The Trimmed Lasso: Sparsity and Robustness

Dimitris Bertsimas, Martin S. Copenhaver, and Rahul Mazumder[*]

August 15, 2017

Abstract

Nonconvex penalty methods for sparse modeling in linear regression have been a topic of fervent interest in recent years. Herein, we study a family of nonconvex penalty functions that we call the trimmed Lasso and that offers exact control over the desired level of sparsity of estimators. We analyze its structural properties and in doing so show the following:

1. Drawing parallels between robust statistics and robust optimization, we show that the trimmed-Lasso-regularized least squares problem can be viewed as a generalized form of total least squares under a specific model of uncertainty. In contrast, this same model of uncertainty, viewed instead through a robust optimization lens, leads to the convex SLOPE (or OWL) penalty.
2. Further, in relating the trimmed Lasso to commonly used sparsity-inducing penalty functions, we provide a succinct characterization of the connection between trimmed-Lasso-like approaches and penalty functions that are coordinate-wise separable, showing that the trimmed penalties subsume existing coordinate-wise separable penalties, with strict containment in general.
3. Finally, we describe a variety of exact and heuristic algorithms, both existing and new, for trimmed Lasso regularized estimation problems. We include a comparison between the different approaches and an accompanying implementation of the algorithms.

1 Introduction

Sparse modeling in linear regression has been a topic of fervent interest in recent years [23, 42]. This interest has taken several forms, from substantial developments in the theory of the Lasso to advances in algorithms for convex optimization. Throughout there has been a strong emphasis on the increasingly high-dimensional nature of linear regression problems; in such problems, where the number of variables p can vastly exceed the number of observations n, sparse modeling techniques are critical for performing inference.

Context

One of the fundamental approaches to sparse modeling in the usual linear regression model of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$, is the best subset selection [57] problem:

$$\min_{\|\boldsymbol{\beta}\|_0 \le k} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{1}$$

[*] Authors' affiliation: Sloan School of Management and Operations Research Center, MIT.
Emails: {dbertsim,mcopen,rahulmaz}@mit.edu.

which seeks to find the best choice of k from among p features that best explain the response in terms of the least squares loss function. The problem (1) has received extensive attention from a variety of statistical and optimization perspectives—see for example [14] and references therein. One can also consider the Lagrangian, or penalized, form of (1), namely,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_0, \tag{2}$$

for a regularization parameter $\mu > 0$. One of the advantages of (1) over (2) is that it offers direct control over estimators' sparsity via the discrete parameter k, as opposed to the Lagrangian form (2) for which the correspondence between the continuous parameter μ and the resulting sparsity of estimators obtained is not entirely clear. For further discussion, see [65].

Another class of problems that have received considerable attention in the statistics and machine learning literature is the following:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + R(\boldsymbol{\beta}), \tag{3}$$

where $R(\boldsymbol{\beta})$ is a choice of regularizer which encourages sparsity in $\boldsymbol{\beta}$. For example, the popularly used Lasso [70] takes the form of problem (3) with $R(\boldsymbol{\beta}) = \mu\|\boldsymbol{\beta}\|_1$, where $\|\cdot\|_1$ is the ℓ_1 norm; in doing so, the Lasso simultaneously selects variables and also performs shrinkage. The Lasso has seen widespread success across a variety of applications.

In contrast to the convex approach of the Lasso, there also has been been growing interest in considering richer classes of regularizers R which include nonconvex functions. Examples of such penalties include the ℓ_q-penalty (for $q \in [0, 1]$), minimax concave penalty (MCP) [74], and the smoothly clipped absolute deviation (SCAD) [33], among others. Many of the nonconvex penalty functions considered are *coordinate-wise separable*; in other words, R can be decomposed as

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{p} \rho(|\beta_i|),$$

where $\rho(\cdot)$ is a real-valued function [75]. There has been a variety of evidence suggesting the promise of such nonconvex approaches in overcoming certain shortcomings of Lasso-like approaches.

One of the central ideas of nonconvex penalty methods used in sparse modeling is that of creating a continuum of estimation problems which bridge the gap between convex methods for sparse estimation (such as Lasso) and subset selection in the form (1). However, as noted above, such a connection does not necessarily offer direct control over the desired level of sparsity of estimators.

The trimmed Lasso

In contrast with coordinate-wise separable penalties as considered above, we consider a family of penalties that are not separable across coordinates. One such penalty which forms a principal object of our study herein is

$$T_k(\boldsymbol{\beta}) := \min_{\|\boldsymbol{\phi}\|_0 \le k} \|\boldsymbol{\phi} - \boldsymbol{\beta}\|_1.$$

The penalty T_k is a measure of the distance from the set of k-sparse estimators as measured via the ℓ_1 norm. In other words, when used in problem (3), the penalty $R = T_k$ controls the amount of shrinkage towards sparse models.

The penalty T_k can equivalently be written as

$$T_k(\boldsymbol{\beta}) = \sum_{i=k+1}^{p} |\beta_{(i)}|,$$

where $|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(p)}|$ are the sorted entries of $\boldsymbol{\beta}$. In words, $T_k(\boldsymbol{\beta})$ is the sum of the absolute values of the $p-k$ smallest magnitude entries of $\boldsymbol{\beta}$. The penalty was first introduced in [39, 43, 69, 72]. We refer to this family of penalty functions (over choices of k) as the *trimmed Lasso.*[1] The case of $k=0$ recovers the usual Lasso, as one would suspect. The distinction, of course, is that for general k, T_k no longer shrinks, or biases towards zero, the k largest entries of $\boldsymbol{\beta}$.

Let us consider the least squares loss regularized via the trimmed lasso penalty—this leads to the following optimization criterion:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k(\boldsymbol{\beta}), \tag{4}$$

where $\lambda > 0$ is the regularization parameter. The penalty term shrinks the smallest $p-k$ entries of $\boldsymbol{\beta}$ and does not impose any penalty on the largest k entries of $\boldsymbol{\beta}$. If λ becomes larger, the smallest $p-k$ entries of $\boldsymbol{\beta}$ are shrunk further; after a certain threshold—as soon as $\lambda \geq \lambda_0$ for some finite λ_0—the smallest $p-k$ entries are set to zero. The existence of a finite λ_0 (as stated above) is an attractive feature of the trimmed Lasso and is known as its *exactness* property, namely, for λ sufficiently large, the problem (4) exactly solves constrained best subset selection as in problem (1) (*c.f.* [39]). Note here the contrast with the separable penalty functions which correspond instead with problem (2); as such, the trimmed Lasso is distinctive in that it offers precise control over the desired level of sparsity vis-à-vis the discrete parameter k. Further, it is also notable that many algorithms developed for separable-penalty estimation problems can be directly adapted for the trimmed Lasso.

Our objective in studying the trimmed Lasso is distinctive from previous approaches. In particular, while previous work on the penalty T_k has focused primarily on its use as a tool for reformulating sparse optimization problems [43, 69] and on how such reformulations can be solved computationally [39, 72], we instead aim to explore the trimmed Lasso's structural properties and its relation to existing sparse modeling techniques.

In particular, a natural question we seek to explore is, what is the connection of the trimmed Lasso penalty with existing separable penalties commonly used in sparse statistical learning? For example, the trimmed Lasso bears a close resemblance to the clipped (or capped) Lasso penalty [76], namely,

$$\sum_{i=1}^{p} \mu \min\{\gamma|\beta_i|, 1\},$$

where $\mu, \gamma > 0$ are parameters (when γ is large, the clipped Lasso approximates $\mu\|\boldsymbol{\beta}\|_0$).

