \alpha b_2]; \\ E_2(x,u), & \text{if } u \in [\alpha b_2, \infty), \end{cases} \tag{4.3}$$

where

$$E_1(x,u) = \frac{1}{2}\left(a_2 - \frac{1}{\alpha} + \frac{1}{\beta}\right)u^2 + \left(b_2 - \frac{1}{\beta}x\right)u + \frac{1}{2\beta}x^2, \tag{4.4}$$

$$E_2(x,u) = \frac{1}{2}\left(\frac{\alpha a_2^2}{\alpha a_2 + 1} + \frac{1}{\beta}\right)u^2 + \left(\frac{\alpha a_2 b_2}{\alpha a_2 + 1} - \frac{1}{\beta}x\right)u + \frac{\alpha b_2^2}{2(\alpha a_2 + 1)} + \frac{1}{2\beta}x^2. \tag{4.5}$$

These two functions match at the line $u = \alpha b_2$, that is, for all $x \geq 0$,

$$E_1(x, \alpha b_2) = E_2(x, \alpha b_2), \tag{4.6}$$

which will facilitate the proofs of technical lemmas given later.

Define

$$s_1(x) = \text{argmin}_{u \in [0, \alpha b_2]} E_1(x,u) \quad \text{and} \quad s_2(x) = \text{argmin}_{u \in [\alpha b_2, \infty)} E_2(x,u).$$

Obviously,

$$\text{prox}_{\beta f_\alpha}(x) \subset s_1(x) \cup s_2(x). \tag{4.7}$$

Therefore, to figure out the expression of $\text{prox}_{\beta f_\alpha}(x)$, there is a need to know the structures of the sets $s_1(x)$ and $s_2(x)$.

Since the quadratic polynomial $E_2(x, \cdot)$ is strictly convex, then we have for each $x \geq 0$, $s_2(x)$ is a singleton set as follows:

$$\begin{aligned} s_2(x) &= \max\left\{\alpha b_2, \frac{\alpha a_2 + 1}{\alpha a_2(a_2\beta + 1) + 1}\left(x - \frac{\alpha a_2 \beta b_2}{\alpha a_2 + 1}\right)\right\} \\ &= \begin{cases} \alpha b_2, & \text{if } 0 \leq x \leq \alpha b_2(a_2\beta + 1); \\ \frac{\alpha a_2 + 1}{\alpha a_2(a_2\beta+1)+1}\left(x - \frac{\alpha a_2 \beta b_2}{\alpha a_2 + 1}\right), & \text{if } x \geq \alpha b_2(a_2\beta + 1), \end{cases} \end{aligned} \tag{4.8}$$

which clearly is a piecewise linear function of x.

Lemma 4.2.3. *Let f be a piecewise quadratic sparsity promoting function as defined by $(\mathcal{Q})$. If $b_2 = 0$, then $\text{prox}_{\beta f_\alpha}(x) = s_2(x)$ for all $x \geq 0$, where s_2 is given by (4.8).*

Proof. This follows from (4.3) and (4.5) that $E(x, u) = E_2(x, u)$ for $(x, u) \in [0, \infty) \times [0, \infty)$. □

Next, we assume that $b_2 > 0$ by Lemma 4.2.1. In view of the form of $E_1(x, \cdot)$ in (4.4), we consider three cases: $a_2 - \frac{1}{\alpha} + \frac{1}{\beta} > 0$, $a_2 - \frac{1}{\alpha} + \frac{1}{\beta} = 0$, and $a_2 - \frac{1}{\alpha} + \frac{1}{\beta} < 0$, which are equivalent to (i) $\alpha b_2(a_2\beta + 1) > \beta b_2$, (ii) $\alpha b_2(a_2\beta + 1) = \beta b_2$, and (iii) $\alpha b_2(a_2\beta + 1) < \beta b_2$, respectively. Accordingly, $E_1(x, \cdot)$ is strongly convex, convex, or concave on $[0, \alpha b_2]$. The result for case (i) is stated in the following lemma.

Lemma 4.2.4. *Let f be a piecewise quadratic sparsity promoting function as defined by $(\mathcal{Q})$.*

If $b_2 > 0$ *and* $\alpha b_2(a_2\beta + 1) > \beta b_2$, *then*

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } 0 \le x < \beta b_2; \\ \frac{\alpha}{(a_2\beta+1)\alpha-\beta}(x - \beta b_2), & \text{if } \beta b_2 \le x \le \alpha b_2(a_2\beta + 1); \\ \frac{\alpha a_2+1}{\alpha a_2(a_2\beta+1)+1}\left(x - \frac{\alpha a_2 \beta b_2}{\alpha a_2+1}\right), & \text{if } x > \alpha b_2(a_2\beta + 1). \end{cases} \tag{4.9}$$

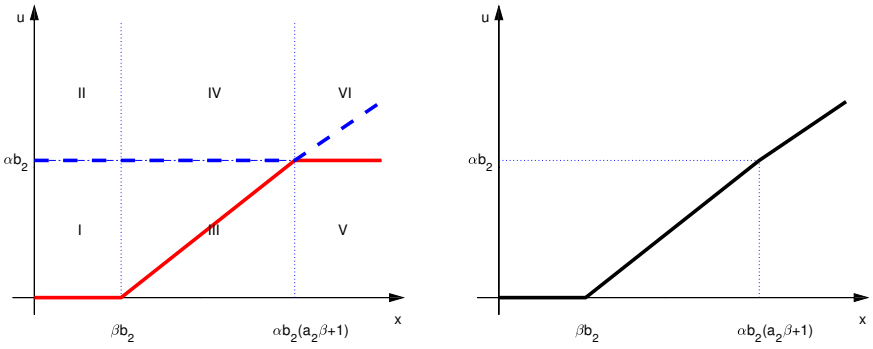
Proof. From (4.7), we first find the set $s_1(x)$ since the set $s_2(x)$ is already given in (4.8). By the assumption of this lemma, for each $x \ge 0$, $s_1(x)$ contains only one element and is given as follows:

$$s_1(x) = \begin{cases} 0, & \text{if } 0 \le x < \beta b_2; \\ \frac{\alpha}{(a_2\beta+1)\alpha-\beta}(x - \beta b_2), & \text{if } \beta b_2 \le x \le \alpha b_2(a_2\beta + 1); \\ \alpha b_2, & \text{if } x > \alpha b_2(a_2\beta + 1). \end{cases}$$

To determine the expression of $\text{prox}_{\beta f_\alpha}(x)$ from the sets $s_1(x)$ and $s_2(x)$, we look at the behaviours of the functions E_1 and E_2 in the first quadrant of the (x, u)-plane.

We use Figure 1 to visualize the minimizers of E_1 and E_2. Three vertical lines $x = 0$, $x = \beta b_2$, and $x = \alpha b_2(a_2\beta + 1)$, and two horizontal lines $u = 0$ and $u = \alpha b_2$ partition the first quadrant into six rectangular regions (I to VI). The solid red line is the graph of $s_1(x)$ while the dashed blue line is the graph of $s_2(x)$.

We know $E_1(x, 0) \le E_1(x, u)$ in region I and $E_2(x, \alpha b_2) \le E_2(x, u)$ in region II, so $E_1(x, 0) < E_2(x, \alpha b_2)$ by Equation (4.6) for $0 \le x \le \beta b_2$. We observe $E_1(x, s_1(x)) \le E_1(x, u)$ in region III and $E_2(x, \alpha b_2) \le E_2(x, u)$ in region IV, so $E_1(x, s_1(x)) < E_2(x, \alpha b_2)$ by Equation (4.6) for $\beta b_2 \le x \le \alpha b_2(a_2\beta + 1)$; Finally, we know $E_1(x, \alpha b_2) \le E_1(x, u)$ in region V and $E_2(x, s_2(x)) \le E_2(x, u)$ in region VI, so $E_2(x, s_2(x)) < E_1(x, \alpha b_2)$ by Equation (4.6) for $x > \alpha b_2(a_2\beta + 1)$. Thus $\text{prox}_{\beta f_\alpha}$ is given by (4.9). □



(a) (b)

Figure 1: An illustration of case (i): $b_2 > 0$ and $\alpha b_2(a_2\beta + 1) > \beta b_2$. The graphs of (a) $s_1(x)$ (solid) and $s_2(x)$ (dashed) and (b) the resulting proximity operator $\text{prox}_{\beta f_\alpha}(x)$.

Next result is for case (ii).

Lemma 4.2.5. *Let f be a piecewise quadratic sparsity promoting function as defined by $(\mathcal{Q})$. If $b_2 > 0$ and $\alpha b_2(a_2\beta + 1) = \beta b_2$, then*

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } 0 \le x < \beta b_2; \\ [0, \alpha b_2], & \text{if } x = \beta b_2; \\ \frac{\alpha a_2 + 1}{\alpha a_2(a_2\beta+1)+1}\left(x - \frac{\alpha a_2 \beta b_2}{\alpha a_2 + 1}\right), & \text{if } x > \beta b_2. \end{cases} \tag{4.10}$$

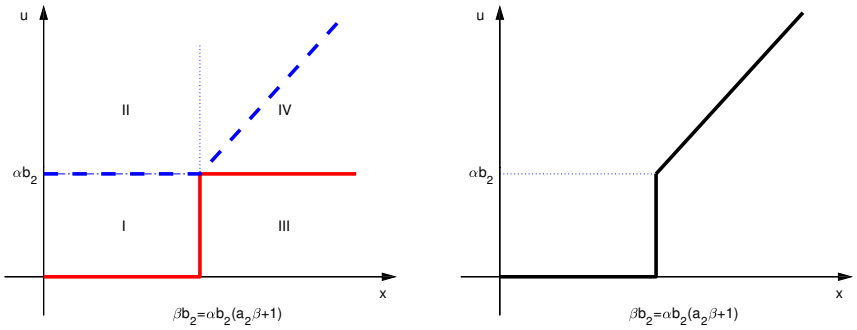
Proof. Similar to the proof of Lemma 4.2.4, we first give the explicit form of the set $s_1(x)$:

$$s_1(x) = \begin{cases} 0, & \text{if } 0 \le x < \beta b_2; \\ [0, \alpha b_2], & \text{if } x = \beta b_2; \\ \alpha b_2, & \text{if } x > \beta b_2. \end{cases}$$

We note that $\text{prox}_{\beta f_\alpha}$ can be set-valued only at βb_2.

In Figure 2, two vertical lines $x = 0$ and $x = \beta b_2$, and two horizontal lines $u = 0$ and $u = \alpha b_2$ partition the first quadrant into four rectangular regions (I to IV). The solid red

line is the graph of $s_1(x)$ while the dashed blue line is the graph of $s_2(x)$. It is identical to Figure 1 with the middle regions collapsed to a line. Following the same reasoning as in Lemma 4.2.4, we see that (4.10) holds. □



(a) (b)

Figure 2: An illustration of case (ii): $b_2 > 0$ and $\alpha b_2(a_2\beta+1) = \beta b_2$. The graphs of (a) $s_1(x)$ (solid) and $s_2(x)$ (dashed) and (b) the resulting proximity operator $\text{prox}_{\beta f_\alpha}(x)$.

Finally, we consider case (iii). Because βb_2 and $\alpha b_2(a_2\beta+1)$ have now switched positions, we see that we must take care when dealing with the intermediate x values.

Lemma 4.2.6. *Let f be a piecewise quadratic sparsity promoting function as defined by $(\mathcal{Q})$. Define*

$$\tau^+ = \frac{\alpha a_2 \beta b_2}{\alpha a_2 + 1} + \frac{\sqrt{\alpha\beta(\alpha a_2^2\beta + \alpha a_2 + 1)}b_2}{\alpha a_2 + 1}.$$

If $b_2 > 0$ and $\alpha b_2(a_2\beta + 1) < \beta b_2$,

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } 0 \le x < \tau^+; \\ \left\{0, \frac{\alpha a_2+1}{\alpha a_2(a_2\beta+1)+1}\left(\tau^+ - \frac{\alpha a_2 \beta b_2}{\alpha a_2+1}\right)\right\}, & \text{if } x = \tau^+; \\ \frac{\alpha a_2+1}{\alpha a_2(a_2\beta+1)+1}\left(x - \frac{\alpha a_2 \beta b_2}{\alpha a_2+1}\right), & \text{if } x > \tau^+. \end{cases} \tag{4.11}$$

Proof. Again, we first give the explicit form of the set $s_1(x)$. Note that $E_1(x,\cdot)$ is concave in this case, so the minimum occurs at the endpoints according to the position of the vertex.

Thus,

$$s_1(x) = \begin{cases} 0, & \text{if } 0 \le x < \frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2); \\ \{0, \alpha b_2\}, & \text{if } x = \frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2); \\ \alpha b_2, & \text{if } x > \frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2). \end{cases}$$

This is set-valued at $\frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2)$.

As before, we plot $s_1(x)$ and $s_2(x)$ in Figure 3. Three vertical lines $x = 0$, $x = \alpha b_2(a_2\beta+1)$, and $x = \frac{1}{2}(\alpha b_2(a_2\beta+1)+\beta b_2)$, and two horizontal lines $u = 0$ and $u = \alpha b_2$ partition the first quadrant into six rectangular regions as shown in Figure 3(a).The solid red line is the graph of $s_1(x)$ while the dashed blue line is the graph of $s_2(x)$. From this figure and (4.6), it is easy to see that regions I, II, V, and VI behave as in the previous cases. That is, $\text{prox}_{\beta f_\alpha}(x) = s_1(x)$ for $0 \le x \le \alpha b_2(a_2\beta + 1)$ and $\text{prox}_{\beta f_\alpha}(x) = s_2(x)$ for $x \ge \frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2)$.

To find the expression of $\text{prox}_{\beta f_\alpha}(x)$ for $\alpha b_2(a_2\beta + 1) < x < \frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2)$, from the solid red line and the dashed blue in regions III and IV, we need to compare the value of $E_1(x, 0)$ with $E_2(x, s_2(x))$. Using (4.8), a direct computation gives

$$E_2(x, s_2(x)) - E_1(x, 0) = -\frac{\alpha a_2 + 1}{2\beta(\alpha a_2(a_2\beta + 1) + 1)} \left(x - \frac{\alpha a_2 \beta b_2}{\alpha a_2 + 1} \right)^2 + \frac{\alpha b_2^2}{2(\alpha a_2 + 1)}.$$

Notice that $E_2(x, s_2(x)) - E_1(x, 0) > 0$ at $x = \alpha b_2(a_2\beta + 1)$ and $E_2(x, s_2(x)) - E_1(x, 0) < 0$ at $x = \frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2)$. Hence, the quadratic polynomial $E_2(x, s_2(x)) - E_1(x, 0)$ has only one root at τ^+ that is between $\alpha b_2(a_2\beta + 1)$ and $\frac{1}{2}(\alpha b_2(a_2\beta + 1) + \beta b_2)$. So, the result of this lemma holds and is illustrated in Figure 3(c). □

With the above results, we know $\text{prox}_{\beta f_\alpha}(x)$ for $x \ge 0$. The following lemma extends these results to $x \le 0$.

Lemma 4.2.7. *Let f be a piecewise quadratic sparsity promoting function as defined by*

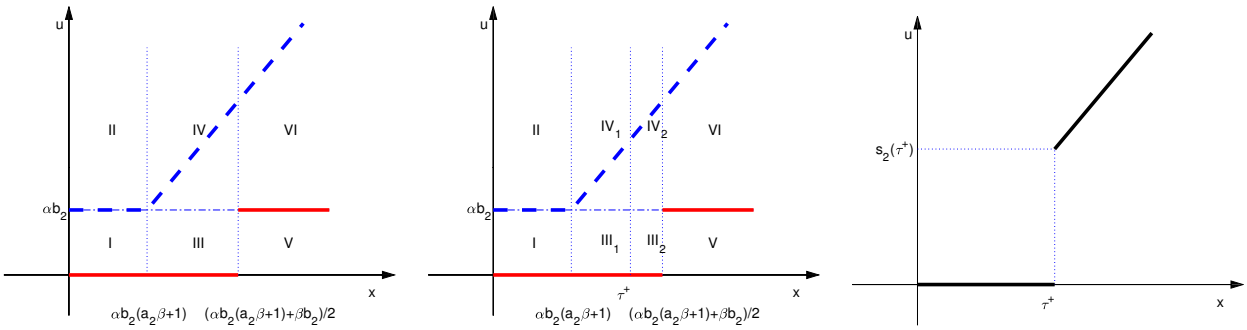


Figure 3: An illustration of case (iii): $b_2 > 0$ and $\alpha b_2(a_2\beta + 1) < \beta b_2$. The graphs of (a), (b) $s_1(x)$ (solid) and $s_2(x)$ (dashed) and (c) the resulting proximity operator $\mathrm{prox}_{\beta f_\alpha}(x)$.