Robustness: robust statistics and robust optimization

A significant thread woven throughout the consideration of penalty methods for sparse modeling is the notion of robustness—in short, the ability of a method to perform in the face of noise. Not surprisingly, the notion of robustness has myriad distinct meanings depending on the context. Indeed, as Huber, a pioneer in the area of robust statistics, aptly noted:

[1] The choice of name is our own and is motivated by the least trimmed squares regression estimator, described below

> "The word 'robust' is loaded with many—sometimes inconsistent—connotations." [45, p. 2]

For this reason, we consider robustness from several perspectives—both the robust statistics [45] and robust optimization [9] viewpoints.

A common premise of the various approaches is as follows: that a robust model should perform well even under small deviations from its underlying assumptions; and that to achieve such behavior, some efficiency under the assumed model should be sacrificed. Not surprisingly in light of Huber's prescient observation, the exact manifestation of this idea can take many different forms, even if the initial premise is ostensibly the same.

Robust statistics and the "min-min" approach

One such approach is in the field of robust statistics [45, 58, 61]. In this context, the primary assumptions are often probabilistic, i.e. distributional, in nature, and the deviations to be "protected against" include possibly gross, or arbitrarily bad, errors. Put simply, robust statistics is primary focused on analyzing and mitigating the influence of outliers on estimation methods.

There have been a variety of proposals of different estimators to achieve this. One that is particularly relevant for our purposes is that of *least trimmed squares* ("LTS") [61]. For fixed $j \in \{1, \ldots, n\}$, the LTS problem is defined as

$$\min_{\boldsymbol{\beta}} \sum_{i=j+1}^{n} |r_{(i)}(\boldsymbol{\beta})|^2, \tag{5}$$

where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ are the residuals and $r_{(i)}(\boldsymbol{\beta})$ are the sorted residuals given $\boldsymbol{\beta}$ with $|r_{(1)}(\boldsymbol{\beta})| \geq |r_{(2)}(\boldsymbol{\beta})| \geq \cdots \geq |r_{(n)}(\boldsymbol{\beta})|$. In words, the LTS estimator performs ordinary least squares on the $n - j$ smallest residuals (discarding the j largest or worst residuals).

Furthermore, it is particularly instructive to express (5) in the equivalent form (*c.f.* [16])

$$\min_{\boldsymbol{\beta}} \min_{\substack{I \subseteq \{1,\ldots,n\}: \\ |I| = n-j}} \sum_{i \in I} |r_i(\boldsymbol{\beta})|^2. \tag{6}$$

In light of this representation, we refer to LTS as a form of "min-min" robustness. One could also interpret this min-min robustness as *optimistic* in the sense the estimation problems (6) and, *a fortiori*, (5) allow the modeler to also choose observations to discard.

Other min-min models of robustness

Another approach to robustness which also takes a min-min form like LTS is the classical technique known as *total least squares* [38, 54]. For our purposes, we consider total least squares in the form

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\Delta}} \frac{1}{2} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2 + \eta \|\boldsymbol{\Delta}\|_2^2, \tag{7}$$

where $\|\boldsymbol{\Delta}\|_2$ is the usual Frobenius norm of the matrix $\boldsymbol{\Delta}$ and $\eta > 0$ is a scalar parameter. In this framework, one again has an optimistic view on error: find the best possible "correction" of the data matrix $\mathbf{X}$ as $\mathbf{X} + \boldsymbol{\Delta}^*$ and perform least squares using this corrected data (with η controlling the flexibility in choice of $\boldsymbol{\Delta}$).

In contrast with the penalized form of (7), one could also consider the problem in a constrained form such as

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\Delta} \in \mathcal{V}} \frac{1}{2} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2, \tag{8}$$

where $\mathcal{V} \subseteq \mathbb{R}^{n \times p}$ is defined as $\mathcal{V} = \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_2 \leq \eta'\}$ for some $\eta' > 0$. This problem again has the min-min form, although now with perturbations $\mathbf{\Delta}$ as restricted to the set $\mathcal{V}$.

Robust optimization and the "min-max" approach

We now turn our attention to a different approach to the notion of robustness known as robust optimization [9, 12]. In contrast with robust statistics, robust optimization typically replaces distributional assumptions with a new primitive, namely, the deterministic notion of an *uncertainty set*. Further, in robust optimization one considers a worst-case or pessimistic perspective and the focus is on perturbations from the nominal model (as opposed to possible gross corruptions as in robust statistics).

To be precise, one possible robust optimization model for linear regression takes form [9, 15, 73]

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta} \in \mathcal{U}} \frac{1}{2}\|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}\|_2^2, \tag{9}$$

where $\mathcal{U} \subseteq \mathbb{R}^{n \times p}$ is a (deterministic) uncertainty set that captures the possible deviations of the model (from the nominal data $\mathbf{X}$). Note the immediate contrast with the robust models considered earlier (LTS and total least squares in (5) and (7), respectively) that take the min-min form; instead, robust optimization focuses on "min-max" robustness. For a related discussion contrasting the min-min approach with min-max, see [8, 49] and references therein.

One of the attractive features of the min-max formulation is that it gives a re-interpretation of several statistical regularization methods. For example, the usual Lasso (problem (3) with $R = \mu\ell_1$) can be expressed in the form (9) for a specific choice of uncertainty set:

Proposition 1.1 (e.g. [9, 73])**.** *Problem* (9) *with uncertainty set* $\mathcal{U} = \{\mathbf{\Delta} : \|\mathbf{\Delta}_i\|_2 \leq \mu \ \forall i\}$ *is equivalent to the Lasso, i.e., problem* (3) *with* $R(\boldsymbol{\beta}) = \mu\|\boldsymbol{\beta}\|_1$*, where* $\mathbf{\Delta}_i$ *denotes the ith column of* $\mathbf{\Delta}$*.*

For further discussion of the robust optimization approach as applied to statistical problems, see [15] and references therein.

Other min-max models of robustness

We close our discussion of robustness by considering another example of min-max robustness that is of particular relevance to the trimmed Lasso. In particular, we consider problem (3) with the SLOPE (or OWL) penalty [18, 35], namely,

$$R_{\mathrm{SLOPE}(\mathbf{w})}(\boldsymbol{\beta}) = \sum_{i=1}^{p} w_i |\beta_{(i)}|,$$

where $\mathbf{w}$ is a (fixed) vector of weights with $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and $w_1 > 0$. In its simplest form, the SLOPE penalty has weight vector $\tilde{\mathbf{w}}$, where $\tilde{w}_1 = \cdots = \tilde{w}_k = 1$, $\tilde{w}_{k+1} = \cdots = \tilde{w}_p = 0$, in which case we have the identity

$$R_{\mathrm{SLOPE}(\tilde{\mathbf{w}})}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 - T_k(\boldsymbol{\beta}).$$

There are some apparent similarities but also subtle differences between the SLOPE penalty and the trimmed Lasso. From a high level, while the trimmed Lasso focuses on the smallest magnitude entries of $\boldsymbol{\beta}$, the SLOPE penalty in its simplest form focuses on the *largest* magnitude entries of $\boldsymbol{\beta}$. As such, the trimmed Lasso is generally nonconvex, while the SLOPE penalty is always convex; consequently, the techniques for solving the related estimation problems will necessarily be different.

Finally, we note that the SLOPE penalty can be considered as a min-max model of robustness for a particular choice of uncertainty set:

Proposition 1.2. *Problem* (9) *with uncertainty set*

$$\mathcal{U} = \left\{ \boldsymbol{\Delta} : \begin{array}{c} \boldsymbol{\Delta} \textit{ has at most } k \textit{ nonzero} \\ \textit{columns and } \|\boldsymbol{\Delta}_i\|_2 \leq \mu \ \forall i \end{array} \right\}$$

is equivalent to problem (3) *with* $R(\boldsymbol{\beta}) = \mu R_{\mathrm{SLOPE}(\tilde{\mathbf{w}})}(\boldsymbol{\beta})$, *where* $\tilde{w}_1 = \cdots = \tilde{w}_k = 1$ *and* $\tilde{w}_{k+1} = \cdots = \tilde{w}_p = 0$.

We return to this particular choice of uncertainty set later. (For completeness, we include a more general min-max representation of SLOPE in Appendix A.)

Computation and Algorithms

Broadly speaking, there are numerous distinct approaches to algorithms for solving problems of the form (1)–(3) for various choices of R. We do not attempt to provide a comprehensive list of such approaches for general R, but we will discuss existing approaches for the trimmed Lasso and closely related problems. Approaches typically take one of two forms: heuristic or exact.

Heuristic techniques

Heuristic approaches to solving problems (1)–(3) often use techniques from convex optimization [21], such as proximal gradient descent or coordinate descent (see [33,55]). Typically these techniques are coupled with an analysis of local or global behavior of the algorithm. For example, global behavior is often considered under additional restrictive assumptions on the underlying data; unfortunately, verifying such assumptions can be as difficult as solving the original nonconvex problem. (For example, consider the analogy with compressed sensing [25, 30, 32] and the hardness of verifying whether underlying assumptions hold [5, 71]).

There is also extensive work studying the local behavior (e.g. stationarity) of heuristic approaches to these problems. For the specific problems (1) and (2), the behavior of augmented Lagrangian methods [4, 68] and complementarity constraint techniques [22, 24, 29, 34] have been considered. For other local approaches, see [52].

Exact techniques

One of the primary drawbacks of heuristic techniques is that it can often be difficult to verify the degree of suboptimality of the estimators obtained. For this reason, there has been an increasing interest in studying the behavior of exact algorithms for providing certifiably optimal solutions to problems of the form (1)–(3) [14, 16, 51, 56]. Often these approaches make use of techniques from *mixed integer optimization* ("MIO") [19] which are implemented in a variety of software, e.g. Gurobi [40]. The tradeoff with such approaches is that they typically carry a heavier computational burden than convex approaches. For a discussion of the application of MIO in statistics, see [14, 16, 51, 56].

What this paper is about

In this paper, we focus on a detailed analysis of the trimmed Lasso, especially with regard to its properties and its relation to existing methods. In particular, we explore the trimmed Lasso from two perspectives: that of sparsity as well as that of robustness. We summarize our contributions as follows:

1. We study the robustness of the trimmed Lasso penalty. In particular, we provide several min-min robustness representations of it. We first show that the same choice of uncertainty set that leads to the SLOPE penalty in the min-max robust model (9) gives rise to the trimmed Lasso in the corresponding min-min robust problem (8) (with an additional regularization term). This gives an interpretation of the SLOPE and trimmed Lasso as a complementary pair of penalties, one under a pessimistic (min-max) model and the other under an optimistic (min-min) model.

   Moreover, we show another min-min robustness interpretation of the trimmed Lasso by comparison with the ordinary Lasso. In doing so, we further highlight the nature of the trimmed Lasso and its relation to the LTS problem (5).

2. We provide a detailed analysis on the connection between estimation approaches using the trimmed Lasso and separable penalty functions. In doing so, we show directly how penalties such as the trimmed Lasso can be viewed as a generalization of such existing approaches in certain cases. In particular, a trimmed-Lasso-like approach always subsumes its separable analogue, and the containment is strict in general. We also focus on the specific case of the clipped (or capped) Lasso [76]; for this we precisely characterize the relationship and provide a necessary and sufficient condition for the two approaches to be equivalent. In doing so, we highlight some of the limitations of an approach using a separable penalty function.

3. Finally, we describe a variety of algorithms, both existing and new, for trimmed Lasso estimation problems. We contrast two heuristic approaches for finding locally optimal solutions with exact techniques from mixed integer optimization that can be used to produce certificates of optimality for solutions found via the convex approaches. We also show that the convex envelope [60] of the trimmed Lasso takes the form

   $$(\|\boldsymbol{\beta}\|_1 - k)_+ ,$$

   where $(a)_+ := \max\{0, a\}$, a "soft-thresholded" variant of the ordinary Lasso. Throughout this section, we emphasize how techniques from convex optimization can be used to find high-quality solutions to the trimmed Lasso estimation problem. An implementation of the various algorithms presented herein can be found at

   https://github.com/copenhaver/trimmedlasso.

Paper structure

The structure of the paper is as follows. In Section 2, we study several properties of the trimmed Lasso, provide a few distinct interpretations, and highlight possible generalizations. In Section 3, we explore the trimmed Lasso in the context of robustness. Then, in Section 4, we study the relationship between the trimmed Lasso and other nonconvex penalties. In Section 5, we study the algorithmic implications of the trimmed Lasso. Finally, in Section 6 we share our concluding thoughts and highlight future directions.

2 Structural properties and interpretations

In this section, we provide further background on the trimmed Lasso: its motivations, interpretations, and generalizations. Our remarks in this section are broadly grouped as follows: in Section 2.1 we summarize the trimmed Lasso's basic properties as detailed in [39, 43, 69, 72]; we then turn our attention to an interpretation of the trimmed Lasso as a relaxation of complementarity constraints problems from optimization (Section 2.2) and as a variable decomposition method (Section 2.3); finally, in Sections 2.4 and 2.5 we highlight the key structural features of the trimmed Lasso by identifying possible generalizations of its definition and its application. These results augment the existing literature by giving a deeper understanding of the trimmed Lasso and provide a basis for further results in Sections 3 and 4.

2.1 Basic observations

We begin with a summary of some of the basic properties of the trimmed Lasso as studied in [39, 43, 69]. First of all, let us also include another representation of T_k:

Lemma 2.1. *For any $\boldsymbol{\beta}$,*

$$T_k(\boldsymbol{\beta}) = \min_{\substack{I\subseteq\{1,\ldots,p\}:\\ |I|=p-k}} \sum_{i\in I} |\beta_i| = \begin{array}{ll} \min\limits_{\mathbf{z}} & \langle \mathbf{z}, |\boldsymbol{\beta}| \rangle \\ \text{s. t.} & \sum\limits_i z_i = p-k \\ & \mathbf{z} \in \{0,1\}^p, \end{array}$$

where $|\boldsymbol{\beta}|$ denotes the vector whose entries are the absolute values of the entries of $\boldsymbol{\beta}$.

In other words, the trimmed Lasso can be represented using auxiliary binary variables.