($\mathcal{Q}$). *Define* $g : x \mapsto f(-x)$. *Then for* $x \leq 0$ *and any positive numbers* α *and* β, *we have* $\mathrm{prox}_{\beta f_\alpha}(x) = -\mathrm{prox}_{\beta g_\alpha}(-x)$ *where* $\mathrm{prox}_{\beta g_\alpha}(-x)$ *can be evaluated using the results in Lemmas 4.2.3-4.2.6.*

Proof. Since f is sparsity promoting, so is g by Lemma 3.2.1. Moreover, $f_\alpha = g_\alpha(-\cdot)$ which leads to $\mathrm{prox}_{\beta f_\alpha}(x) = -\mathrm{prox}_{\beta g_\alpha}(-x)$ for all x. Note that

$$g(x) = \begin{cases} \frac{1}{2}a_2x^2 - b_2x, & \text{if } x \leq 0; \\ \frac{1}{2}a_1x^2 - b_1x, & \text{if } x \geq 0, \end{cases}$$

which is a piecewise quadratic sparsity promoting function. All results developed in Lemmas 4.2.3-4.2.6 can be applied for g. Therefore, the results of this lemma follow immediately. □

In summary, we have the following result.

Theorem 4.2.1. *If* $f \in \Gamma_0(\mathbb{R})$ *is a quadratic sparsity promoting function as defined by* ($\mathcal{Q}$), *then the following statements hold.*

(i) $\text{prox}_{\beta f_\alpha}$ *is set-valued for at most one point on each side of the origin. Moreover,* $\text{prox}_{\beta f_\alpha}$ *is piecewise linear on any interval not containing these possible set-valued points.*

(ii) For any $p \in \text{prox}_{\beta f_\alpha}(x)$, $|p| \leq |x|$. *Furthermore,* $\text{sgn}(p) = \text{sgn}(x)$ *if both* p *and* x *are nonzero.*

Proof. All results follows directly from the expressions of $\text{prox}_{\beta f_\alpha}(x)$ given in Lemma 4.2.3-Lemma 4.2.7. □

Remark 4.2.2. Theorem 4.2.1 guarantees that $\text{prox}_{\beta f_\alpha}$ will be a thresholding operator for any f_α given by $(\mathcal{Q}_\alpha)$. Furthermore, Lemmas 4.2.3-4.2.6 provide detailed and easily customizable forms which can be tailored to applications.

4.3 Piecewise Quadratic on Intervals

Let C be a closed interval containing the origin and f a piecewise quadratic function defined by $(\mathcal{Q})$. We consider a function $\widetilde{f}$ that is the restriction of f on the interval C as follows:

$$\widetilde{f} = f + \iota_C. \tag{$\widetilde{\mathcal{Q}}$}$$

Let f be a piecewise quadratic sparsity promoting function defined by $(\mathcal{Q})$ and let C be a closed interval on $\mathbb{R}$ such that $\{0\} \subsetneq \partial f(0) \cap C$. By Proposition 4.1.1, $\widetilde{f}$ defined above is a sparsity promoting function, and

$$(\widetilde{f})_\alpha = f_\alpha + \iota_C. \tag{$\widetilde{\mathcal{Q}}_\alpha$}$$

For $\widetilde{f}$ defined in $(\widetilde{\mathcal{Q}})$ we always assume that the coefficients in f satisfy (4.2) and that $C = [\lambda_1, \lambda_2]$ with $\lambda_1 \leq 0 \leq \lambda_2$ and $\lambda_2 - \lambda_1 > 0$.

Theorem 4.3.1. *Let $\widetilde{f}$ be defined in $(\widetilde{\mathcal{Q}})$, let $x \in \mathbb{R}$ and let α and β be two positive numbers. Then the following statements hold.*

(i) If the set $\mathrm{prox}_{\beta f_\alpha}(x) \cap C$ is not empty, then $\mathrm{prox}_{\beta f_\alpha}(x) \cap C \subseteq \mathrm{prox}_{\beta \widetilde{f}_\alpha}(x)$;

(ii) If $\lambda_2 \in \mathrm{prox}_{\beta \widetilde{f}_\alpha}(x)$, then $\lambda_2 \in \mathrm{prox}_{\beta \widetilde{f}_\alpha}(y)$ for all $y > x$;

(iii) If $\lambda_1 \in \mathrm{prox}_{\beta \widetilde{f}_\alpha}(x)$, then $\lambda_1 \in \mathrm{prox}_{\beta \widetilde{f}_\alpha}(y)$ for all $y < x$;

Proof. (i): Assume p is an element in $\mathrm{prox}_{\beta f_\alpha}(x) \cap C$. We have

$$\begin{aligned} f_\alpha(p) + \frac{1}{2\beta}(p-x)^2 &= \min_{u \in \mathbb{R}} \left\{ f_\alpha(u) + \frac{1}{2\beta}(u-x)^2 \right\} \\ &= \min_{u \in C} \left\{ f_\alpha(u) + \frac{1}{2\beta}(u-x)^2 \right\} \\ &= \min_{u \in \mathbb{R}} \left\{ \widetilde{f}_\alpha(u) + \frac{1}{2\beta}(u-x)^2 \right\}, \end{aligned}$$

where the first equation is due to $p \in \mathrm{prox}_{\beta f_\alpha}(x)$, the second equation is due to $p \in C$, the last one is due to Theorem 4.3.1, hence, $p \in \mathrm{prox}_{\beta \widetilde{f}_\alpha}(x)$.

(ii): Since $\lambda_2 \geq 0$, the inclusion $\lambda_2 \in \mathrm{prox}_{\beta \widetilde{f}_\alpha}(x)$ together with Lemma 3.2.2 implies that $x \geq 0$ and for all $u \in [\lambda_1, \lambda_2]$,

$$\widetilde{f}_\alpha(u) + \frac{1}{2\beta}(u-x)^2 \geq \widetilde{f}_\alpha(\lambda_2) + \frac{1}{2\beta}(\lambda_2 - x)^2.$$

With the above inequality, when $y > x$, we have that

$$\begin{aligned}\widetilde{f}_\alpha(\lambda_2) + \frac{1}{2\beta}(\lambda_2 - y)^2 &= \widetilde{f}_\alpha(\lambda_2) + \frac{1}{2\beta}(\lambda_2 - x)^2 + \frac{1}{2\beta}(y - x)(y + x - 2\lambda_2) \\ &\leq \widetilde{f}_\alpha(u) + \frac{1}{2\beta}(u - x)^2 + \frac{1}{2\beta}(y - x)(y + x - 2u) \\ &= \widetilde{f}_\alpha(u) + \frac{1}{2\beta}(u - y)^2\end{aligned}$$

hold for all $u \in [\lambda_1, \lambda_2]$. This yields $\lambda_2 \in \operatorname{prox}_{\beta\widetilde{f}_\alpha}(y)$.

(iii): The proof is similar to (ii). □

Theorem 4.3.1 tells us if $\widetilde{f}$ is defined by $(\widetilde{\mathcal{Q}})$, $\operatorname{prox}_{\beta\tilde{f}_\alpha}$ will resemble the proximity operator of f_α around the origin and the proximity operator of ι_C elsewhere. Due to the number of parameters, there are a huge number of possible combinations. Rather than list all of the combinations here, we provide the details for a specific function in Example 4 of Section 4.4.

We have shown that sparsity promoting quadratic and indicator functions have thresholding proximity operators. The results essentially rely on the fact that $\operatorname{env}_\alpha f$ is quadratic for these functions. In fact, quadratic and indicator functions are the only ones with this property [41], so our discussion is a comprehensive method for obtaining thresholding rules.

4.4 Examples

In this section, we illustrate our theory by presenting several examples that are of practical interest.

For the first example, we collect and expand upon the previous discussion of $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$ for $x \in \mathbb{R}^n$. The ℓ_1-norm has been extensively used in myriad applications for promoting sparsity.

The second example is the ReLU (Rectified Linear Unit) function. It is the most commonly used activation function in convolutional neural networks or deep learning. The ReLU function on $\mathbb{R}^n$ is defined as follows: $f(x) = \sum_{i=1}^{n} \max\{0, x_i\}$, where $x \in \mathbb{R}^n$.

The third example is the elastic net penalty function, which is widely used in statistics (see [54]). The general form of the elastic net is the linear combination of the ℓ_1 and ℓ_2 norms as follows: $f(x) = \frac{\lambda_1}{2}\|x\|^2 + \lambda_2\|x\|_1$, where λ_1 and λ_2 are two nonnegative parameters. In our discussion, we will simply choose $\lambda_1 = \lambda_2 = 1$. This is known as the naive elastic net.

The last example is similar to the first one, but restricted to a cube centered at the origin. The function f is given as follows: $f(x) = \|x\|_1 + \iota_C(x)$, where $C = [-\lambda, \lambda]^n$. Generally speaking, this function promotes the sparsity on C.

We notice that the function f in the above four examples can be written as

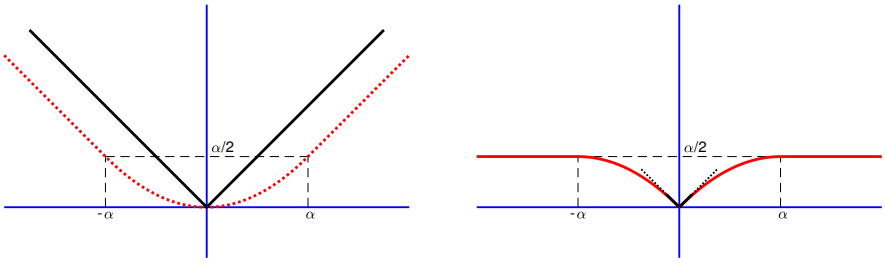
$$f(x) = \sum_{i=1}^{n} g(x_i)$$

for $x \in \mathbb{R}^n$ and some specific function g. For example, g is $|\cdot|$, $\max\{0, \cdot\}$, $\frac{1}{2}|\cdot|^2 + |\cdot|$, or $|\cdot| + \iota_{[-\lambda,\lambda]}$, in examples 1, 2, 3, or 4, an analogue of f when $\mathbb{R}^n$ reduces to $\mathbb{R}$. We further have that $\operatorname{prox}_{\alpha f}(x) = \operatorname{prox}_{\alpha g}(x_1) \times \operatorname{prox}_{\alpha g}(x_2) \times \cdots \times \operatorname{prox}_{\alpha g}(x_n)$, $\operatorname{env}_\alpha f(x) = \sum_{i=1}^{n} \operatorname{env}_\alpha g(x_i)$, $\operatorname{prox}_{\beta f_\alpha}(x) = \operatorname{prox}_{\beta g_\alpha}(x_1) \times \operatorname{prox}_{\beta g_\alpha}(x_2) \times \cdots \times \operatorname{prox}_{\beta g_\alpha}(x_n)$, and $\operatorname{env}_{\beta f_\alpha}(x) = \sum_{i=1}^{n} \operatorname{env}_{\beta g_\alpha}(x_i)$. Therefore, in the following discussion we will restrict ourself on $n = 1$.

4.4.1 Example 1: The Absolute Value Function

The first example is the absolute value function $f : \mathbb{R} \to \mathbb{R} : x \longmapsto |x|$, which is a special case of the piecewise quadratic function in $(\mathcal{Q})$ with $a_1 = a_2 = 0$, $b_1 = -1$, and $b_2 = 1$. This function is nondifferentiable at the origin with $\operatorname{argmin}_{x\in\mathbb{R}} f(x) = \{0\}$ and $\partial f(0) = \partial|\cdot|(0) =$

$[-1, 1]$.



(a) (b)

Figure 4: Example 1. (a) The graphs of f (solid), $\mathrm{env}_\alpha f$ (dotted), and (b) the graph of $f_\alpha = f(x) - \mathrm{env}_\alpha f(x)$. Near the origin f_α retains the structure of f, which is emphasized in black (solid-dotted).

The proximity operator and the Moreau envelope of f with parameter $\alpha > 0$ are

$$\mathrm{prox}_{\alpha|\cdot|}(x) = \mathrm{sgn}(x)\max\{0, |x| - \alpha\} \quad \text{and} \quad \mathrm{env}_\alpha |\cdot|(x) = \begin{cases} \frac{1}{2\alpha}x^2, & \text{if } |x| \le \alpha; \\ |x| - \frac{1}{2}\alpha, & \text{otherwise,} \end{cases}$$

respectively. It is well know that $\mathrm{prox}_{\alpha|\cdot|}$ is called the soft thresholding in literature of wavelet [19] and $\mathrm{env}_\alpha |\cdot|$ is Huber's function in robust statistics [28]. Figure 5 shows the typical shape of the proximity operator of f.

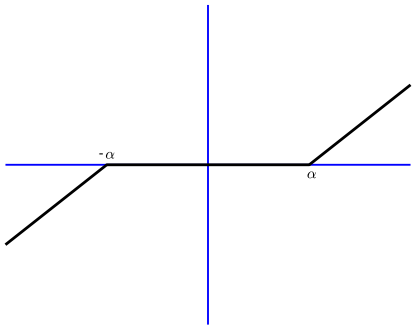


Figure 5: Example 1. The typical shape of $\mathrm{prox}_{\alpha f}$.

As defined in $(\mathcal{F}_\alpha)$, for the absolute value function f,

$$f_\alpha(x) := |x| - \mathrm{env}_\alpha\,|\cdot|(x) = \begin{cases} |x| - \frac{1}{2\alpha}x^2, & \text{if } |x| \le \alpha; \\ \frac{1}{2}\alpha, & \text{otherwise.} \end{cases} \tag{4.12}$$

This function f_α (see Figure 4(b)) is identical to the minimax convex penalty (MCP) function given in [53], but motivated from statistics perspective.

The expression of $\mathrm{prox}_{\beta f_\alpha}$ depends on the relative values of α and β. If $\beta < \alpha$, Lemma 4.2.4 gives

$$\mathrm{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| \le \beta; \\ \frac{\alpha}{\alpha-\beta}(|x| - \beta)\,\mathrm{sgn}(x), & \text{if } \beta < |x| \le \alpha; \\ x, & \text{if } |x| \ge \alpha. \end{cases} \tag{4.13}$$

This is the firm thresholding operator [11]. If $\beta = \alpha$, Lemma 4.2.5 gives

$$\mathrm{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| < \alpha; \\ [0, \alpha], & \text{if } |x| = \alpha; \\ x, & \text{if } |x| > \alpha, \end{cases} \tag{4.14}$$

Finally, if $\beta > \alpha$, Lemma 4.2.6 gives

$$\mathrm{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| < \sqrt{\alpha\beta}; \\ \{0, x\}, & \text{if } |x| = \sqrt{\alpha\beta}; \\ x, & \text{if } |x| > \sqrt{\alpha\beta}; \end{cases} \tag{4.15}$$

The proximity operator $\mathrm{prox}_{\beta f_\alpha}$ for different values of α and β is plotted in Figure 6.

To end this example, we give several remarks on the proximity operators of $\mathrm{prox}_{\alpha f}$ and

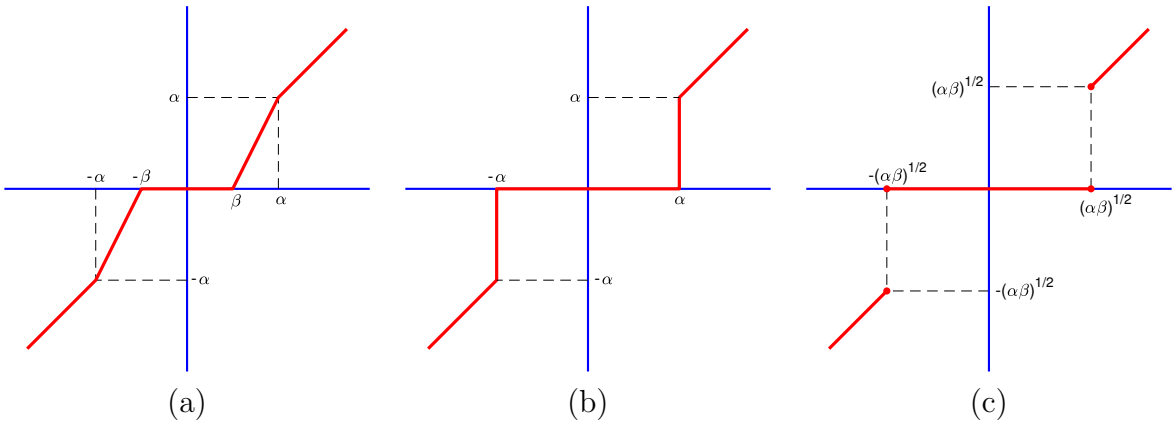


Figure 6: Typical shapes of the proximity operator of $|\cdot|_\alpha$ for (a) $\beta < \alpha$, (b) $\beta = \alpha$, (c) $\beta > \alpha$. The sparsity threshold and the thresholding behavior depend on the relationship between α and β.