Now let us consider the problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k(\boldsymbol{\beta}), \tag{$\mathrm{TL}_{\lambda,k}$}$$

where $\lambda > 0$ and $k \in \{0, 1, \ldots, p\}$ are parameters. Based on the definition of T_k, we have the following:

Lemma 2.2. *The problem* $(\mathrm{TL}_{\lambda,k})$ *can be rewritten exactly in several equivalent forms:*

$$\begin{aligned} (\mathrm{TL}_{\lambda,k}) &= \min_{\substack{\boldsymbol{\beta},\boldsymbol{\phi}:\\ \|\boldsymbol{\phi}\|_0\le k}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}-\boldsymbol{\phi}\|_1 \\ &= \min_{\substack{\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\epsilon}:\\ \boldsymbol{\beta}=\boldsymbol{\phi}+\boldsymbol{\epsilon}\\ \|\boldsymbol{\phi}\|_0\le k}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\epsilon}\|_1 \\ &= \min_{\substack{\boldsymbol{\phi},\boldsymbol{\epsilon}:\\ \|\boldsymbol{\phi}\|_0\le k}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}(\boldsymbol{\phi}+\boldsymbol{\epsilon})\|^2 + \lambda\|\boldsymbol{\epsilon}\|_1 \end{aligned}$$

Exact penalization

Based on the definition of T_k, it follows that $T_k(\boldsymbol{\beta}) = 0$ if and only if $\|\boldsymbol{\beta}\|_0 \le k$. Therefore, one can rewrite problem (1) as

$$\min_{T_k(\boldsymbol{\beta})=0} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2.$$

In Lagrangian form, this would suggest an approximation for (1) of the form

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k(\boldsymbol{\beta}),$$

where $\lambda > 0$. As noted in the introduction, this approximation is in fact exact (in the sense of [10, 11]), summarized in the following theorem; for completeness, we include in Appendix B a full proof that is distinct from that in [39].[2]

Theorem 2.3 (*c.f.* [39])**.** *For any fixed $k \in \{0, 1, 2, \dots, p\}$, $\eta > 0$, and problem data $\mathbf{y}$ and $\mathbf{X}$, there exists some $\overline{\lambda} = \overline{\lambda}(\mathbf{y}, \mathbf{X}) > 0$ so that for all $\lambda > \overline{\lambda}$, the problems*

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k\left(\boldsymbol{\beta}\right) + \eta\|\boldsymbol{\beta}\|_1$$

and

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}} & \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 \\ \text{s. t.} & \|\boldsymbol{\beta}\|_0 \leq k \end{array}$$

have the same optimal objective value and the same set of optimal solutions.

The direct implication is that trimmed Lasso leads to a continuum (over λ) of relaxations to the best subset selection problem starting from ordinary least squares estimation; further, best subset selection lies on this continuum for λ sufficiently large.

2.2 A complementary constraints viewpoint

We now turn our attention to a new perspective on the trimmed Lasso as considered via mathematical programming with complementarity constraints ("MPCCs") [24, 44, 47, 48, 50, 62], sometimes also referred to as mathematical programs with equilibrium constraints [27]. By studying this connection, we will show that a penalized form of a common relaxation scheme for MPCCs leads directly to the trimmed Lasso penalty. This gives a distinctly different optimization perspective on the trimmed Lasso penalty.

As detailed in [22, 24, 34], the problem (1) can be exactly rewritten as

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\mathbf{z}} & \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s. t.} & \sum_i z_i = p-k \\ & \mathbf{z} \in [0,1]^p \\ & z_i\beta_i = 0. \end{array} \tag{10}$$

by the inclusion of auxiliary variables $\mathbf{z} \in [0, 1]^p$. In essence, the auxiliary variables replace the combinatorial constraint $\|\boldsymbol{\beta}\|_0 \leq k$ with *complementarity* constraints of the form $z_i\beta_i = 0$. Of course, the problem as represented in (10) is still not directly amenable to convex optimization techniques.

As such, relaxation schemes can be applied to (10). One popular method from the MPCC literature is the Scholtes-type relaxation [44]; applied to (10) as in [24, 34], this takes the form

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\mathbf{z}} & \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s. t.} & \sum_i z_i = p-k \\ & \mathbf{z} \in [0,1]^p \\ & |z_i\beta_i| \leq t, \end{array} \tag{11}$$

[2] The presence of the additional regularizer $\eta\|\boldsymbol{\beta}\|_1$ can be interpreted in many ways. For our purposes, it serves to make the problems well-posed.

where $t > 0$ is some fixed numerical parameter which controls the strength of the relaxation, with $t = 0$ exactly recovering (10). In the traditional MPCC context, it is standard to study local optimality and stationarity behavior of solutions to (11) as they relate to the original problem (1), *c.f.* [34].

Instead, let us consider a different approach. In particular, consider a penalized, or Lagrangian, form of the Scholtes relaxation (11), namely,

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\mathbf{z}} & \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum\limits_i(|z_i\beta_i| - t) \\ \text{s. t.} & \sum_i z_i = p-k \\ & \mathbf{z}\in[0,1]^p \end{array} \tag{12}$$

for some fixed $\lambda \geq 0$.[3] Observe that we can minimize (12) with respect to $\mathbf{z}$ to obtain the equivalent problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k(\boldsymbol{\beta}) - p\lambda t,$$

which is precisely problem ($\mathrm{TL}_{\lambda,\ell}$) (up to the fixed additive constant). In other words, the trimmed Lasso can also be viewed as arising directly from a penalized form of the MPCC relaxation, with auxiliary variables eliminated. This gives another view on Lemma 2.1 which gave a representation of T_k using auxiliary binary variables.

2.3 Variable decomposition

To better understand the relation of the trimmed Lasso to existing methods, it is also useful to consider alternative representations. Here we focus on representations which connect it to variable decomposition methods. Our discussion here is an extended form of related discussions in [39,43,72].

To begin, we return to the final representation of the trimmed Lasso problem as shown in Lemma 2.2, viz.,

$$(\mathrm{TL}_{\lambda,k}) = \min_{\substack{\boldsymbol{\phi},\boldsymbol{\epsilon}:\\ \|\boldsymbol{\phi}\|_0\leq k}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}(\boldsymbol{\phi}+\boldsymbol{\epsilon})\|^2 + \lambda\|\boldsymbol{\epsilon}\|_1. \tag{13}$$

We will refer to ($\mathrm{TL}_{\lambda,k}$) in the form (13) as the *split* or *decomposed* representation of the problem. This is because in this form it is clear that we can think about estimators $\boldsymbol{\beta}$ found via ($\mathrm{TL}_{\lambda,k}$) as being decomposed into two different estimators: a sparse component $\boldsymbol{\phi}$ and another component $\boldsymbol{\epsilon}$ with small ℓ_1 norm (as controlled via λ).

Several remarks are in order. First, the decomposition of $\boldsymbol{\beta}$ into $\boldsymbol{\beta} = \boldsymbol{\phi}+\boldsymbol{\epsilon}$ is truly a decomposition in that if $\boldsymbol{\beta}^*$ is an optimal solution to ($\mathrm{TL}_{\lambda,k}$) with $(\boldsymbol{\phi}^*,\boldsymbol{\epsilon}^*)$ a corresponding optimal solution to the split representation of the problem (13), then one must have that $\phi_i^*\epsilon_i^* = 0$ for all $i\in\{1,\ldots,p\}$. In other words, the supports of $\boldsymbol{\phi}$ and $\boldsymbol{\epsilon}$ do not overlap; therefore, $\boldsymbol{\beta}^* = \boldsymbol{\phi}^*+\boldsymbol{\epsilon}^*$ is a genuine decomposition.