$\text{prox}_{\beta f_\alpha}$ as follows:

- Note that $\partial f(0) = [-1, 1]$. The results given in (4.13) (for $\beta < \alpha$) and (4.14) (for $\beta = \alpha$) exactly match the first two statements of Theorem 3.2.2. For $\beta > \alpha$, the $\text{prox}_{\beta f_\alpha}(x) = 0$ for all $x \in [-\sqrt{\alpha\beta}, \sqrt{\alpha\beta}]$, which includes the interval $[-\alpha, \alpha] = \alpha\partial f(0)$ as indicated in the third statement of Theorem 3.2.2.

- The operator $\text{prox}_{\alpha f}$ forces its variable to zero when the absolute value is less than a given threshold, and otherwise reduces the variable, in absolute value, by the amount of the threshold. Like $\text{prox}_{\alpha f}$, $\text{prox}_{\beta f_\alpha}$ forces its variable to zero when the absolute value is less than a given threshold, but it fixes variables whose absolute value exceeds a certain threshold.

- For $\beta \geq \alpha$ the proximity operator $\text{prox}_{\beta f_\alpha}$ is almost identical to the hard threshold operator. Let $|\cdot|_0$ be the ℓ_0-norm on $\mathbb{R}$, that is, $|x|_0$ equals 1 if x is nonzero, 0 otherwise.

The proximity operator of $|\cdot|_0$ with parameter γ at x is

$$\text{prox}_{\gamma|\cdot|_0}(x) = \begin{cases} \{0\}, & \text{if } |x| < \sqrt{2\gamma}; \\ \{0, x\}, & \text{if } |x| = \sqrt{2\gamma}; \\ \{x\}, & \text{if } |x| > \sqrt{2\gamma}, \end{cases}$$

which is called the hard thresholding operator with threshold $\sqrt{2\gamma}$. We can see that $\text{prox}_{\gamma|\cdot|_0} = \text{prox}_{\beta f_\alpha}$ as long as $2\gamma = \alpha\beta$ and $\beta > \alpha$. It is interesting that although $|\cdot|_0$ is discontinuous and f_α is continuous, they have the same proximity operator. Moreover, by fixing α and varying the parameter β, the proximity operator $\text{prox}_{\beta f_\alpha}$ changes from the firm thresholding operator to the hard thresholding operator.

4.4.2 Example 2: ReLU Function

The ReLU (Rectified Linear Unit) function on $\mathbb{R}$ is

$$f(x) := \max\{0, x\},$$

which is a special case of the piecewise quadratic function in $(\mathcal{Q})$ with $a_1 = b_1 = a_2 = 0$ and $b_2 = 1$. The proximity operator and the Moreau envelope of f with parameter $\alpha > 0$ are

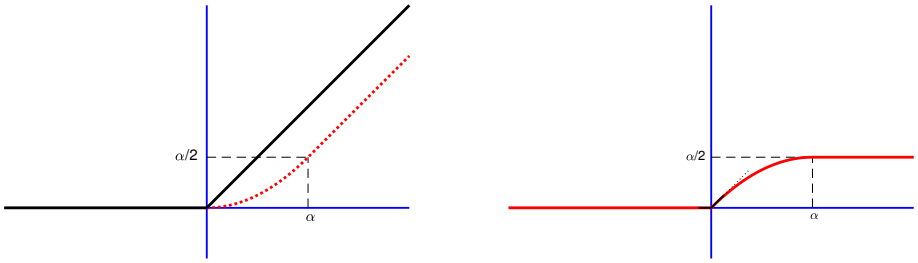
$$\text{prox}_{\alpha f}(x) = \min\{x, \max\{0, x - \alpha\}\},$$

$$\text{env}_\alpha f(x) = \begin{cases} 0 & \text{if } x \le 0; \\ \frac{1}{2\alpha}x^2, & \text{if } 0 \le x \le \alpha; \\ x - \frac{1}{2}\alpha, & \text{if } x \ge \alpha, \end{cases}$$

respectively. By $(\mathcal{F}_\alpha)$, $f_\alpha(x) = f(x) - \mathrm{env}_\alpha f(x)$ is

$$f_\alpha(x) = \begin{cases} 0, & \text{if } x < 0; \\ x - \frac{1}{2\alpha}x^2, & \text{if } 0 \le x \le \alpha; \\ \frac{\alpha}{2}, & \text{if } x > \alpha. \end{cases} \tag{4.16}$$

Figure 7(a) depicts the graphs of f and $\mathrm{env}_\alpha f$ while Figure 7(b) presents the function f_α. The graph of $\mathrm{prox}_{\alpha f}$ is given in Figure 8.



(a) (b)

Figure 7: Example 2. (a) The graphs of f (solid), $\mathrm{env}_\alpha f$ (dotted), and (b) their difference $f_\alpha = f - \mathrm{env}_\alpha f$. The singularity of f_α at zero is emphasized in black (solid-dotted).

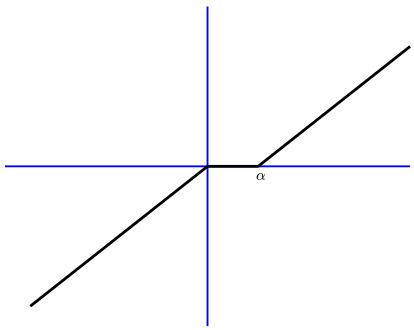


Figure 8: Example 2. The typical shape of $\mathrm{prox}_{\alpha f}$. The parameter α is the sparsity threshold.

As in example 1, the expression of $\mathrm{prox}_{\beta f_\alpha}$ depends on the relative values of α and β. If

$\beta < \alpha$,

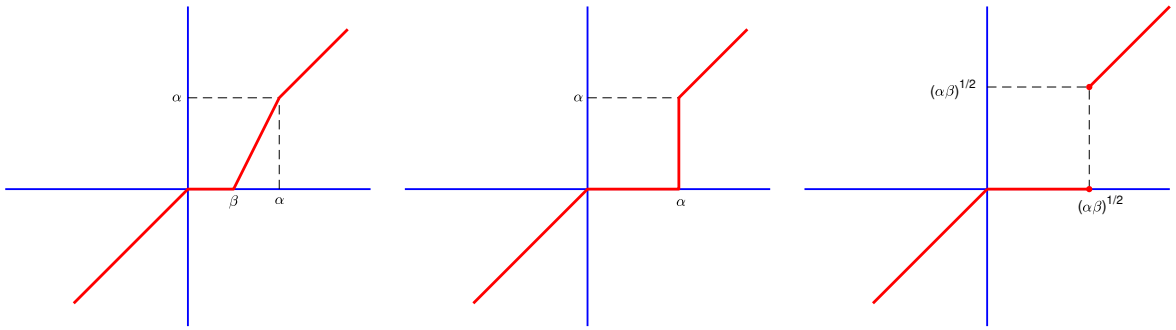
$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} x, & \text{if } x \leq 0 \text{ or } x \geq \alpha; \\ 0, & \text{if } 0 \leq x \leq \beta; \\ \frac{\alpha(x-\beta)}{\alpha-\beta}; & \text{if } \beta \leq x \leq \alpha. \end{cases} \tag{4.17}$$

If $\beta = \alpha$,

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} x, & \text{if } x \leq 0 \text{ or } x > \alpha; \\ 0, & \text{if } 0 \leq x < \alpha; \\ [0, \alpha] & \text{if } x = \alpha. \end{cases} \tag{4.18}$$

Finally, if $\beta > \alpha$,

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} x, & \text{if } x \leq 0 \text{ or } x > \sqrt{\alpha\beta}; \\ 0, & \text{if } 0 \leq x < \sqrt{\alpha\beta}; \\ \{0, \sqrt{\alpha\beta}\}, & \text{if } x = \sqrt{\alpha\beta}. \end{cases} \tag{4.19}$$



(a) (b) (c)

Figure 9: Example 2. Typical shapes of the proximity operator of f_α for (a) $\beta < \alpha$; (b) $\beta = \alpha$; and (c) $\beta > \alpha$.

Note that $\partial f(0) = [0, 1]$. The results given in (4.17) (for $\beta < \alpha$) and (4.18) (for $\beta = \alpha$)

exactly match the first two statements of Theorem 3.2.2. For $\beta > \alpha$, equation (4.19) shows that $\operatorname{prox}_{\beta f_\alpha}(x) = 0$ for all $x \in [0, \sqrt{\alpha\beta}]$, which includes the interval $[0, \alpha] = \alpha\partial f(0)$ as indicated in the third statement of Theorem 3.2.2.

4.4.3 Example 3: Elastic Net

The elastic net is a regularized regression method in data analysis that linearly combines the ℓ_1 and ℓ_2 penalties of the LASSO and ridge methods. In this example, we consider a special case of the elastic net in $\mathbb{R}$:

$$f(x) = \frac{1}{2}x^2 + |x|.$$

This is an instance of the piecewise quadratic function given in $(\mathcal{Q})$ with $a_1 = a_2 = 1$, $b_1 = -1$ and $b_2 = 1$. Clearly, f is nondifferentiable at the origin with $\operatorname{argmin}_{x\in\mathbb{R}} f(x) = \{0\}$. Moreover, $\partial f(0) = \partial|\cdot|(0) = [-1, 1]$.

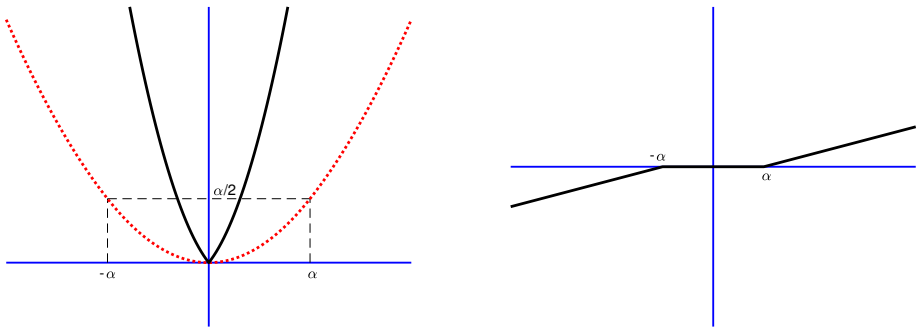
The proximity operator and the Moreau envelope of f with parameter $\alpha > 0$ are

$$\operatorname{prox}_{\alpha f}(x) = \max\left\{0, \frac{1}{\alpha+1}(|x| - \alpha)\right\}\operatorname{sgn}(x),$$

$$\operatorname{env}_\alpha f(x) = \begin{cases} \frac{1}{2\alpha}x^2, & \text{if } |x| \le \alpha; \\ \frac{1}{\alpha+1}(\frac{1}{2}x^2 + |x| - \frac{\alpha}{2}), & \text{if } |x| \ge \alpha, \end{cases}$$

respectively.

The graphs of f and $\operatorname{env}_\alpha f$ are plotted in Figure 10 (a). The graph of $\operatorname{prox}_{\alpha f}$ is plotted in Figure 10 (b). As in the case of the absolute value function, $\operatorname{prox}_{\alpha f}$ sends all values between α and $-\alpha$ to zero. Unlike the absolute value, it also contracts elements outside of this interval toward the origin.



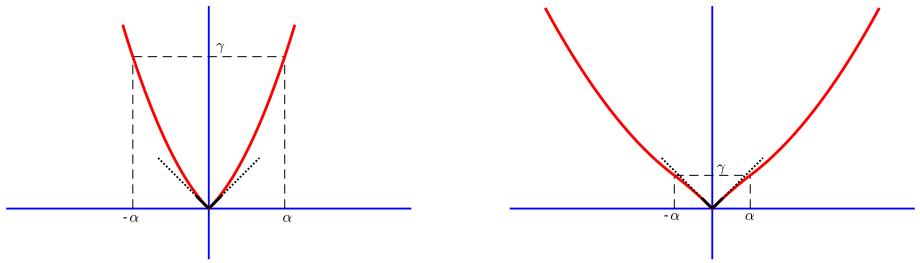
(a) (b)

Figure 10: Example 3. (a) The graphs of f (solid) and $\mathrm{env}_\alpha f$ (dotted); and (b) the graph of $\mathrm{prox}_{\alpha f}$.

Now f_α, the difference between f and its Moreau envelope $\mathrm{env}_\alpha f$, is

$$f_\alpha(x) = \begin{cases} \frac{\alpha-1}{2\alpha}x^2 + |x|, & \text{if } |x| \le \alpha; \\ \frac{\alpha}{2(\alpha+1)}x^2 + \frac{\alpha}{\alpha+1}|x| + \frac{\alpha}{2(\alpha+1)}, & \text{if } |x| \ge \alpha. \end{cases} \tag{4.20}$$

We remark that f_α is convex when $\alpha \ge 1$ and nonconvex when $\alpha < 1$. The graph of f_α for $\alpha \ge 1$ and $\alpha < 1$ are shown in Figure 11(a) and (b), respectively.



(a) (b)

Figure 11: Example 3. The graph of f_α when (a) $\alpha \ge 1$ and (b) $\alpha < 1$. The singularity of f_α at zero is emphasized in black (solid-dotted).

According to the discussion given in subsection 4.2, we consider three cases: $\beta(\alpha-1)+\alpha > 0$, $\beta(\alpha-1)+\alpha = 0$, and $\beta(\alpha-1)+\alpha < 0$. These cases are equivalent to $\alpha(\beta+1) > \beta$,

$\alpha(\beta+1)=\beta$, and $\alpha(\beta+1)<\beta$ respectively. Recall that these cases correspond to the convexity (or lack thereof) of $f_\alpha(u)+\frac{1}{2\beta}(u-x)^2$ for u close to zero.

Case 1: $\alpha(\beta+1)>\beta$. In this case, by Lemma 4.2.4 we have

$$\operatorname{prox}_{\beta f_\alpha}(x)=\begin{cases}0 & \text{if } |x|\le\beta;\\ \frac{\alpha}{\alpha\beta-\beta+\alpha}(x-\beta\operatorname{sgn}(x)) & \text{if } \beta\le|x|\le\alpha(\beta+1);\\ \frac{\alpha+1}{\alpha\beta+\alpha+1}(x-\frac{\alpha\beta}{\alpha+1}\operatorname{sgn}(x)) & \text{if } \alpha(\beta+1)\le|x|.\end{cases} \tag{4.21}$$

Case 2: $\alpha(\beta+1)=\beta$. By Lemma 4.2.5 we have

$$\operatorname{prox}_{\beta f_\alpha}(x)=\begin{cases}0 & \text{if } |x|\le\beta;\\ [0,\alpha]\operatorname{sgn}(x) & \text{if } |x|=\beta;\\ \frac{\alpha+1}{\alpha\beta+\alpha+1}(x-\frac{\alpha\beta}{\alpha+1}\operatorname{sgn}(x)) & \text{if } \beta\le|x|.\end{cases} \tag{4.22}$$

Case 3: $\alpha(\beta+1)<\beta$. Define

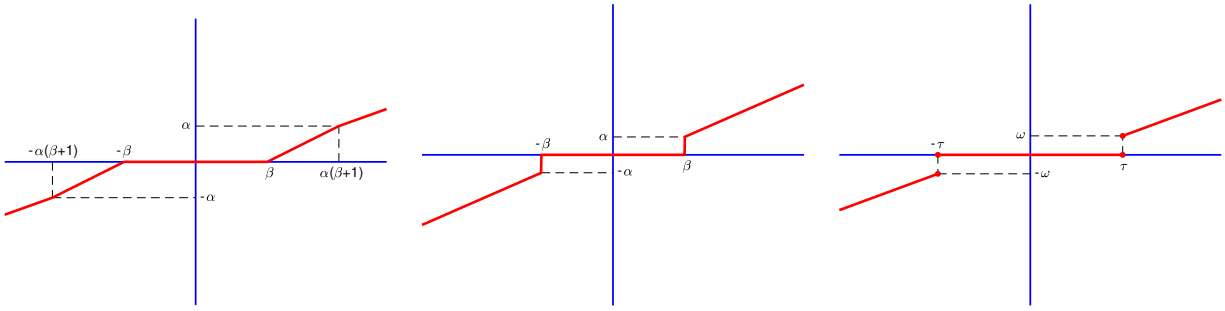
$$\tau=\frac{\alpha\beta}{\alpha+1}+\frac{\sqrt{\alpha\beta(\alpha\beta+\alpha+1)}}{\alpha+1}. \tag{4.23}$$

as in Lemma 4.2.6. Then we have

$$\operatorname{prox}_{\beta f_\alpha}(x)=\begin{cases}0 & \text{if } |x|\le\tau;\\ \{0,\omega\} & \text{if } |x|=\tau;\\ \frac{(\alpha+1)x-\alpha\beta\operatorname{sgn}(x)}{\alpha\beta+\alpha+1}, & \text{if } |x|>\tau,\end{cases} \tag{4.24}$$

where $\omega=\frac{(\alpha+1)\tau-\alpha\beta}{\alpha\beta+\alpha+1}$. The graphs of $\operatorname{prox}_{\beta f_\alpha}$ in the above three cases are plotted in Fig-

ure 4.4.3.



(a) (b) (c)

Figure 12: Example 3. Typical shapes of $\text{prox}_{\beta f_\alpha}$ when (a) $\alpha(\beta+1) > \beta$, (b) $\alpha(\beta+1) = \beta$, and (c) $\alpha(\beta+1) < \beta$.

Below are some comments on this example.

- The function f_α in the first two examples is nonconvex for any $\alpha > 0$, however, by Proposition 3.2.1 it is convex if $\alpha \geq 1$ due to our elastic net function f being 1-strongly convex.

- The computation of the proximity operator $\text{prox}_{\beta f_\alpha}$ is discussed under three different situations, namely, $\alpha(\beta+1) > \beta$, $\alpha(\beta+1) = \beta$, and $\alpha(\beta+1) < \beta$. These situations are quite natural from Proposition 3.2.1. Since f is 1-strongly convex, hence, the function $f_\alpha + \frac{1}{2\beta}(\cdot - x)^2$ is $(1+\beta^{-1}-\alpha^{-1})$-strongly convex if $\alpha(\beta+1) > \beta$, convex if $\alpha(\beta+1) = \beta$, and $(\alpha^{-1} - 1 - \beta^{-1})$-semiconvex if $\alpha(\beta+1) < \beta$.

- For the case of $\beta \leq \alpha$, we know that $\alpha(1+\beta) > \beta$, so the proximity operator given (4.21) covers both statements 1 and 2 in Theorem 3.2.2.

- For the case of $\beta > \alpha$, there are three possible related cases. If $\alpha < \beta < \alpha(\beta+1)$ (resp. $\alpha < \beta = \alpha(\beta+1)$), the proximity operator given (4.21) (resp. (4.22)) shows that this operator vanishes at all elements in $\beta\partial f(0) = [-\beta, \beta] \supset \alpha\partial f(0)$, fulfilling the

third statement of Theorem 3.2.2. If $\beta > \alpha(\beta+1)$, we know that $\alpha < 1$, $\beta > \frac{\alpha}{1-\alpha}$, and τ defined in (4.23) satisfying

$$\tau = \frac{\alpha\beta}{\alpha+1} + \frac{\sqrt{\alpha\beta(\alpha\beta+\alpha+1)}}{\alpha+1} > \frac{\alpha^2}{1-\alpha^2} + \frac{\alpha}{1-\alpha^2} > \alpha.$$

Hence, the proximity operator given (4.24) annihilates all elements in $\tau\partial f(0) \supset \alpha\partial f(0)$, once again fulfilling the third statement of Theorem 3.2.2.

4.4.4 Example 4: Absolute Value on an Interval Centered at the Origin

Let λ be a positive parameter. The absolute function on the interval $[-\lambda, \lambda]$ centered at the origin is

$$f(x) := |x| + \iota_{[-\lambda,\lambda]}(x),$$

which is a special case given in $(\widetilde{\mathcal{Q}})$ with $a_1 = a_2 = 0$, $b_1 = -1$, $b_2 = 1$, and $C = [-\lambda, \lambda]$. Its proximity operator and Moreau envelope with parameter α at point x, respectively, are

$$\operatorname{prox}_{\alpha f}(x) = \begin{cases} 0, & \text{if } |x| \le \alpha; \\ \operatorname{sgn}(x)(|x| - \alpha), & \text{if } \alpha < |x| \le \alpha + \lambda; \\ \lambda \operatorname{sgn}(x), & \text{if } \alpha + \lambda < |x|; \end{cases}$$

and

$$\operatorname{env}_\alpha f(x) = \begin{cases} |x| - \frac{\alpha}{2} + \frac{1}{2\alpha}(|x| - \alpha)^2, & \text{if } |x| \le \alpha; \\ |x| - \frac{\alpha}{2}, & \text{if } \alpha < |x| \le \alpha + \lambda; \\ |x| - \frac{\alpha}{2} + \frac{1}{2\alpha}(|x| - (\lambda + \alpha))^2, & \text{if } \alpha + \lambda < |x|. \end{cases}$$

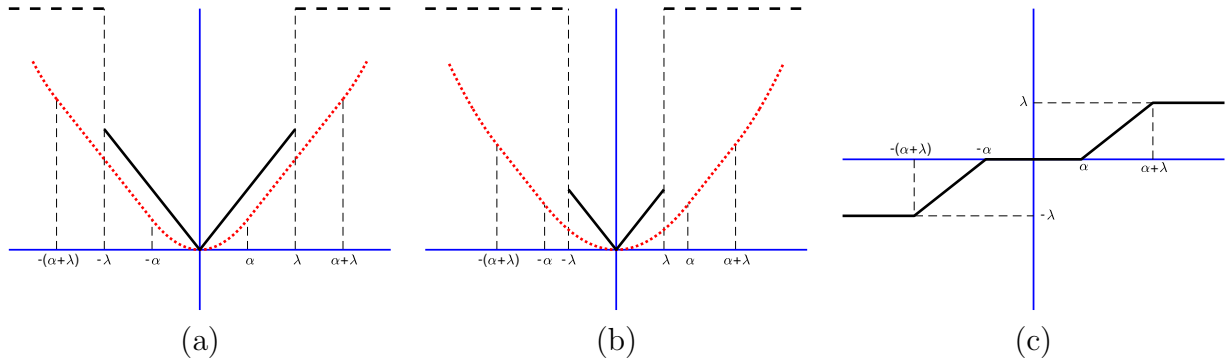


Figure 13: Example 4. The graphs of f (solid, dashed) and $\operatorname{env}_\alpha f$ (dotted) when (a) $\alpha < \lambda$ and (b) $\alpha > \lambda$. The graph of $\operatorname{prox}_{\alpha f}$ is shown in (c). Between $-(\alpha + \lambda)$ and $\alpha + \lambda$, $\operatorname{prox}_{\alpha f}$ is the soft thresholding operator with sparsity parameter α; otherwise it projects onto this interval.

Figure 4.4.4 depicts the graphs of f, $\operatorname{env}_\alpha f$, and $\operatorname{prox}_{\alpha f}$. We observe that on the interval $[-\lambda, \lambda]$ (the domain of f_α) the envelope $\operatorname{env}_\alpha f$ is a piecewise quadratic polynomial (Figure 4.4.4(a)) if $\alpha < \lambda$ and is simply a quadratic polynomial (Figure 4.4.4(b)) if $\alpha \ge \lambda$. It turns out that the expression of $\operatorname{prox}_{\beta f_\alpha}$ for $\alpha < \lambda$ is much more complicated than that for $\alpha \ge \lambda$ as we will see below.

As both f and $\operatorname{env}_\alpha f$ depend on α and λ, the explicit expression for f_α will depend on the values of these parameters. To compute the proximity operator $\operatorname{prox}_{\beta f_\alpha}$, we consider separately two main cases: $\alpha < \lambda$ and $\alpha \ge \lambda$.

Case 1: $\alpha < \lambda$. In this case, we get (see Figure 4.4.4)

$$f_\alpha(x) = f(x) - \operatorname{env}_\alpha f(x) = \begin{cases} \frac{\alpha}{2} - \frac{1}{2\alpha}(|x| - \alpha)^2, & \text{if } |x| \le \alpha; \\ \frac{\alpha}{2}, & \text{if } a \le |x| \le \lambda; \\ +\infty, & \text{if } \lambda < |x|. \end{cases} \tag{4.25}$$

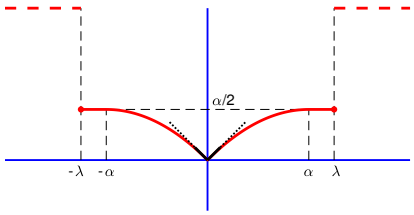


Figure 14: Example 4. The graph of f_α when $\alpha < \lambda$ with the singularity of f_α at zero emphasized in black (solid-dotted). Further, we see that f_α agrees with Example 1 on $[-\lambda, \lambda]$.

Depending on the values of α, β, and λ, we consider four possible cases: $\beta < \alpha < \lambda$, $\beta = \alpha < \lambda$, $\alpha < \beta \le \lambda$, and $\lambda < \beta$.

Case 1.1: $\beta < \alpha < \lambda$. In this case, we have

$$\operatorname{prox}_{\beta f_\alpha}(x) = \begin{cases} \max\{0, \frac{\alpha(|x|-\beta)}{\alpha-\beta}\} \operatorname{sgn}(x), & \text{if } |x| \le \alpha; \\ \min\{|x|, \lambda\} \operatorname{sgn}(x), & \text{if } |x| > \alpha. \end{cases} \tag{4.26}$$

Case 1.2: $\beta = \alpha < \lambda$. In this case, we have

$$\operatorname{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| < \alpha; \\ \operatorname{sgn}(x)[0, \alpha], & \text{if } |x| = \alpha; \\ \operatorname{sgn}(x) \min\{|x|, \lambda\}, & \text{if } \alpha < |x|, \end{cases} \tag{4.27}$$

Case 1.3: $\alpha < \beta \leq \lambda$. In this case, we have

$$\operatorname{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| < \sqrt{\alpha\beta}; \\ \{0, \operatorname{sgn}(x)\sqrt{\alpha\beta}\}, & \text{if } |x| = \sqrt{\alpha\beta}; \\ \min\{|x|, \lambda\} \operatorname{sgn}(x), & \text{if } \sqrt{\alpha\beta} < |x|, \end{cases} \tag{4.28}$$

Case 1.4: $\alpha < \lambda < \beta$. We have

$$\operatorname{prox}_{\beta f_\alpha}(x) = \begin{cases} \{0\}, & \text{if } |x| < \frac{\alpha\beta+\lambda^2}{2\lambda}; \\ \{0, \lambda \operatorname{sgn}(x)\}, & \text{if } |x| = \frac{\alpha\beta+\lambda^2}{2\lambda}; \\ \{\lambda \operatorname{sgn}(x)\}, & \text{if } \frac{\alpha\beta+\lambda^2}{2\lambda} < |x|, \end{cases} \tag{4.29}$$

We now move on to the second main case.

Case 2: $\lambda \leq \alpha$. In this case, we get (see Figure 4.4.4)

$$f_\alpha(x) = \begin{cases} \frac{\alpha}{2} - \frac{1}{2\alpha}(|x| - \alpha)^2, & \text{if } |x| \leq \lambda; \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.30}$$

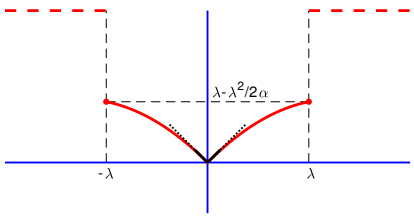


Figure 15: Example 4. The graph of f_α when $\lambda \leq \alpha$ with the singularity of f_α at zero emphasized in black (solid-dotted). As before, f_α agrees with Example 1 on $[-\lambda, \lambda]$, but is cut off before it plateaus.

To compute $\text{prox}_{\beta f_\alpha}$, we consider three situations: $\beta < \alpha$, $\beta = \alpha$, and $\beta > \alpha$.

Case 2.1: $\beta < \alpha$. In this case, we have that

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| \leq \beta; \\ \frac{\alpha(|x|-\beta)}{\alpha-\beta}\,\text{sgn}(x), & \text{if } \beta \leq |x| \leq \beta + \frac{\alpha-\beta}{\alpha}\lambda; \\ \lambda\,\text{sgn}(x), & \text{if } \beta + \frac{\alpha-\beta}{\alpha}\lambda \leq |x|, \end{cases} \tag{4.31}$$

Case 2.2: $\beta = \alpha$. In this case, we have

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| < \alpha; \\ \text{sgn}(x)[0,\lambda], & \text{if } |x| = \alpha; \\ \lambda\,\text{sgn}(x), & \text{if } \alpha < |x|, \end{cases} \tag{4.32}$$

Case 2.3: $\beta > \alpha$. Similar to Case 1.4, we get

$$\text{prox}_{\beta f_\alpha}(x) = \begin{cases} 0, & \text{if } |x| \leq \beta - \frac{\beta-\alpha}{2\alpha}\lambda; \\ \text{sgn}(x)\{0,\lambda\}, & \text{if } |x| = \beta - \frac{\beta-\alpha}{2\alpha}\lambda; \\ \lambda\,\text{sgn}(x), & \text{if } \beta - \frac{\beta-\alpha}{2\alpha}\lambda < |x|, \end{cases} \tag{4.33}$$

Table of Functions and Proximity Operators

To close this chapter, we provide a reference table collecting our results.

Table 1: Functions and proximity operators for all examples

$f(x)$	$f_\alpha(x)$	$\beta < \alpha$	$\beta = \alpha$	$\beta > \alpha$
$\lvert x\rvert$	(4.12)	(4.13)	(4.14)	(4.15)
$\max\{0, x\}$	(4.16)	(4.17)	(4.18)	(4.19)
	$\alpha < \lambda$ $\alpha \geq \lambda$	$\alpha < \lambda$ $\alpha \geq \lambda$	$\alpha < \lambda$ $\alpha \geq \lambda$	$\beta \leq \lambda$ $\beta > \lambda$ $\alpha \geq \lambda$
$\lvert x\rvert + \iota_{[-\lambda,\lambda]}$	(4.25) (4.30)	(4.26) (4.31)	(4.27) (4.32)	(4.28) (4.29) (4.33)
		$\beta < \alpha(\beta+1)$	$\beta = \alpha(\beta+1)$	$\beta > \alpha(\beta+1)$
$\frac{1}{2}x^2 + \lvert x\rvert$	(4.20)	(4.21)	(4.22)	(4.24)

Chapter 5

Algorithms

We explore several methods for solving the f_α-penalized least squares problem

$$\min_{x \in X} \; f_\alpha(Dx) + \frac{1}{2\lambda}\|x - z\|^2, \tag{P}$$

where X is Euclidean space with its usual norm and f_α is one of our sparsity promoting functions. As discussed in Chapter 3, this model may be convex or nonconvex depending on the parameters α and λ, and it can be decomposed as a difference of convex functions. We consider three algorithms which highlight each of these cases: Primal-Dual Splitting, Difference of Convex, and the Alternating Directions Method of Multipliers. We connect properties of our functions with known convergence analysis and provide improvements where possible.

5.1 Primal-Dual Splitting

The Primal-Dual Splitting Algorithm (PDA) was introduced by Chambolle and Pock to minimize the sum of two convex functions, one of which is composed with a bounded linear operator [14]. Condat later extended this to include a third term with a Lipschitz gradient [17], and it is this framework that we discuss. PDA is an example of a proximal splitting algorithm: the problem is split into simpler subproblems which can be solved using the proximity (or proximal) operators of individual functions in the objective. The term primal-dual comes from the fact that it outputs both a primal and a dual solution.

The generic model considered here is

$$\operatorname{argmin}\{F(x) + G(x) + H(Bx) : x \in X\} \tag{5.1}$$

such that

- F is convex and differentiable with L-Lipschitz gradient,
- G and H are prox-friendly: that is, their proximity operators have an explicit form or can be easily computed,
- and B is a bounded linear operator with adjoint B^* and induced norm $\|B\| = \sup\{\|Bx\| : \|x\| \leq 1\}$,
- The set of minimizers is nonempty.

The dual problem is then

$$\operatorname{argmin}\{(F + G)^*(-B^*y) + H^*(y) : y \in Y\}. \tag{5.2}$$

Generalized Karush-Kuhn-Tucker conditions provide necessary first order conditions for primal-dual solution pairs (x^*, y^*). If $x^* \in X$ and $y^* \in Y$ satisfy the variational inclusion

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G(x^*) + B^*(y^*) + \nabla F(x^*) \\ -Bx^* + \partial H^*(y^*) \end{pmatrix}, \tag{5.3}$$

then x^* solves (5.1) and y^* solves (5.2). To find such a pair, PDA iteratively solves the above inclusions. Given initial points (x_0, y_0), positive parameters σ and τ, and a sequence of a positive averaging weights $\{\rho_n\}$, the iterates are computed by:

$$x_{n+1}^+ \coloneqq \operatorname{prox}_{\tau G}(x_n - \tau \nabla F(x_n) - \tau B^* y_n) \tag{5.4}$$

$$y_{n+1}^+ \coloneqq \operatorname{prox}_{\sigma H*}(y_n + \sigma B(2x_{n+1}^+ - x_n)) \tag{5.5}$$

$$(x_{n+1}, y_{n+1}) \coloneqq \rho_k(x_{n+1}^+, y_{n+1}^+) + (1 - \rho_n)(x_n, y_n). \tag{5.6}$$

The following theorem guarantees weak convergence to a solution pair $(x^*, y^*) \in X \times Y$.