Secondly, the variable decomposition (13) suggests that the problem of finding the k largest entries of $\boldsymbol{\beta}$ (i.e., finding $\boldsymbol{\phi}$) can be solved as a best subset selection problem with a (possibly different) convex loss function (without $\boldsymbol{\epsilon}$). To see this, observe that the problem of finding $\boldsymbol{\phi}$ in (13) can be written as the problem

$$\min_{\|\boldsymbol{\phi}\|_0\leq k} \widetilde{L}(\boldsymbol{\phi}),$$

[3] To be precise, this is a *weaker* relaxation than if we had separate dual variables λ_i for each constraint $|z_i\beta_i| \leq t$, at least in theory.

where

$$\widetilde{L}(\boldsymbol{\phi}) = \min_{\boldsymbol{\epsilon}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}(\boldsymbol{\phi} + \boldsymbol{\epsilon})\|_2^2 + \lambda\|\boldsymbol{\epsilon}\|_1.$$

Using theory on duality for the Lasso problem [59], one can argue that $\widetilde{L}$ is itself a convex loss function. Hence, the variable decomposition gives some insight into how the largest k loadings for the trimmed Lasso relates to solving a related sparse estimation problem.

A view towards matrix estimation

Finally, we contend that the variable decomposition of $\boldsymbol{\beta}$ as a sparse component $\boldsymbol{\phi}$ plus a "noise" component $\boldsymbol{\epsilon}$ with small norm is a natural and useful analogue of corresponding decompositions in the matrix estimation literature, such as in factor analysis [3,6,53] and robust Principal Component Analysis [26]. For the purposes of this paper, we will focus on the analogy with factor analysis.

Factor analysis is a classical multivariate statistical method for decomposing the covariance structure of random variables; see [13] for an overview of modern approaches to factor analysis. Given a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, one is interested in describing it as the sum of two distinct components: a low-rank component $\boldsymbol{\Theta}$ (corresponding to a low-dimensional covariance structure common across the variables) and a diagonal component $\boldsymbol{\Phi}$ (corresponding to individual variances unique to each variable)—in symbols, $\boldsymbol{\Sigma} = \boldsymbol{\Theta} + \boldsymbol{\Phi}$.

In reality, this *noiseless* decomposition is often too restrictive (see e.g. [41,63,67]), and therefore it is often better to focus on finding a decomposition $\boldsymbol{\Sigma} = \boldsymbol{\Theta}+\boldsymbol{\Phi}+\mathcal{N}$, where $\mathcal{N}$ is a noise component with small norm. As in [13], a corresponding estimation procedure can take the form

$$\begin{array}{ll} \min\limits_{\boldsymbol{\Theta},\boldsymbol{\Phi}} & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| \\ \text{s. t.} & \operatorname{rank}(\boldsymbol{\Theta}) \le k \\ & \boldsymbol{\Phi} = \operatorname{diag}(\Phi_{11}, \ldots, \Phi_{pp}) \succcurlyeq \mathbf{0} \\ & \boldsymbol{\Theta} \succcurlyeq \mathbf{0}, \end{array} \tag{14}$$

where the constraint $\mathbf{A} \succcurlyeq \mathbf{0}$ denotes that $\mathbf{A}$ is symmetric, positive semidefinite, and $\|\cdot\|$ is some norm. One of the attractive features of the estimation procedure (14) is that for common choices of $\|\cdot\|$, it is possible to completely eliminate the combinatorial rank constraint and the variable $\boldsymbol{\Theta}$ to yield a smooth (nonconvex) optimization problem with compact, convex constraints (see [13] for details).

This exact same argument can be used to motivate the appearance of the trimmed Lasso penalty. Indeed, instead of considering estimators $\boldsymbol{\beta}$ which are exactly k-sparse (i.e., $\|\boldsymbol{\beta}\|_0 \le k$), we instead consider estimators which are approximately k-sparse, i.e., $\boldsymbol{\beta} = \boldsymbol{\phi} + \boldsymbol{\epsilon}$, where $\|\boldsymbol{\phi}\|_0 \le k$ and $\boldsymbol{\epsilon}$ has small norm. Given fixed $\boldsymbol{\beta}$, such a procedure is precisely

$$\min_{\|\boldsymbol{\phi}\|_0 \le k} \|\boldsymbol{\beta} - \boldsymbol{\phi}\|.$$

Just as the rank constraint is eliminated from (14), the sparsity constraint can be eliminated from this to yield a continuous penalty which precisely captures the quality of the approximation $\boldsymbol{\beta} \approx \boldsymbol{\phi}$. The trimmed Lasso uses the choice $\|\cdot\| = \ell_1$, although other choices are possible; see Section 2.4.

This analogy with factor analysis is also useful in highlighting additional benefits of the trimmed Lasso. One of particular note is that it enables the direct application of existing convex optimization techniques to find high-quality solutions to $(\mathrm{TL}_{\lambda,k})$.

2.4 Generalizations

We close this section by considering some generalizations of the trimmed Lasso. These are particularly useful for connecting the trimmed Lasso to other penalties, as we will see later in Section 4.

As noted earlier, the trimmed Lasso measures the distance (in ℓ_1 norm) from the set of k-sparse vectors; therefore, it is natural to inquire what properties other measures of distance might carry. In light of this, we begin with a definition:

Definition 2.4. *Let $k \in \{0, 1, \dots, p\}$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ be any unbounded, continuous, and strictly increasing function with $g(0) = 0$. Define the corresponding kth projected penalty function, denoted π_k^g, as*

$$\pi_k^g(\boldsymbol{\beta}) = \min_{\|\boldsymbol{\phi}\|_0 \le k} \sum_i g(|\phi_i - \beta_i|).$$

It is not difficult to argue that π_k^g has as an equivalent definition

$$\pi_k^g(\boldsymbol{\beta}) = \sum_{i>k} g(|\beta_{(i)}|).$$

As an example, π_k^g is the trimmed Lasso penalty when g is the absolute value, viz. $g(x) = |x|$, and so it is a special case of the projected penalties. Alternatively, suppose $g(x) = x^2/2$. In this case, we get a trimmed version of the ridge regression penalty: $\sum_{i>k} |\beta_{(i)}|^2/2$.

This class of penalty functions has one notable feature, summarized in the following result:[4]

Proposition 2.5. *If $g : \mathbb{R}_+ \to \mathbb{R}_+$ is an unbounded, continuous, and strictly increasing function with $g(0) = 0$, then for any $\boldsymbol{\beta}$, $\pi_k^g(\boldsymbol{\beta}) = 0$ if and only if $\|\boldsymbol{\beta}\|_0 \le k$. Hence, the problem $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \pi_k^g(\boldsymbol{\beta})$ converges in objective value to $\min_{\|\boldsymbol{\beta}\|_0 \le k} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ as $\lambda \to \infty$.*

Therefore, any projected penalty π_k^g results in the best subset selection problem (1) asymptotically. While the choice of g as the absolute value gives the trimmed Lasso penalty and leads to exact sparsity in the non-asymptotic regime (*c.f.* Theorem 2.3) , Proposition 2.5 suggests that the projected penalty functions have potential utility in attaining approximately sparse estimators. We will return to the penalties π_k^g again in Section 4 to connect the trimmed Lasso to nonconvex penalty methods.

Before concluding this section, we briefly consider a projected penalty function that is different than the trimmed Lasso. As noted above, if $g(x) = x^2/2$, then the corresponding penalty function is the trimmed ridge penalty $\sum_{i>k} |\beta_{(i)}|^2/2$. The estimation procedure is then

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2} \sum_{i>k} |\beta_{(i)}|^2,$$

or equivalently in decomposed form (*c.f.* Section 2.3),[5]

$$\min_{\substack{\boldsymbol{\phi}, \boldsymbol{\epsilon}: \\ \|\boldsymbol{\phi}\|_0 \le k}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}(\boldsymbol{\phi} + \boldsymbol{\epsilon})\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\epsilon}\|_2^2.$$

[4] An extended statement of the convergence claim is included in Appendix B.

[5] Interestingly, if one considers this trimmed ridge regression problem and uses convex envelope techniques [21,60] to relax the constraint $\|\boldsymbol{\phi}\|_0 \le k$, the resulting problem takes the form $\min_{\boldsymbol{\phi}, \boldsymbol{\epsilon}} \|\mathbf{y} - \mathbf{X}(\boldsymbol{\phi} + \boldsymbol{\epsilon})\|_2^2/2 + \lambda\|\boldsymbol{\epsilon}\|_2^2 + \tau\|\boldsymbol{\phi}\|_1$, a sort of "split" variant of the usual elastic net [77], another popular convex method for sparse modeling.

It is not difficult to see that the variable $\boldsymbol{\epsilon}$ can be eliminated to yield

$$\min_{\|\boldsymbol{\phi}\|_0 \le k} \frac{1}{2} \|\mathbf{A}(\mathbf{y} - \mathbf{X}\boldsymbol{\phi})\|_2^2 , \tag{15}$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}')^{1/2}$. It follows that the largest k loadings are found via a modified best subset selection problem under a different loss function—precisely a variant of the ℓ_2 norm. This is in the same spirit of observations made in Section 2.3.

Observation 2.6. *An obvious question is whether the norm in (15) is genuinely different. Observe that this loss function is the same as the usual ℓ_2^2 loss if and only if $\mathbf{A}'\mathbf{A}$ is a non-negative multiple of the identity matrix. It is not difficult to see that this is true iff $\mathbf{X}'\mathbf{X}$ is a non-negative multiple of the identity. In other words, the loss function in (15) is the same as the usual ridge regression loss if and only if $\mathbf{X}$ is (a scalar multiple of) an orthogonal design matrix.*

2.5 Other applications of the trimmed Lasso: the (Discrete) Dantzig Selector

The above discussion which pertains to the least squares loss data-fidelity term can be generalized to other loss functions as well. For example, let us consider a data-fidelity term given by the maximal absolute inner product between the features and residuals, given by $\|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty$. An ℓ_1-penalized version of this data-fidelity term, popularly known as the Dantzig Selector [17, 46], is given by the following linear optimization problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty + \mu\|\boldsymbol{\beta}\|_1. \tag{16}$$

Estimators found via (16) have statistical properties similar to the Lasso. Further, problem (16) may be interpreted as an ℓ_1-approximation to the cardinality constrained version:

$$\min_{\|\boldsymbol{\beta}\|_0 \le k} \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty, \tag{17}$$

that is, the Discrete Dantzig Selector, recently proposed and studied in [56]. The statistical properties of (17) are similar to the best-subset selection problem (1), but may be more attractive from a computational viewpoint as it relies on mixed integer *linear* optimization as opposed to mixed integer *conic* optimization (see [56]).

The trimmed Lasso penalty can also be applied to the data-fidelity term $\|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty$, leading to the following estimator:

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty + \lambda T_k(\boldsymbol{\beta}) + \mu\|\boldsymbol{\beta}\|_1.$$

Similar to the case of the least squares loss function, the above estimator yields k-sparse solutions for any $\mu > 0$ and for $\lambda > 0$ sufficiently large.[6] While this claim follows *a fortiori* by appealing to properties of the Dantzig selector, it nevertheless highlights how any exact penalty method with a separable penalty function can be turned into a trimmed-style problem which offers direct control over the sparsity level.

[6] For the same reason, but instead with the usual Lasso objective, the proof of Theorem 2.3 (see Appendix B) could be entirely omitted; yet, it is instructive to see in the proof there that the trimmed Lasso truly does set the *smallest* entries to zero, and not simply all entries (when λ is large) like the Lasso.

3 A perspective on robustness

We now turn our attention to a deeper exploration of the robustness properties of the trimmed Lasso. We begin by studying the min-min robust analogue of the min-max robust SLOPE penalty; in doing so, we show under which circumstances this analogue is the trimmed Lasso problem. Indeed, in such a regime, the trimmed Lasso can be viewed as an optimistic counterpart to the robust optimization view of the SLOPE penalty. Finally, we turn our attention to an additional min-min robust interpretation of the trimmed Lasso in direct correspondence with the least trimmed squares estimator shown in (5), using the ordinary Lasso as our starting point.

3.1 The trimmed Lasso as a min-min robust analogue of SLOPE

We begin by reconsidering the uncertainty set that gave rise to the SLOPE penalty via the min-max view of robustness as considered in robust optimization:

$$\mathcal{U}_k^\lambda := \left\{ \boldsymbol{\Delta} : \begin{array}{c} \boldsymbol{\Delta} \text{ has at most } k \text{ nonzero} \\ \text{columns and } \|\boldsymbol{\Delta}_i\|_2 \leq \lambda \ \forall i \end{array} \right\}.$$

As per Proposition 1.2, the min-max problem (9), viz.,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda} \frac{1}{2} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2$$

is equivalent to the SLOPE-penalized problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda R_{\mathrm{SLOPE}(\tilde{\mathbf{w}})}(\boldsymbol{\beta}). \tag{18}$$

for the specific choice of $\tilde{\mathbf{w}}$ with $\tilde{w}_1 = \cdots = \tilde{w}_k = 1$ and $\tilde{w}_{k+1} = \cdots = \tilde{w}_p = 0$.

Let us now consider the form of the min-min robust analogue of the the problem (9) for this specific choice of uncertainty set. As per the discussion in Section 1, the min-min analogue takes the form of problem (8), i.e., a variant of total least squares:

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda} \frac{1}{2} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2,$$

or equivalently as the linearly homogenous problem[7]

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \tag{19}$$

It is useful to consider problem (19) with an explicit penalization (or regularization) on $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 + r(\boldsymbol{\beta}), \tag{20}$$

where $r(\cdot)$ is, say, a norm (the use of lowercase is to distinguish from the function R in Section 1).

As described in the following theorem, this min-min robustness problem (20) is equivalent to the trimmed Lasso problem for specific choices of r. The proof is contained in Appendix B.

[7] In what follows, the linear homogeneity is useful primarily for simplicity of analysis, *c.f.* [9, ch. 12]. Indeed, the conversion to linear homogeneous functions is often hidden in equivalence results like Proposition 1.2.