Proposition 5.1.1 (Condat [17])**.** *Let τ, σ, and the sequence $\{\rho_n\}$ be the parameters in* (5.4)–(5.6)*. Suppose that the functions F, G, and H in* (5.1) *are convex, the gradient of F is L-Lipschitz with $L > 0$, and the following hold:*

(i) $\frac{1}{\tau} - \sigma\|B\|^2 > \frac{L}{2}$*;*

(ii) $\forall n \in \mathbb{N}, \rho_n \in (0, \delta)$*, where we set* $\delta := 2 - \frac{L}{2}(\frac{1}{\tau} - \sigma\|B\|^2)^{-1} \in [1, 2)$*;*

(iii) $\sum_{i \in \mathbb{N}} \rho_n(\delta - \rho_n) = +\infty$*.*

Then there exists a solution $(x^*, y^*) \in X \times Y$ to (5.3) *such that $\{x_n\}$ and $\{y_n\}$ converge weakly to x^* and y^* respectively.*

Note that $f_\alpha(D\cdot) + \frac{1}{2\lambda}\|\cdot - z\|^2$ can be expressed as (5.1) by making the following identifications:

$$F(x) = \frac{1}{2\lambda}\|x - z\|^2 - \operatorname{env}_\alpha f(Dx),\ G(x) = 0,\ \text{and}\ H(Bx) = f(Dx). \tag{5.7}$$

It is straightforward to see F is convex if $\lambda < \frac{\alpha}{\|D\|^2}$. As the difference of differentiable functions, F is differentiable with gradient

$$\nabla F(x) = \frac{1}{\lambda}(x - z) + D^\top \nabla \operatorname{env}_\alpha f(Dx).$$

To reduce the number of operations, we use the Moreau identity to write $\nabla \operatorname{env}_\alpha f(Dx) = \operatorname{prox}_{\alpha^{-1}f^*}(\alpha^{-1}Dx)$ (see Section 2.4). We summarize the ADMM for (P) in Algorithm 1. To apply Proposition 5.1.1 to Algorithm 1, we first determine the Lipschitz constant of ∇F.

Algorithm 1: Primal-Dual Splitting Algorithm for (P)

Input: Initialization: Choose the positive parameters τ, σ, the sequence of pos- itive relaxation parameters $(\rho_n)_{n\in\mathbb{N}}$ and the initial estimates $x_0 \in X$, $y_0 \in Y$.

for $n = 0, 1, \ldots$ **do**

$$\begin{aligned} x_{n+1}^+ &:= x_n - \tau\left(\frac{1}{\lambda}(x_n - z) - D^\top \operatorname{prox}_{\alpha^{-1}f^*}(\alpha^{-1}Dx_n)\right) - \tau D^\top y_n \\ y_{n+1}^+ &:= \operatorname{prox}_{\sigma f^*}\left(y_n + \sigma D(2x_{n+1}^+ - x_n)\right) \\ (x_{n+1}, y_{n+1}) &:= \rho_n(x_{n+1}^+, y_{n+1}^+) + (1 - \rho_n)(x_n, y_n) \end{aligned}$$

Lemma 5.1.1. *Let F be defined as in* (5.7) *and suppose that $\lambda < \frac{\alpha}{\|D\|^2}$, then F is $\frac{1}{\lambda}$-smooth.*

Proof. We know that

$$\nabla F = \frac{1}{\lambda}(\cdot - z) - D^\top \operatorname{prox}_{\alpha^{-1}(\|\cdot\|_1)^*}(\alpha^{-1}D\cdot).$$

For any x and y in $\mathbb{R}^n$, let us denote $p = \text{prox}_{\alpha^{-1}(\|\cdot\|_1)^*}(\alpha^{-1}Dx)$ and $q = \text{prox}_{\alpha^{-1}(\|\cdot\|_1)^*}(\alpha^{-1}Dy)$. Then, one has

$$\begin{aligned}\|\nabla F(x) - \nabla F(y)\|^2 &= \frac{1}{\lambda^2}\|x-y\|^2 - \frac{2\alpha}{\lambda}\langle \alpha^{-1}D(x-y), p-q\rangle + \|D^\top(p-q)\|^2 \\ &\leq \frac{1}{\lambda^2}\|x-y\|^2 - \frac{2\alpha}{\lambda}\|p-q\|^2 + \|D^\top(p-q)\|^2 \\ &= \frac{1}{\lambda^2}\|x-y\|^2 + (p-q)^\top(DD^\top - \frac{2\alpha}{\lambda}\,\text{Id})(p-q).\end{aligned}$$

The inequality above follows from the nonexpansiveness of the proximity operator. By assumption, $\|D\|^2 < \frac{\alpha}{\lambda}$, so $DD^\top - \frac{2\alpha}{\lambda}\,\text{Id}$ is semi-negative. Thus,

$$\|\nabla F(x) - \nabla F(y)\| \leq \frac{1}{\lambda}\|x-y\|.$$

□

Corollary 5.1.1. *Let λ, α, and z be as in problem* (P)*, and let τ, σ, and the sequence $\{\rho_n\}_{n\in\mathbb{N}}$ be the parameters in Algorithm 1. Suppose that $\lambda < \frac{\alpha}{\|D\|^2}$ and the following hold:*

(i) $\frac{1}{\tau} - \sigma\|D\|^2 > \frac{1}{2\lambda}$;

(ii) $\forall n \in \mathbb{N}, \rho_n \in (0, \delta)$, *where we set* $\delta := 2 - \frac{1}{2\lambda}(\frac{1}{\tau} - \sigma\|D\|^2)^{-1} \in [1, 2)$;

(iii) $\sum_{n\in\mathbb{N}} \rho_n(\delta - \rho_n) = +\infty$.

Let $\{x_n\}$ and $\{y_n\}$ be the sequences produced by Algorithm 1. Then $\{x_n\}$ and $\{y_n\}$ converge weakly to a primal solution x^* and a dual solution y^* respectively.

We remark here that Primal-Dual Splitting as in (5.4)–(5.6) extends the Douglas-Rachford Splitting method [24] of which the Alternating Directions Method of Multipliers is a special

case [21]. However ADMM may converge even when the model is nonconvex, as we discuss in Section 5.3.

5.2 Difference of Convex

The general difference of convex (DC) problem and difference of convex algorithm (DCA) was introduced and extensively developed by Le Thi et al [31, 38, 30]. Many applications involve nonconvex functions which can be decomposed as a difference of convex functions. A generic DC problem has the form

$$\operatorname{argmin}\{G(x) - H(x) : x \in X\} \tag{5.8}$$

where $G, H \in \Gamma_0(X)$. The DC dual problem is

$$\operatorname{argmin}\{H^*(y) - G^*(y) : y \in Y\}, \tag{5.9}$$

where H^* and G^* are the conjugates of H and G respectively.

Of course, any decomposition of a DC function $F \coloneqq G - H$ is not unique; for example, we can force strong convexity in each term by writing $F(x) = \big(G(x) + \frac{\rho}{2}\|x\|^2\big) - \big(H(x) + \frac{\rho}{2}\|x\|^2\big)$. Following the convention in [30], we define the modulus of strong convexity of a function G as the largest $\rho > 0$ such that $G - \frac{\rho}{2}\|\cdot\|^2$ is strongly convex and denote this by $\rho(G)$. Similarly, we denote by $\rho(G, C)$ the modulus of strong convexity of G on the set C.

Given an initial point $x_0 \in X$, DCA iterates by solving the following first order approxi-

mations of (5.8) and (5.9):

$$y_n \in \operatorname{argmin}\{H^*(y) - (G^*(y_{n-1}) + \langle x_n, y - y_{n-1}\rangle) : y \in Y\} \tag{5.10}$$

$$x_{n+1} \in \operatorname{argmin}\{G(x) - (H(x_n) + \langle x - x_n, y_n\rangle) : x \in X\}. \tag{5.11}$$

Now because y_n minimizes (5.10), we must have $0 \in \partial H^*(y_n) - x_n$, i.e. $x_n \in \partial H^*(y_n)$. Recall from Section 2.3 that this is equivalent to $y_n \in \partial H(x_n)$. Similar computations can be applied to (5.11), simplifying the above to:

$$y_n \in \partial H(x_n)$$

$$x_{n+1} \in \partial G^*(y_n).$$

We summarize the convergence analysis for (5.10)–(5.11).

Proposition 5.2.1 ([30, Theorem 3.3])**.** *Let C and D be two convex sets containing the sequences $\{x_n\}$ and $\{y_n\}$ respectively.*

(i) The sequences $\{G(x_n) - H(x_n)\}$ and $\{H^*(y_n) - G^*(y_n)\}$ are decreasing and converge to the same limit. If G or H is strictly convex on C, then $\{x_n\}$ converges in finite steps. Similarly, if H^* or G^* is strictly convex on D, then $\{y_n\}$ converges in finite steps.

(ii) If $\rho(G, C) + \rho(H, C) > 0$ (resp. $\rho(G^*, D) + \rho(H^*, D) > 0$), then the series $\{\|x_{n+1} - x_n\|^2\}$ (resp. $\{\|y_{n+1} - y_n\|^2\}$) is convergent.

(iii) If the optimal value of (5.8) is finite and the sequences $\{x_n\}$ and $\{y_n\}$ are bounded, then every limit point $\tilde{x}$ (resp. $\tilde{y}$) of the sequence $\{x_n\}$ (resp. $\{y_n\}$) is a critical point of $G - H$ (resp. $H^* - G^*$).

Proposition 5.2.2 ([30, Theorem 3.4]). *Let $\{x_n\}$ and $\{y_n\}$ be the sequences generated by DCA. Then the following properties hold:*

(i) Suppose that the DC function $F := G - H$ is subanalytic; $\mathrm{dom}(F)$ is closed; $f|_{\mathrm{dom}(F)}$ is continuous; and around every critical point of (5.8), either G or H is differentiable with locally Lipschitz derivative. Assume that $\rho := \rho(G) + \rho(H) > 0$. If either the sequence $\{x_n\}$ or $\{y_n\}$ is bounded, then $\{x_n\}$ and $\{y_n\}$ are convergent to critical points of (5.8) and (5.9) respectively.

(ii) Similarly, if $H^* - G^*$ is subanalytic; $\mathrm{dom}(H^* - G^*)$ is closed; $(H^* - G^*)|_{\mathrm{dom}(H^* - G^*)}$ is continuous; and around critical points of (5.9), either G^* or H^* is differentiable with locally Lipschitz derivative. If $\rho(G^*) + \rho(H^*) > 0$ and either sequence $\{x_n\}$ or $\{y_n\}$ is bounded, then $\{x_n\}$ and $\{y_n\}$ are convergent to critical points of (5.8) and (5.9) respectively.

We decompose (P) as a DC problem by identifying $G = \frac{1}{2\lambda}\|\cdot - z\|^2 + f(D\cdot)$ and $H = \mathrm{env}_\alpha f(D\cdot)$. In this case, the primal DC problem (5.8) becomes

$$\mathrm{argmin}\left\{(f \circ D + \frac{1}{2\lambda}\|\cdot - z\|^2)(x) - \mathrm{env}_\alpha f(Dx) : x \in X\right\} \tag{5.12}$$

and the dual problem becomes

$$\mathrm{argmin}\left\{(\mathrm{env}_\alpha f \circ D)^*(y) - (f \circ D + \frac{1}{2\lambda}\|\cdot - z\|^2)^*(y) : y \in Y\right\}. \tag{5.13}$$

In this case, DCA can be computed by Algorithm 2. Note that the first inclusion becomes equality because $\mathrm{env}_\alpha f$ is differentiable. As in the previous section, we write $\nabla \mathrm{env}_\alpha f(Dx) = \mathrm{prox}_{\alpha^{-1} f^*}(\alpha^{-1} D x_n)$.

Algorithm 2: Difference of Convex Algorithm (DCA) for (P)

Input: Choose initial estimate $\mathrm{x}_0 \in X$

for $i = 0, 1, \ldots$ **do**

$$y_n = D^\top \operatorname{prox}_{\alpha^{-1} f^*}\left(\alpha^{-1} D x_n\right) \tag{5.14}$$

$$x_{n+1} \in \operatorname{argmin}\left\{\frac{1}{2\lambda}\|x - z\|^2 + f(Dx) - \langle x, y_k\rangle : x \in X\right\} \tag{5.15}$$

Applying results from Chapter 3, Proposition 5.2.1, and Proposition 5.2.2, the convergence analysis for Algorithm 2 is given by the following theorem.

Theorem 5.2.1. *Let $\{x_n\}$ and $\{y_n\}$ be the sequences generated by Algorithm 2, and let $G = \frac{1}{2\lambda}\|\cdot - z\|^2 + f(D\cdot)$ and $H = \operatorname{env}_\alpha f(D\cdot)$. The sequences $\{G(x_n) - H(x_n)\}$ and $\{H^*(y_n) - G^*(y_n)\}$ decrease to the same limit, and the sequence $\{x_n\}$ converges in finite steps.*

Furthermore, if f is subanalytic with closed domain such that $f|_{\mathrm{dom}(f)}$ is continuous, then if either sequence $\{x_n\}$ or $\{y_n\}$ is bounded, then they converge to critical points of (5.8) *and* (5.13) *respectively.*

Recent work shows that DCA can be boosted by taking advantage of the differentiability of the entire objective $G - H$ [2] or of the first term G [3]. Both methods accelerate convergence by introducing a backtracking line search in the direction $y_n - x_n$ at each step. Based on results in Section 3.3, we have hope that while the original problem is certainly not differentiable, the dual problem may be. We give two necessary conditions for the dual problem to be differentiable below. Both rely on the fact that the conjugate of a function is essentially smooth if and only if the function is essentially strictly convex [44].

Lemma 5.2.1. *(i) If* $\mathrm{env}_\alpha f \circ D$ *is strictly convex, then the dual objective*

$$(\mathrm{env}_\alpha f \circ D)^*(y) - (f \circ D + \frac{1}{2\lambda}\|\cdot - z\|^2)^*(y)$$

is differentiable.

(ii) If we decompose the primal problem as $G + \frac{\rho}{2}\|\cdot\|^2 - H - \frac{\rho}{2}\|\cdot\|^2$*, then the dual problem* $(H + \frac{\sigma}{2}\|\cdot\|^2)^* - (G + \frac{\sigma}{2}\|\cdot\|^2)^*$ *is differentiable.*

Theoretically, we can apply boosted DCA to our problem (5.13) to find both primal and dual solutions. However, it may be very difficult to compute the terms H^* and G^* in this case, especially if D is not invertible. We discuss future plans to explore this topic in the Conclusion.

5.3 Alternating Direction Method of Multipliers

The Alternating Direction Method of Multipliers (ADMM), as introduced by Gabay and Mercier in [24], is a proximal splitting algorithm for finding zeros of monotone operators and, like Primal-Dual splitting, is a special case of the Douglas-Rachford splitting method [20]. In the context of optimization, the monotone operators in question are the subdifferentials of convex functions. To illustrate the method, we first consider the generic constrained optimization problem

$$\min_{(x,y)\in X\times Y} F(x) + G(y)$$
$$\text{subject to } \; y = Bx.$$

Recall that the augmented Lagrangian with parameter $\eta > 0$ is

$$\mathcal{L}_\eta(x, y, d) \coloneqq F(x) + G(y) - \langle d, Bx - y\rangle + \frac{\eta}{2}\|Bx - y\|^2.$$

Given some starting point (x_0, d_0), the ADMM algorithm iterates as follows:

$$y_{n+1} \in \operatorname{argmin}\{\mathcal{L}_\eta(x_n, y, d_n) : y \in Y\}, \tag{5.16}$$

$$x_{n+1} = \operatorname{argmin}\{\mathcal{L}_\eta(x, y_{n+1}, d_n) : x \in X\}, \tag{5.17}$$

$$d_{n+1} = d_n - \eta(Bx_{n+1} - y_{n+1}). \tag{5.18}$$

Here we assume that F and G are continuous and subdifferentiable, but we do not require any convexity.

To apply this method to (P), we first reformulate the problem as

$$\begin{aligned} &\min_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m} \ \frac{1}{2\lambda}\|x - z\|^2 + f_\alpha(y) \\ &\text{subject to} \ \ y = Dx. \end{aligned} \tag{5.19}$$

The augmented Lagrangian with parameter $\eta > 0$ for the constrained problem (5.19) is

$$\mathcal{L}_\eta(x, y, d) \coloneqq \frac{1}{2\lambda}\|x - z\|^2 + f_\alpha(y) - \langle d, Dx - y\rangle + \frac{\eta}{2}\|Dx - y\|^2. \tag{5.20}$$

As above, we assume that the parent function f (and therefore f_α) is continuous. Moreover, for technical reasons, we assume that the smallest eigenvalue of $DD^\top$, denoted $\lambda_{\min}(DD^\top)$, is nonzero. Now the algorithm (5.16)–(5.18) can be written as follows.

Algorithm 3: Proximal ADMM for (5.19) (i.e., (P))

Input: Initialization: Input $(x^{(0)}, d^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^m$ and $\eta > 0$.

for $n = 0, 1, \ldots$ **do**

$$y_{n+1} \in \operatorname{prox}_{\frac{1}{\eta} f_\alpha}(Dx_n - \frac{1}{\eta} d_n), \tag{5.21}$$

$$x_{n+1} = \left(\frac{1}{\lambda}\operatorname{Id} + \eta D^\top D\right)^{-1} \left(\frac{1}{\lambda} z + D^\top (d_n + \eta y_{n+1})\right), \tag{5.22}$$

$$d_{n+1} = d_n - \eta(Dx_{n+1} - y_{n+1}). \tag{5.23}$$

Lemma 5.3.1. *Let $\{x_k, y_k, d_k\}$ be generated by Algorithm 3. Then we have*

$$x_{k+1} = z + \lambda D^\top d_{k+1},$$
$$y_{k+1} = Dx_{k+1} + \frac{1}{\eta}(d_{k+1} - d_k).$$

Proof. The expression for y_{k+1} follows immediately from (5.23).

Now since $x_{k+1} = \arg\min_x \mathcal{L}_\eta(x_k, y_k, d_k)$, we must have

$$0 = \frac{1}{\lambda}(x_{k+1} - z) - D^\top d_k + \eta D^\top (Dx_{k+1} - y_k).$$

Again by (5.23), this is equivalent to

$$0 = \frac{1}{\lambda}(x_{k+1} - z) - D^\top d_{k+1}.$$

The result follows. □

We will show that the sequence $\{x_n\}$ converges to a stationary point of the (P), but first we require several technical lemmas.

Lemma 5.3.2. *Let $\{x_k, y_k, d_k\}$ be the sequence generated by Algorithm 3. If*

$$\lim_{k\to\infty} \|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2 + \|d_{k+1} - d_k\|^2 = 0 \tag{5.24}$$

and the sequence has a cluster point (x^*, y^*, d^*), then x^* is a stationary point of (P).

Proof. Because (x^*, y^*, d^*) is a cluster point of $\{x_k, y_k, d_k\}$, there is a subsequence $\{x_{k_j}, y_{k_j}, d_{k_j}\}$ such that

$$\lim_{j\to\infty} \|x_{k_j} - x^*\|^2 = 0, \ \lim_{j\to\infty} \|y_{k_j} - y^*\|^2 = 0, \ \lim_{j\to\infty} \|d_{k_j} - d_{k_j}\|^2 = 0.$$

Now since y_{k_j} is a solution of $\arg\min_y \mathcal{L}_\eta(x_{k_j-1}, y, d_{k_j-1})$, we get

$$0 \in \partial f_\alpha(y_{k_j}) + d_{k_j-1} + \eta(y_{k_j} - Dx_{k_j-1}).$$

Using (5.23), the above inclusion becomes

$$0 \in \partial f_\alpha(y_{k_j}) + d_{k_j} + \eta D(x_{k_j} - x_{k_j-1}).$$

By taking $j \to \infty$, this becomes $0 \in \partial f_\alpha(y^*) + d^*$.Applying Lemma 5.3.1 and letting j approach infinity, we see that

$$D^\top d^* = \frac{1}{\lambda}(x^* - z) \ \text{ and } \ y^* = Dx^*.$$

Therefore $0 \in \frac{1}{\lambda}(x^* - z) + D^\top \partial f_\alpha(Dx^*)$, i.e. x^* is a stationary point of the problem.

□

The convergence analysis for Algorithm 3 is standard and closely follows both [32] and [52]. In particular, (P) is a special case of the problem considered in [52], which allows us to greatly simplify the analysis. As a result of Lemma 5.3.1, we can bound the dual updates using the primal updates.

Lemma 5.3.3. *Let $\{x_k\}$ and $\{d_k\}$ be the sequences defined by Algorithm 3. Assume that $\sigma \coloneqq \lambda_{\min}(DD^\top) > 0$. Then for every $k \in \mathbb{N}$, $\|d_{k+1} - d_k\|^2 \leq \frac{1}{\sigma\lambda^2}\|x_{k+1} - x_k\|^2$.*

Proof. By Lemma 5.3.1, we see that $D^\top(d_{k+1} - d_k) = \frac{1}{\lambda}(x_{k+1} - x_k)$, so we can control the convergence of $\{d_k\}$ through $\{x_k\}$. Because $\sigma > 0$, we get

$$\sigma\|d_{k+1} - d_k\|^2 \leq \|D^\top(d_{k+1} - d_k)\|^2 \leq \frac{1}{\lambda^2}\|x_{k+1} - x_k\|^2. \tag{5.25}$$

□

In order to show that the sequence $\{\mathcal{L}_\eta(x_k, y_k, d_k)\}$ is decreasing, we provide descent guarantees for the y update step in Lemma 5.3.4 and the x and d updates in Lemma 5.3.5.

Lemma 5.3.4. *Let $\{x_k, y_k, d_k\}$ be the sequence generated by Algorithm 3. Then for every $k \in \mathbb{N}$ we have*

$$\mathcal{L}_\eta(x_k, y_{k+1}, d_k) - \mathcal{L}_\eta(x_k, y_k, d_k) \leq -\frac{(\eta - \frac{1}{\alpha})}{2}\|y_{k+1} - y_k\|^2.$$

Proof. For ease of notation, we set $\mathcal{L}_{\hat{y}}(y) = \mathcal{L}_\eta(x_k, y, d_k)$. From (5.20), we have

$$\mathcal{L}_{\hat{y}}(y_k) - \mathcal{L}_{\hat{y}}(y_{k+1}) = f_\alpha(y_k) + \frac{\eta}{2}\|Dx_k - y_k\|^2 - f_\alpha(y_{k+1}) - \frac{\eta}{2}\|Dx_k - y_{k+1}\|^2 - \langle d_k, y_{k+1} - y_k\rangle. \tag{5.26}$$

Because f_α is $\frac{1}{\alpha}$-semiconvex, we have the following subgradient inequality

$$f_\alpha(y_k) + \frac{\eta}{2}\|Dx_k - y_k\|^2 \geq f_\alpha(y_{k+1}) + \frac{\eta}{2}\|Dx_k - y_{k+1}\|^2 + \langle \xi, y_k - y_{k+1}\rangle + \frac{(\eta - \frac{1}{\alpha})}{2}\|y_{k+1} - y_k\|^2$$

where $\xi \in \partial f_\alpha(y_{k+1}) - \eta(Dx_k - y_{k+1})$. Then (5.26) becomes

$$\mathcal{L}_{\hat{y}}(y_k) - \mathcal{L}_{\hat{y}}(y_{k+1}) \geq \langle \xi, y_{k+1} - y_k\rangle - \frac{(\eta - \frac{1}{\alpha})}{2}\|y_{k+1} - y_k\|^2.$$

Because $y_{k+1} \in \arg\min \mathcal{L}_\eta(x_k, y, d_k)$, we know that $0 \in \partial f_\alpha(y_{k+1}) - \eta(Dx_k - y_{k+1}) + d_k$. That is, $-d_k \in \partial f_\alpha(y_{k+1}) - \eta(Dx_k - y_{k+1})$. Combining this with (5.26) we see that

$$\mathcal{L}_{\hat{y}}(y_k) - \mathcal{L}_{\hat{y}}(y_{k+1}) \geq \frac{(\eta - \frac{1}{\alpha})}{2}\|y_{k+1} - y_k\|^2. \tag{5.27}$$

□

Lemma 5.3.5. *Let $\{x_k, y_k, d_k\}$ be the sequence generated by Algorithm 3. Then for every $k \in \mathbb{N}$ we have*

$$\mathcal{L}_\eta(x_k, y_{k+1}, d_k) - \mathcal{L}_\eta(x_{k+1}, y_{k+1}, d_{k+1}) \geq \left(\frac{1}{2\lambda} + \frac{\eta}{2}\lambda_{\min}(D^\top D) - \frac{1}{\eta\sigma\lambda^2}\right)\|x_k - x_{k+1}\|^2.$$

Proof. For ease of notation, we let $h(x) = \frac{1}{2\lambda}\|x - z\|^2$ and denote $(x_{k+1}, y_{k+1}, d_{k+1}) = (x_+, y_+, d_+)$.

The difference $\mathcal{L}_\eta(x_k, y_k, d_k) - \mathcal{L}_\eta(x_+, y_+, d_+)$ is

$$h(x_k) - h(x_+) + \underbrace{\langle -d_k, Dx_k - y_+\rangle}_{a_1} + \underbrace{\langle d_+, Dx_+ - y_+\rangle}_{a_2} + \underbrace{\frac{\eta}{2}\|Dx_k - y_+\|^2}_{a_3} - \underbrace{\frac{\eta}{2}\|Dx_+ - y_+\|^2}_{a_4} \tag{5.28}$$

Rewriting $d_k = d_+ + \eta(Dx_+ - y_+)$ (5.23), we see that

$$a_1 + a_2 = \langle d_+, Dx_+ - Dx_k\rangle - \eta\langle Dx_+ - y_+, Dx_k - y_+\rangle.$$

Now, by completing the square, we get that

$$a_1 + a_2 + a_3 = \langle d_+, Dx_+ - Dx_k\rangle + \frac{\eta}{2}\|Dx_+ - Dx_k\|^2 - \frac{\eta}{2}\|Dx_+ - y_+\|^2.$$

Noting that $Dx_+ - y_+ = \frac{1}{\eta}(d_+ - d_k)$, we get

$$a_1 + a_2 + a_3 + a_4 = \langle d_+, Dx_+ - Dx_k\rangle + \frac{\eta}{2}\|Dx_+ - Dx_k\|^2 - \frac{1}{\eta}\|d_+ - d_k\|^2.$$

By Lemma 5.3.1, we know that $\nabla h(x_+) = \frac{1}{\lambda}(x_+ - z) = D^\top d_+$, so (5.28) can be rewritten

$$\underbrace{h(x_k) - h(x_+) - \langle \nabla h(x_+), Dx_k - Dx_+\rangle}_{b_1} + \underbrace{\frac{\eta}{2}\|Dx_+ - Dx_k\|^2}_{b_2} - \underbrace{\frac{1}{\eta}\|d_+ - d_k\|^2}_{b_3},$$

noting the sign change due to flipping Dx_k and Dx_+.

Since h is strongly convex, we know that $b_1 \geq \frac{1}{2\lambda}\|x_k - x_+\|^2$. We also know that $b_2 \geq \frac{\eta}{2}\lambda_{\min}(D^\top D)\|x_+ - x_k\|^2$. Finally, we apply the bound from Lemma 5.3.3, and get

$$b_1 + b_2 + b_3 \geq \left(\frac{1}{2\lambda} + \frac{\eta}{2}\lambda_{\min}(D^\top D) - \frac{1}{\eta\sigma\lambda^2}\right)\|x_k - x_+\|^2.$$

□

Remark 5.3.1. While we assume that $\lambda_{\min}(DD^\top) > 0$, we make no such assumption on $\lambda_{\min}(D^\top D)$. For the applications discussed in Chapter 6, we will have $\lambda_{\min}(D^\top D) = 0$.

Finally, Proposition 5.3.1 and Theorem 5.3.1 provide the convergence analysis for Algorithm 3.

Proposition 5.3.1. *Let $\{x_k, y_k, d_k\}$ be the sequence generated by Algorithm 3. If $\sigma := \lambda_{\min}(DD^\top) > 0$ and the parameter η satisfies $\eta \geq \max\{\frac{2}{\sigma\lambda}, \frac{1}{\alpha}\}$, then the following statements hold:*

(i) The sequence $\{\mathcal{L}_\eta(x_k, y_k, d_k)\}$ is decreasing.

(ii) The sequence $\{x_k, y_k, d_k\}$ has a convergent subsequence.

(iii) The sequence $\{x_k, y_k, d_k\}$ satisfies (5.24).

Proof. Item (i): Add together the results of Lemmas 5.3.4 and 5.3.5 to get

$$\begin{aligned}&\mathcal{L}_\eta(x_k, y_k, d_k) - \mathcal{L}_\eta(x_{k+1}, y_{k+1}, d_{k+1}) \\ &\qquad \geq \left(\frac{1}{2\lambda} + \frac{\eta}{2}\lambda_{\min}(D^\top D) - \frac{1}{\sigma\lambda^2\eta}\right)\|x_k - x_{k+1}\|^2 + \frac{\eta - \frac{1}{\alpha}}{2}\|y_k - y_{k+1}\|^2.\end{aligned}$$

By the assumption on η, we see that the above is greater than zero.

Item (ii): By Lemma 5.3.1, $D^\top d_k = \frac{1}{\lambda}(x_k - z)$. It follows that $\sigma\|d_k\|^2 \leq \|D^\top d_k\|^2 \leq \frac{1}{\lambda^2}\|x_k - z\|^2$. Therefore, the boundedness of $\{x_k\}$ implies the boundedness of $\{d_k\}$. The boundedness of $\{y_k\}$ can be derived from the monotonicity of $\mathcal{L}_\eta$ and the semiconvexity of f_α. By item (i),

$$\begin{aligned}\mathcal{L}_\eta(x_0, y_0, d_0) &\geq \mathcal{L}_\eta(x_k, y_k, d_k) \\ &= \frac{1}{2\lambda}(1 - \frac{1}{\sigma\eta\lambda})\|x_k - z\|^2 + f_\alpha(y_k) + \frac{\eta}{2}\|Dx_k - y_k - \frac{d_k}{\eta}\|^2.\end{aligned} \tag{5.29}$$

Now since $y_{k+1} \in \operatorname{prox}_{\frac{1}{\eta} f_\alpha}(Dx_k - \frac{d_k}{\eta})$,

$$f_\alpha(y_k) + \frac{\eta}{2}\|Dx_k - \frac{d_k}{\eta} - y_k\|^2 \geq f_\alpha(y_{k+1}) + \frac{\eta}{2}\|Dx_k - \frac{d_k}{\eta} - y_{k+1}\|^2 + \frac{(\eta - \frac{1}{\lambda})}{2}\|y_{k+1} - y_k\|^2$$
$$\geq \frac{(\eta - \frac{1}{\lambda})}{2}\|y_{k+1} - y_k\|^2.$$

Then

$$\mathcal{L}_\eta(x_k, y_k, d_k) \geq \frac{1}{2}(-\frac{1}{\sigma\eta\lambda^2} + \eta)\min\{\|x_k - z\|^2, \|y_{k+1} - y_k\|^2\} \geq 0.$$

Item (iii): Suppose $\{x_{k_j}, y_{k_j}, d_{k_j}\}$ is a convergent subsequence of $\{x_k, y_k, d_k\}$ such that

$$\lim_{j\to\infty} \|x_{k_j} - x^*\|^2 = 0, \ \lim_{j\to\infty} \|y_{k_j} - y^*\|^2 = 0, \ \lim_{j\to\infty} \|d_{k_j} - d^*\|^2 = 0.$$

The continuity of $\mathcal{L}_\eta$ yields

$$\lim_{j\to\infty} \mathcal{L}_\eta(x_{k_j}, y_{k_j}, d_{k_j}) = \mathcal{L}_\eta(x^*, y^*, d^*) > -\infty,$$

From item (ii), (5.29), and $\eta \geq \max\{\frac{1}{\sigma\lambda}, \frac{1}{\alpha}\}$,

$$\mathcal{L}_\eta(x_0, y_0, d_0) - \mathcal{L}_\eta(x_{k_j}, y_{k_j}, d_{k_j})$$
$$\geq \left(\frac{1}{2\lambda} + \frac{\eta}{2}\lambda_{\min}(D^\top D) - \frac{1}{\sigma\lambda^2\eta}\right)\sum_{i=0}^{k_j-1}\|x_i - x_{i+1}\|^2 + \frac{\eta - \frac{1}{\alpha}}{2}\sum_{i=0}^{k_j-1}\|y_i - y_{i+1}\|^2.$$

Thus $\sum_{i=0}^{\infty}\|x_{i+1} - x_i\|^2$ and $\sum_{i=0}^{\infty}\|y_{i+1} - y_i\|^2$ converge.