Theorem 3.1. *For any k, $\lambda > 0$, and norm r, the problem* (20) *can be rewritten exactly as*

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + r(\boldsymbol{\beta}) - \lambda \sum_{i=1}^{k} |\beta_{(i)}| \\ \text{s. t.} \quad & \lambda \sum_{i=1}^{k} |\beta_{(i)}| \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2. \end{aligned}$$

We have the following as an immediate corollary:

Corollary 3.2. *For the choice of $r(\boldsymbol{\beta}) = \tau\|\boldsymbol{\beta}\|_1$, where $\tau > \lambda$, the problem* (20) *is precisely*

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + (\tau - \lambda)\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\beta}) \\ \text{s. t.} \quad & \lambda \sum_{i=1}^{k} |\beta_{(i)}| \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2. \end{aligned} \tag{21}$$

In particular, when $\lambda > 0$ is small, it is approximately equal (in a precise sense)[8] *to the trimmed Lasso problem*

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + (\tau - \lambda)\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\beta}).$$

In words, the min-min problem (20) (with an ℓ_1 regularization on $\boldsymbol{\beta}$) can be written as a variant of a trimmed Lasso problem, subject to an additional constraint. It is instructive to consider both the objective and the constraint of problem (21). To begin, the objective has a combined penalty on $\boldsymbol{\beta}$ of $(\tau - \lambda)\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\beta})$. This can be thought of as the more general form of the penalty T_k. Namely, one can consider the penalty $T_{\mathbf{x}}$ (with $0 \leq x_1 \leq x_2 \leq \cdots \leq x_p$ fixed) defined as

$$T_{\mathbf{x}}(\boldsymbol{\beta}) := \sum_{i=1}^{p} x_i |\beta_{(i)}|.$$

In this notation, the objective of (21) can be rewritten as $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + T_{\mathbf{x}}(\boldsymbol{\beta})$, with

$$\mathbf{x} = (\underbrace{\tau - \lambda, \ldots, \tau - \lambda}_{k \text{ times}}, \underbrace{\tau, \ldots, \tau}_{p-k \text{ times}}).$$

In terms of the constraint of problem (21), note that it takes the form of a model-fitting constraint: namely, λ controls a trade-off between model fit $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2$ and model complexity measured via the SLOPE norm $\sum_{i=1}^{k} |\beta_{(i)}|$.

Having described the structure of problem (21), a few remarks are in order. First of all, the trimmed Lasso problem (with an additional ℓ_1 penalty on $\boldsymbol{\beta}$) can be interpreted as (a close approximation to) a min-min robust problem, at least in the regime when λ is small; this provides an interesting contrast to the sparse-modeling regime when λ is large (*c.f.* Theorem 2.3). Moreover, the trimmed Lasso is a min-min robust problem in a way that is the *optimistic* analogue of its min-max counterpart, namely, the SLOPE-penalized problem (18). Finally, Theorem 3.1 gives a natural representation of the trimmed Lasso problem in a way that directly suggests why methods from difference-of-convex optimization [2] are relevant (see Section 5).

[8] For a precise characterization and extended discussion, see Appendix B and Theorem B.2. The informal statement here is sufficient for the purposes of our present discussion.

The general SLOPE penalty

Let us briefly remark upon SLOPE in its most general form (with general $\mathbf{w}$); again we will see that this leads to a more general trimmed Lasso as its (approximate) min-min counterpart. In its most general form, the SLOPE-penalized problem (18) can be written as the min-max robust problem (9) with choice of uncertainty set

$$\mathcal{U}_{\mathbf{w}}^{\lambda} = \left\{ \mathbf{\Delta} : \|\mathbf{\Delta}\boldsymbol{\phi}\|_2 \leq \lambda \sum_i w_i |\phi_{(i)}| \ \forall \boldsymbol{\phi} \right\}$$

(see Appendix A). In this case, the penalized, homogenized min-min robust counterpart, analogous to problem (20), can be written as follows:

Proposition 3.3. *For any k, $\lambda > 0$, and norm r, the problem*

$$\min_{\boldsymbol{\beta}} \min_{\mathbf{\Delta} \in \mathcal{U}_{\mathbf{w}}^{\lambda}} \|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}\|_2 + r(\boldsymbol{\beta}) \tag{22}$$

can be rewritten exactly as

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}} & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + r(\boldsymbol{\beta}) - \lambda R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\beta}) \\ \text{s. t.} & \lambda R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\beta}) \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2. \end{array}$$

For the choice of $r(\boldsymbol{\beta}) = \tau\|\boldsymbol{\beta}\|_1$, where $\tau > \lambda w_1$, the problem (22) is

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}} & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + T_{\tau\mathbf{1} - \lambda\mathbf{w}}(\boldsymbol{\beta}) \\ \text{s. t.} & \lambda R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\beta}) \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2. \end{array}$$

In particular, when $\lambda > 0$ is sufficiently small, problem (22) is approximately equal to the generalized trimmed Lasso problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + T_{\tau\mathbf{1} - \lambda\mathbf{w}}(\boldsymbol{\beta}).$$

Put plainly, the general form of the SLOPE penalty leads to a generalized form of the trimmed Lasso, precisely as was true for the simplified version considered in Theorem 3.1.

3.2 Another min-min interpretation

We close our discussion of robustness by considering another min-min representation of the trimmed Lasso. We use the ordinary Lasso problem as our starting point and show how a modification in the same spirit as the min-min robust least trimmed squares estimator in (5) leads directly to the trimmed Lasso.

To proceed, we begin with the usual Lasso problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{23}$$

As per Proposition 1.1, this problem is equivalent to the min-max robust problem (9) with uncertainty set $\mathcal{U} = \mathcal{L}^{\lambda} = \{\mathbf{\Delta} : \|\mathbf{\Delta}_i\|_2 \leq \lambda \ \forall i\}$:

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta} \in \mathcal{L}^{\lambda}} \frac{1}{2}\|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}\|_2^2. \tag{24}$$

In this view, the usual Lasso (23) can be thought of as a least squares method which takes into account certain feature-wise adversarial perturbations of the matrix $\mathbf{X}$. The net result is that the adversarial approach penalizes all loadings equally (with coefficient λ).

Using this setup and Theorem 2.3, we can re-express the trimmed Lasso problem ($\mathrm{TL}_{\lambda,k}$) in the equivalent min-min form

$$\min_{\boldsymbol{\beta}} \min_{\substack{I\subseteq\{1,\ldots,p\}:\\ |I|=p-k}} \max_{\boldsymbol{\Delta}\in\mathcal{L}_I^\lambda} \frac{1}{2}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2, \tag{25}$$

where $\mathcal{L}_I^\lambda \subseteq \mathcal{L}^\lambda$ requires that the columns of $\boldsymbol{\Delta} \in \mathcal{L}_I^\lambda$ are supported on I:

$$\mathcal{L}_I^\lambda = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}_i\|_2 \le \lambda \ \forall i,\ \boldsymbol{\Delta}_i = \mathbf{0} \ \forall i \notin I\}.$$

While the adversarial min-max approach in problem (24) would attempt to “corrupt” all p columns of $\mathbf{X}$, in estimating $\boldsymbol{\beta}$ we have the power to optimally discard k out of the p corruptions to the columns (corresponding to I^c). In this sense, the trimmed Lasso in the min-min robust form (25) acts in a similar spirit to the min-min, robust-statistical least trimmed squares estimator shown in problem (6).