Finally, this implies $\lim_{i\to\infty}\|x_{i+1} - x_i\|^2 = 0$ and $\lim_{i\to\infty}\|y_{i+1} - y_i\|^2 = 0$, and from (5.25), $\lim_{i\to\infty}\|d_{i+1} - d_i\|^2 = 0$. □

Theorem 5.3.1. *Suppose that* $\sigma \coloneqq \lambda_{\min}(DD^\top) > 0$ *and* $\eta \geq \max\{\frac{1}{\sigma\lambda}, \frac{1}{\alpha}\}$. *Then the sequence*

$\{x_k, y_k, d_k\}$ converges to (x^*, y^*, d^*) and x^* is a stationary point of (P).

Moreover, $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < \infty$.

Proof. Therefore the sequence $\{x_k, y_k, d_k\}$ converges, denote this limit point (x^*, y^*, d^*). It follows from Lemma 5.3.2 and item (iii) of Proposition 5.3.1 that x^* is a stationary point of (P).

Let $\mathcal{L}_\eta(x^*, y^*, d^*) = \ell^*$. Note that $\lim_{i\to\infty} \mathcal{L}_\eta(x_i, y_i, d_i) = \ell^*$. If $\mathcal{L}_\eta(x_i, y_i, d_i) = \ell^*$ for some i, we show that Algorithm 3 terminates in $i+1$ iterations. Actually, by the monotonicity of $\{\mathcal{L}_\eta(x_k, y_k, d_k)\}$, we must have $\mathcal{L}_\eta(x_j, y_j, d_j) = \ell^*$ for all $j \geq i$. Otherwise, infinitely many terms of the convergent sequence $\{\mathcal{L}_\eta(x_k, y_k, d_k)\}$ will be less than ℓ^*, which is a contradiction. We conclude from Proposition 5.3.1 (i) and $\eta \geq \max\{\frac{1}{\sigma\lambda}, \frac{1}{\alpha}\}$, we conclude that $x_i = x_j$ for $j \geq i$. As a consequence, we have $d_i = d_j$ and $y_{i+1} = y_{j+1}$.

□

Chapter 6

Applications

In this chapter, we study the application of our structured sparsity promoting functions to the problems of signal and image denoising. The total variation denoising (TVD) model is given by

$$\operatorname{argmin}\{\frac{1}{2\lambda}\|x-z\|^2 + \|x\|_{TV} : x \in X\}. \tag{TVD}$$

Here X is either $\mathbb{R}^n$ (1D signals) or $\mathbb{R}^{m\times n}$ (grayscale images), z is the noisy input, and $\|\cdot\|_{TV}$ is the corresponding total variation function which we define in each section. Variational problems are widespread in image processing due to their sensitivity to geometric features of images, and these geometric features can be captured by sparsity in the appropriate basis.

By a slight abuse of notation, we modify this problem by replacing $\|\cdot\|_{TV}$ with $(\|\cdot\|_{TV})_\alpha$:

$$\operatorname{argmin}\{\frac{1}{2\lambda}\|x-z\|^2 + (\|\cdot\|_{TV})_\alpha(x) : x \in X\}. \tag{TVD-α}$$

This is a special case of the problem (P) discussed in Chapter 5 and can therefore be solved using the algorithms described there. We give numerical results comparing the performance

of (TVD) against (TVD-α) and offer insight into the parameter choices for these models.

6.1 Total Variation Denoising: Signals

In the case of one-dimensional signals $x \in \mathbb{R}^n$, the TV function $\|x\|_{TV} = \|Dx\|_1$, where $D \in \mathbb{R}^{m\times n}$ is the difference matrix defined by

$$D = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

which measures the difference between one measurement of the signal and the next. Then the problem (TVD) becomes

$$\operatorname{argmin}\ \{\frac{1}{2\lambda}\|x-z\|_2^2 + \|Dx\|_1 : x \in \mathbb{R}^n\}, \tag{6.1}$$

where $z \in \mathbb{R}^n$ is a noisy signal. From our previous discussion of $\|\cdot\|_1$ as a sparsity promoting function, it is clear that the TV penalty removes small fluctuations from the signal. The parameter λ both controls how closely the solution $\hat{x}$ fits the noisy data z and determines the threshold of what is considered noise.

We propose replacing the ℓ_1-norm with the function $(\|\cdot\|_1)_\alpha(x) = \|x\|_1 - \operatorname{env}_{\alpha\|\cdot\|_1}(x)$. Then the problem (TVD-$\alpha$) becomes

$$\arg\min\ \{\frac{1}{2\lambda}\|x-z\|_2^2 + (\|\cdot\|_1)_\alpha(Dx) : x \in \mathbb{R}^n\}. \tag{6.2}$$

As above, λ controls how closely $\hat{x}$ must fit the data z, but the parameter α allows greater

customization of the thresholding behavior (see Section 4.4).

Note that, for the given matrix D, $\sigma := \lambda_{\min}(D^\top D) = 2-2\cos(\frac{\pi}{n})$ where n is the length of the signal z. For n large this is approximately $4(\frac{\pi}{2n})^2$, and the bound $\eta > \frac{1}{\sigma\lambda}$ for Algorithm 3 is not optimal. Experimental results in [32] suggest the much smaller $\eta = \frac{1}{n\lambda\sigma} \approx \frac{1}{\lambda\sqrt{\sigma}}$, which we use in our experiments.

We examine the performance of Algorithm 1 (PD), Algorithm 2 (DCA), and Algorithm 3 (ADMM) for this model on piecewise constant signals with added Gaussian noise. Table 2 contains the average recovery error over 25 samples for each method. We compare the performance of the convex model ($\alpha = 1.1\lambda\|D\|^2$) and the nonconvex model ($\alpha = 0.5\lambda\|D\|^2$). We note that the nonconvex model generally outperforms both the convex model and the ℓ_1 model.

Table 2: Relative recovery error of L1-PD, PD, DCA, and ADMM methods for signal denoising.

λ	L1-PD	DCA	ADMM	PD	DCA	ADMM	PD
		$\alpha = 0.5\lambda\|D\|^2$			$\alpha = 1.1\lambda\|D\|^2$		
White Gaussian noise with standard deviation 0.25							
3	0.0273	0.1767	0.0167	0.0221	0.0302	0.0285	0.0315
4	0.0316	0.0215	0.0187	0.0278	0.0220	0.0199	0.0235
5	0.0322	0.0208	0.0174	0.0280	0.0319	0.0287	0.0347
White Gaussian noise with standard deviation 0.50							
5	0.0459	0.0351	0.0356	0.0416	0.0464	0.0468	0.0489
6	0.0545	0.0434	0.0423	0.0503	0.0415	0.0403	0.0443
7	0.0509	0.0387	0.0353	0.0469	0.0665	0.0657	0.0713

Figure 16 shows the convergence of each algorithm in terms of the true recovery error. While DCA is the slowest of the four, it converges in only a few iterations. We see that while L1-PD and PD achieve a low recovery error relatively quickly, ADMM eventually achieves a lower recovery error. Finally, Figure 17 shows examples of these results for a given signal.

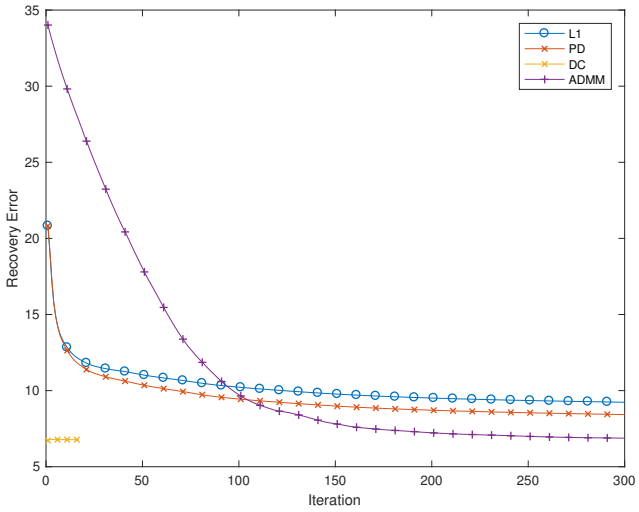
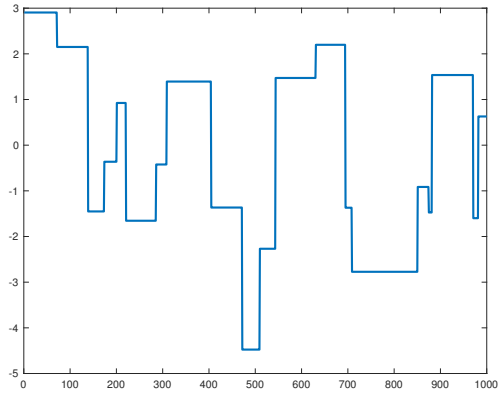
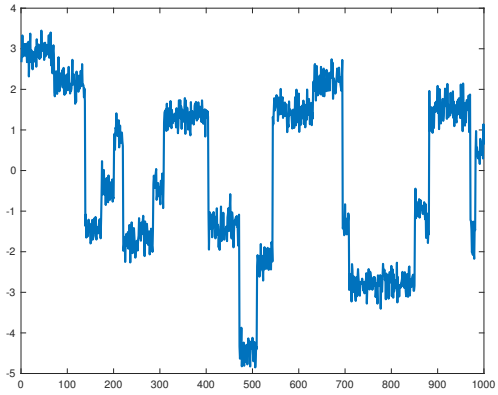


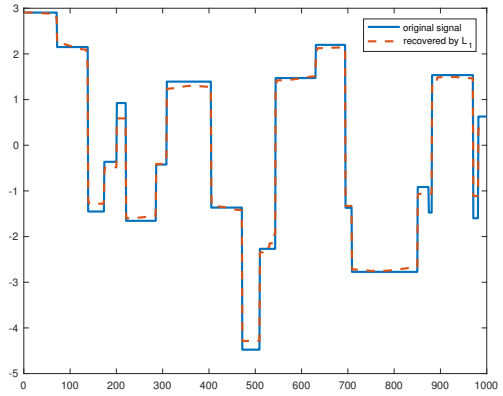
Figure 16: Recovery error of L1-PD, DC, ADMM, and PD vs. the number of iterations.



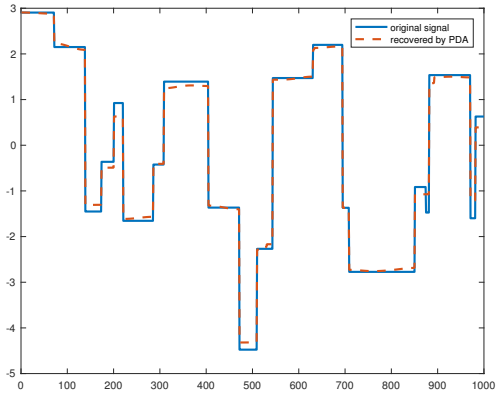
(a)



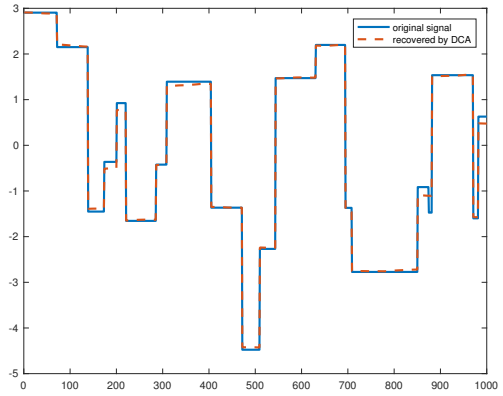
(b)



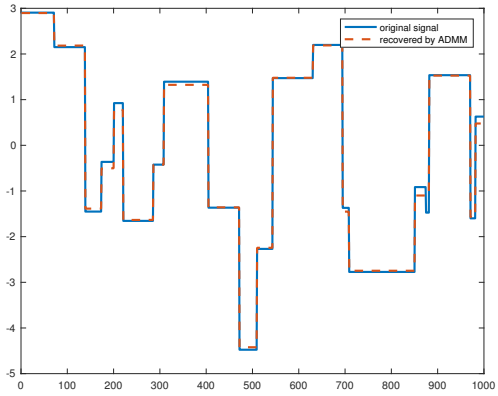
(c)



(d)



(e)



(f)

Figure 17: Figure (a) is the original signal, and figure (b) is the noisy signal. Figure (c) shows the recovered signal using the usual $\|\cdot\|_1$ penalty for PDA. The remaining figures show the results of the $(\|\cdot\|_1)_\alpha$ penalty for (d) PDA, (c) DCA, and (e) ADMM.

6.2 Total Variation Denoising: Images

To apply our models to the problem of image denoising, we must specify the appropriate function $\|\cdot\|_{TV}$. We begin by extending the difference matrix D to the two-dimensional case. Let B denote the $N \times N$ matrix defined by the equation

$$B := \begin{bmatrix} 0 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

and let D be the $2N^2 \times N^2$ matrix given by

$$D := \begin{bmatrix} I_N \otimes B \\ B \otimes I_N \end{bmatrix} \tag{6.3}$$

where I_N is the $N \times N$ identity matrix and the notation $P \otimes Q$ denotes the Kronecker product of matrices P and Q.

Let u be an image in $\mathbb{R}^{N^2}$. We choose $f : \mathbb{R}^{2N^2} \to \mathbb{R}$ as

$$f(z) := \sum_{i=1}^{N^2} \left\| \begin{bmatrix} z_i \\ z_{N^2+i} \end{bmatrix} \right\|_2, \quad z \in \mathbb{R}^{2N^2}. \tag{6.4}$$

With the function f and D given above, $f \circ D(u) = \|u\|_{TV}$ is the well-known Rudin-Osher-Fatemi total variation, which measures the two dimensional variation in pixel values.

The corresponding function f_α is given by Lemma 6.2.1, and its proximity operator is given by Proposition 6.2.2. But first, we compute $(\|\cdot\|_2)_\alpha$ and its proximity operator.

Proposition 6.2.1. *Let $q = \|\cdot\|_2$. Then, it holds for any $x \in \mathbb{R}^d$ that*

$$\operatorname{prox}_{\beta q_\alpha}(x) = \operatorname{prox}_{\beta|\cdot|_\alpha}(\|x\|_2)\frac{x}{\|x\|_2}.$$

Proof. First, a direct computation gives

$$q_\alpha(x) := \begin{cases} \|x\|_2 - \frac{1}{2\alpha}\|x\|_2^2, & \text{if } \|x\|_2 \leq \alpha; \\ \frac{1}{2}\alpha, & \text{otherwise.} \end{cases}$$

Clearly, $q_\alpha(x) = |\cdot|_\alpha(\|x\|_2)$. Therefore, the result holds. □

Lemma 6.2.1. *For the function f defined in* (6.4)*, we have that for $z \in \mathbb{R}^{2N^2}$*

$$f_\alpha(z) = \sum_{i=1}^{N^2} q_\alpha\left(\begin{bmatrix} z_i \\ z_{N^2+i} \end{bmatrix}\right),$$

where $q = \|\cdot\|_2$ the ℓ_2 norm of $\mathbb{R}^2$.

Proof. This is a direct consequence from (6.4) and the definition of proximity operator. □

Proposition 6.2.2. *Let the function f be defined in* (6.4)*, and let α and β be two positive parameters. If $y = \operatorname{prox}_{\beta f_\alpha}(z)$ for $z \in \mathbb{R}^{2N^2}$, then*

$$\begin{bmatrix} y_i \\ y_{N^2+i} \end{bmatrix} = \operatorname{prox}_{\beta|\cdot|_\alpha}\left(\left\|\begin{bmatrix} z_i \\ z_{N^2+i} \end{bmatrix}\right\|\right)\frac{\begin{bmatrix} z_i \\ z_{N^2+i} \end{bmatrix}}{\left\|\begin{bmatrix} z_i \\ z_{N^2+i} \end{bmatrix}\right\|}, \quad i = 1, 2, \ldots, N^2.$$

Proof. It is simply the result of Lemma 6.2.1 and Proposition 6.2.1. □

In our experiments, we choose the "Cameraman" image as the original image x. Cameraman is a 256×256 grayscale image commonly used in image processing. The noisy images are modeled as

$$z = x + \epsilon$$

with ϵ being Gaussian noise. The noise levels of ϵ with the standard deviations 15, 20, and 25 are added to the original image to evaluate the proposed model and the corresponding algorithms.

The quality of denoised images $\widetilde{x}$ obtained from various denoising algorithms is evaluated by the peak-signal-to-noise ratio (PSNR)

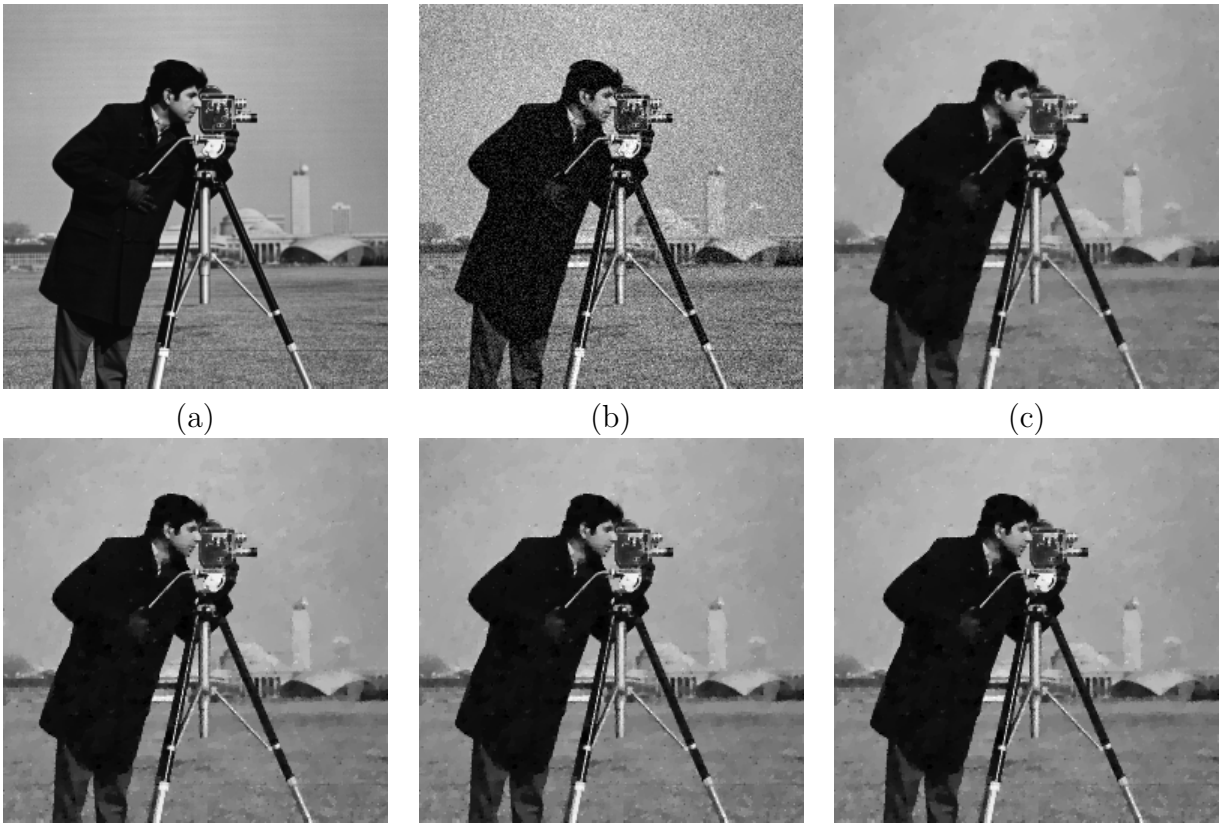
$$\text{PSNR} := 20 \log_{10} \left(\frac{255}{\|x - \widetilde{x}\|_2} \right).$$

Generally, the PSNR value of an image is the ratio between the maximum possible power of the signal (255 for grayscale images) and the power of the noise (given here by the mean squared error). A higher PSNR value indicates greater fidelity between the recovered image and the original.

The average PSNR values for the denoised images for various values of λ and α over 30 trials are listed in Table 3. For noise with standard deviation 20, the regularization parameter λ being 16, and the parameter α being $1.6\lambda\|D\|^2$, the denoised images are shown in Figure 18. For the same parameters λ and α, we report the PSNR values and CPU times (in seconds) of each noise realization in Figure 19

Table 3: Numerical results of ROF-PD, DCA, ADMM, and PD method for the image of "Cameraman".

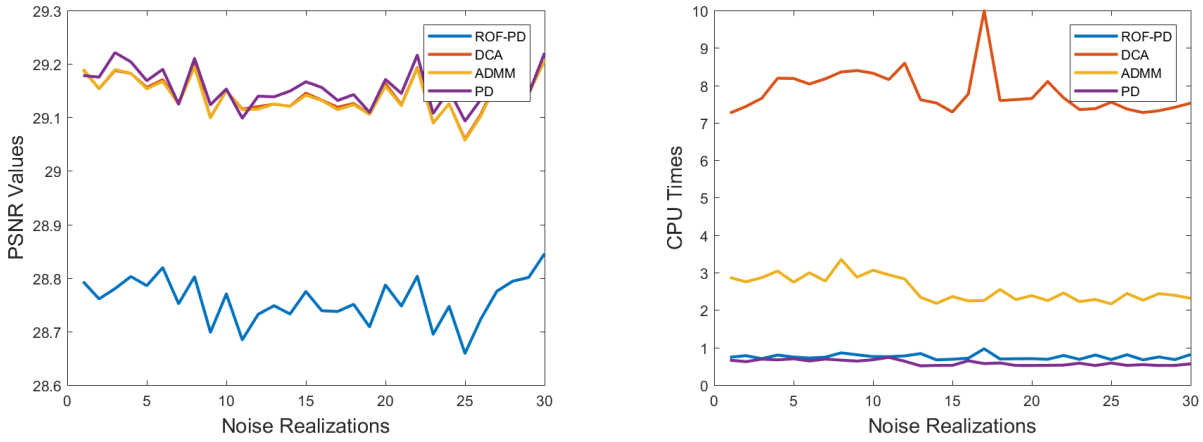
λ	ROF-PD	DCA	ADMM	PD	DCA	ADMM	PD
		$\alpha = 1.6\lambda\|D\|^2$			$\alpha = 0.9\lambda\|D\|^2$		
White Gaussian noise with standard deviation 15							
8	30.26	29.77	29.69	29.77	28.57	28.34	28.58
9	30.39	30.26	30.21	30.27	29.40	29.19	29.41
10	30.39	30.54	30.29	30.55	30.00	30.52	30.02
11	30.29	30.64	30.64	30.66	30.37	30.29	30.38
12	30.12	30.62	30.62	30.63	30.52	30.49	30.53
13	29.91	30.50	30.51	30.51	30.53	30.53	30.54
White Gaussian noise with standard deviation 20							
13	28.88	28.83	28.80	28.85	28.12	27.92	28.14
14	28.90	29.03	29.01	29.05	28.56	28.41	28.57
15	28.85	29.13	29.12	29.15	28.84	28.74	28.86
16	28.76	29.14	29.14	29.16	29.00	28.95	29.02
17	28.64	29.09	29.10	29.11	29.07	28.95	29.02
18	28.49	29.00	29.01	29.01	29.06	29.06	29.08
White Gaussian noise with standard deviation 25							
17	27.77	27.74	27.70	27.76	27.11	26.89	27.12
18	27.79	27.90	27.88	27.92	27.45	27.29	27.47
19	27.78	27.99	27.98	28.01	27.70	27.59	27.72
20	27.73	28.03	28.02	28.05	27.87	27.79	27.89
21	27.65	28.01	28.01	28.03	27.96	27.92	27.98
22	27.56	27.96	27.97	27.98	27.99	27.97	28.01



(a) (b) (c)

(d) (e) (f)

Figure 18: (a) Cameraman; (b) Cameraman with Gaussian noise of standard deviation 20; the denoisied images by using (c) ROF-PD; (d) DCA; (e) ADMM; and (f) PD. The regularization parameter λ is 16 and the parameter α is $1.6\lambda\|D\|^2$.



(a) (b)

Figure 19: (a) The PSNR value of each Gaussian noise realization and (b) the cpu time consumed with standard deviation 20. The regularization parameter λ is 16 and the parameter α is $1.6\lambda\|D\|^2$.

Chapter 7

Conclusions and Future Directions

Motivated by the need for nonconvex penalties in sparse optimization, we provide a simple and intuitive definition of sparsity promoting functions and introduce a method of constructing semiconvex sparsity promoting functions. Theoretical properties of these functions are developed throughout Chapter 3. In particular, we show that our construction preserves properties of the parent function, and both functions are characterized by the thresholding effects of their proximity operators. A basic study of geometric properties related to optimization is included. A deeper examination of certain classes of functions is given in Chapter 4, and several examples of practical interest are given in detail. Chapter 5 highlights the model flexibility and algorithmic performance of our construction. Finally, we demonstrate the applicability of our work by applying these results to the total variation denoising problems in signal and image processing in Chapter 6.

Future Directions

Chapter 3. Because of the relationship between a function f and its envelope $\text{env}_\alpha f$, we believe that f_α may have even more structure than described here. For example, we show that our constructed functions are semiconvex, but they may also be quasiconvex or pseudoconvex. This would open up a variety of results in the theory of quasiconvex functions and quasimonotone operators. There is also more work to be done regarding the Łojasiewicz property. There are many characterizations of Łojasiewicz (or Kurdkyka-Łojasiewicz) functions which provide avenues into other applications. Even more generally, we are also interested in studying whether this construction might be of use in contexts beyond sparsity promotion.

Chapter 5. We believe that the structure of our functions can further improve convergence analysis for the algorithms given here. Recent work shows that the difference of convex algorithm can be boosted if part or all of the objective function is differentiable. While we have shown that the dual objective for our problem is differentiable under certain circumstances, the functions involved may be very difficult to compute. Instead, we hope to modify the boosting algorithm to suit the primal problem. We also suspect that we can improve the parameters in Algorithm 3 by leveraging properties of our functions.

Chapter 6. We would like to extend the results from this chapter to the problem of image restoration, in which the TVD model is adapted to include a blurring kernel. We will also be exploring how our sparsity promoting functions can be used for functional compression, which has implications for computational efficiency and security. Functional compression considers the problem of compressing source data in such a way that a function of the sources can be computed at the receiver using only the compressed data, where here the compression will be achieved using our functions. Questions of interest include how

to incorporate knowledge of the end function into our model and whether we can provide theoretical compression guarantees.

Bibliography

[1] Pierre-Antoine Absil, Robert Mahoney, and Ben Andrews. Convergence of the Iterates of Descent Methods for Analytic Cost Functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[2] Francisco J. Aragón Artacho, Ronan M.T. Fleming, and Phan T. Vuong. Accelerating the DC algorithm for smooth functions. *Mathematical Programming*, 169(1):95–118, 2018.

[3] Francisco J Aragón Artacho and Phan T Vuong. The Boosted DC Algorithm for nonsmooth functions. 2018.

[4] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[5] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017.

[6] Dimitri P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[7] Jacek Bochnak, Michel Coste, and Marie-Francoise Roy. *Real Algebraic Geometry*. Springer-Verlag Berlin Heidelberg, 1 edition, 1998.

[8] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz Inequality for Nonsmooth Subanalytic Functions with Applications to Subgradient Dynamical Systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[9] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Lojasiewicz Inequalities: Subgradient Flows, Talweg, Convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] Andrew G. Bruce and Hong-Ye Gao. WaveShrink: shrinkage functions and thresholds. *Proceedings of the SPIE*, 2569:270–281, 1995.

[12] Emmanuel Candes and Terence Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52:5406–5425, 2006.

[13] Piermarco Cannarsa and Carlo Sinestrari. *Semiconcave Functions, Hamilton-Jacobi Equations, and Optimal Control*. Birkhauser, Boston, 2004.

[14] Antonin Chambolle and Thomas Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[15] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[16] Patrick L. Combettes and Valérie R. Wajs. Signal Recovery by Proximal Forward-Backward Splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[17] Laurent Condat. A primal-dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[18] Michel Coste. An Introduction to o-minimal Geometry. 2000.

[19] David L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[20] Jonathan Eckstein and Dimitri P. Bertsekas. On the DouglasRachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[21] Jonathan Eckstein and Dimitri P. Bertsekas. On the Splitting Methods and the Proximal Point Algorithm for Maximal Monotone Operators. *Mathematical Programming*, 12:83–92, 2004.

[22] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[23] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.

[24] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[25] Nicolas Hadjisavvas, Sándor Komlósi, and Siegfried S. Schaible, editors. *Handbook of Generalized Convexity and Generalized Monotonicity*. Springer-Verlag New York, 2005.

[26] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms II*. Springer-Verlag Berlin Heidelberg, 1993.

[27] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag, Berlin, 2001.

[28] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2nd edition, 2009.

[29] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48:769–783, 1998.

[30] Hoai An Le Thi, Van Ngai Huynh, and Tao Pham Dinh. Convergence Analysis of Difference-of-Convex Algorithm with Subanalytic Data. *Journal of Optimization Theory and Applications*, 179(1):103–126, 2018.

[31] Hoai An Le Thi, Tao Pham Dinh, and Muu Le Dung. Numerical solution for optimization over the efficient set by d.c. optimization algorithms. *Operations Research Letters*, 19(3):117–128, 1996.

[32] Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization arXiv : 1407 . 0753v6 [math . OC] 4 Nov 2015. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

[33] Guoyin Li and Ting Kei Pong. Calculus of the Exponent of KurdykaŁojasiewicz Inequality and Its Applications to Linear Convergence of First-Order Methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018.

[34] Stanislaw Łojasiewicz. Sur les trajectoires de gradient d'une fonction analytique, 1984.

[35] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2nd edition, 1999.

[36] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[37] Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[38] Tao Pham Dinh and Hoai An Le Thi. Convex Analysis Approach to D.C. Programming: Theory, Algorithms, and Applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

[39] Chayne Planiden. *Theory and Algorithmic Applications of the Proximal Mapping and Moreau Envelope*. PhD thesis, The University of British Columbia, Okanagan, 2018.

[40] Chayne Planiden and Xianfu Wang. Proximal mappings and Moreau envelopes of convex piecewise cubic functions and gauge functions. 2017.

[41] Chayne Planiden and Xianfu Wang. Epi-convergence: The Moreau Envelope and Generalized Linear-Quadratic Functions. *Journal of Optimization Theory and Applications*, 177(1):21–63, 2018.

[42] René A. Poliquin. Integration of subdifferentials of nonconvex function. *Nonlinear Analysis: Theory, Methods & Applications*, 17(4):385–398, 1991.

[43] Rene A. Poliquin and R. Tyrrell Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.

[44] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[45] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[46] Ivan Selesnick. Total Variation Denoising Via the Moreau Envelope. *IEEE Signal Processing Letters*, 24(2):216–220, 2017.

[47] Lixin Shen, Bruce W Suter, and Erin E Tripp. Structured Sparsity Promoting Functions. *Journal of Optimization Theory and Applications*, pages 1–29, 2019.

[48] Emmanuel Soubies, Laure Blanc-Féraud, and Gilles Aubert. A Continuous Exact l_0 penalty (CEL0) for least squares regularized problem . *SIAM Journal on Imaging Science*, 8(3):1607–1639, 2015.

[49] Emmanuel Soubies, Laure Blanc-Féraud, and Gilles Aubert. A Unified View of Exact Continuous Penalties for l2-l0 Minimization. *SIAM Journal on Optimization*, 27(3), 2017.

[50] Bruce W. Suter. *Multirate and wavelet signal processing.* Academic Press, 1997.

[51] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[52] Yu Wang, Wotao Yin, and Jinshan Zeng. Global Convergence of ADMM in Nonconvex Nonsmooth Optimization. *Journal of Scientific Computing*, pages 1–35, 2018.

[53] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 28:894–942, 2010.

[54] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Vita

Erin E. Tripp

Education

Master of Science in Mathematics, Syracuse University, May 2017

Bachelor of Science in Mathematics, University of California, Santa Barbara, June 2013

Employment

Graduate Teaching Assistant, Syracuse University, 2014-Present

Research Intern, Air Force Research Laboratory, Rome, New York, 2015-Present

Publications

Lixin Shen, Bruce W. Suter, and Erin E. Tripp. *Structured Sparsity Promoting Functions.* Accepted, Journal of Optimization Theory and Applications.

Service and Memberships

Association for Women in Mathematics, Student Chapter President, 2018-2019

Graduate Student Organization Senator, Syracuse University, 2018-2019

Math Graduate Student Organization Colloquium Organizer, Syracuse University, 2016-2018

Awards and Honors

The Kibbey Award, Department of Mathematics, Syracuse University, 2019

American Mathematical Society Graduate Student Travel Grant, 2019

Speaker Honorarium, University of Wisconsin, Superior, 2018

Outstanding Teaching Assistant Award, The Graduate School, Syracuse University, 2017

Women in Science and Engineering Future Professionals Program, Syracuse University, 2015-2017