4 Connection to nonconvex penalty methods

In this section, we explore the connection between the trimmed Lasso and existing, popular nonconvex (component-wise separable) penalty functions used for sparse modeling. We begin in Section 4.1 with a brief overview of existing approaches. In Section 4.2 we then highlight how these relate to the trimmed Lasso, making the connection more concrete with examples in Section 4.3. Then in Section 4.4 we exactly characterize the connection between the trimmed Lasso and the clipped Lasso [76]. In doing so, we show that the trimmed Lasso subsumes the clipped Lasso; further, we provide a necessary and sufficient condition for when the containment is strict. Finally, in Section 4.5 we comment on the special case of unbounded penalty functions.

4.1 Setup and Overview

Our focus throughout will be the penalized M-estimation problem of the form

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \sum_{i=1}^{p} \rho(|\beta_i|; \mu, \gamma), \tag{26}$$

where μ represents a (continuous) parameter controlling the desired level of sparsity of $\boldsymbol{\beta}$ and γ is a parameter controlling the quality of the approximation of the indicator function $I\{|\beta| > 0\}$. A variety of nonconvex penalty functions and their description in this format is shown in Table 1 (for a general discussion, see [75]). In particular, for each of these functions we observe that

$$\lim_{\gamma\to\infty} \rho(|\beta|; \mu, \gamma) = \mu \cdot I\{|\beta| > 0\}.$$

It is particularly important to note the *separable* nature of the penalty functions appearing in (26)—namely, each coordinate β_i is penalized (via ρ) independently of the other coordinates.

Our primary focus will be on the bounded penalty functions (clipped Lasso, MCP, and SCAD), all of which take the form

$$\rho(|\beta|; \mu, \gamma) = \mu \min\{g(|\beta|; \mu, \gamma), 1\} \tag{27}$$

Name	Definition	Auxiliary Functions
Clipped Lasso [76]	$\mu\min\{\gamma\lvert\beta\rvert,1\}$	$g_1(\lvert\beta\rvert) = \begin{cases} 2\gamma\lvert\beta\rvert - \gamma^2\beta^2, & \lvert\beta\rvert \le 1/\gamma, \\ 1, & \lvert\beta\rvert > 1/\gamma. \end{cases}$ $g_2(\lvert\beta\rvert) = \begin{cases} \lvert\beta\rvert/(\gamma\mu), & \lvert\beta\rvert \le 1/\gamma, \\ \frac{\beta^2+(2/\gamma-4\mu\gamma)\lvert\beta\rvert+1/\gamma^2}{4\mu-4\mu^2\gamma^2}, & 1/\gamma < \lvert\beta\rvert \le 2\mu\gamma - 1/\gamma, \\ 1, & \lvert\beta\rvert > 2\mu\gamma - 1/\gamma. \end{cases}$
MCP [74]	$\mu\min\{g_1(\lvert\beta\rvert),1\}$	
SCAD [33]	$\mu\min\{g_2(\lvert\beta\rvert),1\}$	
ℓ_q $(0<q<1)$ [36, 37]	$\mu\lvert\beta\rvert^{1/\gamma}$	
Log [37]	$\mu\log(\gamma\lvert\beta\rvert+1)/\log(\gamma+1)$	

Table 1: Nonconvex penalty functions $\rho(|\beta|;\mu,\gamma)$ represented as in (26). The precise parametric representation is different than their original presentation but they are equivalent. We have taken care to normalize the different penalty functions so that μ is the sparsity parameter and γ corresponds to the approximation of the indicator $I\{|\beta|>0\}$. For SCAD, it is usually recommended to set $2\mu > 3/\gamma^2$.

where g is an increasing function of $|\beta|$. We will show that in this case, the problem (26) can be rewritten exactly as an estimation problem with a (non-separable) trimmed penalty function:

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \mu \sum_{i=\ell+1}^{p} g(|\beta_{(i)}|) \tag{28}$$

for some $\ell \in \{0,1,\ldots,p\}$ (note the appearance of the projected penalties π_k^g as considered in Section 2.4). In the process of doing so, we will also show that, in general, (28) cannot be solved via the separable-penalty estimation approach of (26), and so the trimmed estimation problem leads to a richer class of models. Throughout we will often refer to (28) (taken generically over all choices of ℓ) as the *trimmed counterpart* of the separable estimation problem (26).

4.2 Reformulating the problem (26)

Let us begin by considering penalty functions ρ of the form (27) with g a non-negative, increasing function of $|\beta|$. Observe that for any $\boldsymbol{\beta}$ we can rewrite $\sum_{i=1}^{p} \min\{g(|\beta_i|),1\}$ as

$$\begin{aligned} &\min\left\{\sum_{i=1}^{p} g(|\beta_{(i)}|), 1+\sum_{i=2}^{p} g(|\beta_{(i)}|), \ldots, p-1+g(|\beta_{(p)}|), p\right\} \\ &= \min_{\ell\in\{0,\ldots,p\}} \left\{\ell + \sum_{i>\ell} g(|\beta_{(i)}|)\right\}. \end{aligned}$$

It follows that (26) can be rewritten *exactly* as

$$\min_{\substack{\boldsymbol{\beta}, \\ \ell\in\{0,\ldots,p\}}} \left(L(\boldsymbol{\beta}) + \mu \sum_{i>\ell} g(|\beta_{(i)}|) + \mu\ell \right) \tag{29}$$

An immediate consequence is the following theorem:

Theorem 4.1. *If $\boldsymbol{\beta}^*$ is an optimal solution to (26), where $\rho(|\beta|;\mu,\gamma) = \mu\min\{g(|\beta|;\mu,\gamma),1\}$, then there exists some $\ell^* \in \{0,\ldots,p\}$ so that $\boldsymbol{\beta}^*$ is optimal to its trimmed counterpart*

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \mu \sum_{i>\ell^*} g(|\beta_{(i)}|).$$

In particular, the choice of $\ell^* = |\{i : g(|\beta_i^*|) \geq 1\}|$ suffices. Conversely, if $\boldsymbol{\beta}^*$ is an optimal solution to (29), then $\boldsymbol{\beta}^*$ in an optimal solution to (26).

It follows that the estimation problem (26), which decouples each loading β_i in the penalty function, can be solved using "trimmed" estimation problems of the form (28) with a trimmed penalty function that couples the loadings and only penalizes the $p - \ell^*$ smallest. Because the trimmed penalty function is generally nonconvex by nature, we will focus on comparing it with other nonconvex penalties for the remainder of the section.

4.3 Trimmed reformulation examples

We now consider the structure of the estimation problem (26) and the corresponding trimmed estimation problem for the clipped Lasso and MCP penalties. We use the ℓ_2^2 loss throughout.

Clipped Lasso

The clipped (or capped, or truncated) Lasso penalty [64, 76] takes the component-wise form

$$\rho(|\beta|;\mu,\gamma) = \mu\min\{\gamma|\beta|,1\}.$$

Therefore, in our notation, g is a multiple of the absolute value function. A plot of ρ is shown in Figure 1a. In this case, the estimation problem with ℓ_2^2 loss is

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mu\sum_i \min\{\gamma|\beta_i|,1\}. \tag{30}$$

It follows that the corresponding trimmed estimation problem (*c.f.* Theorem 4.1) is exactly the trimmed Lasso problem studied earlier, namely,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mu\gamma T_k(\boldsymbol{\beta}). \tag{31}$$

A distinct advantage of the trimmed Lasso formulation (31) over the traditional clipped Lasso formulation (30) is that it offers direct control over the desired level of sparsity vis-à-vis the discrete parameter k. We perform a deeper analysis of the two problems in Section 4.4.

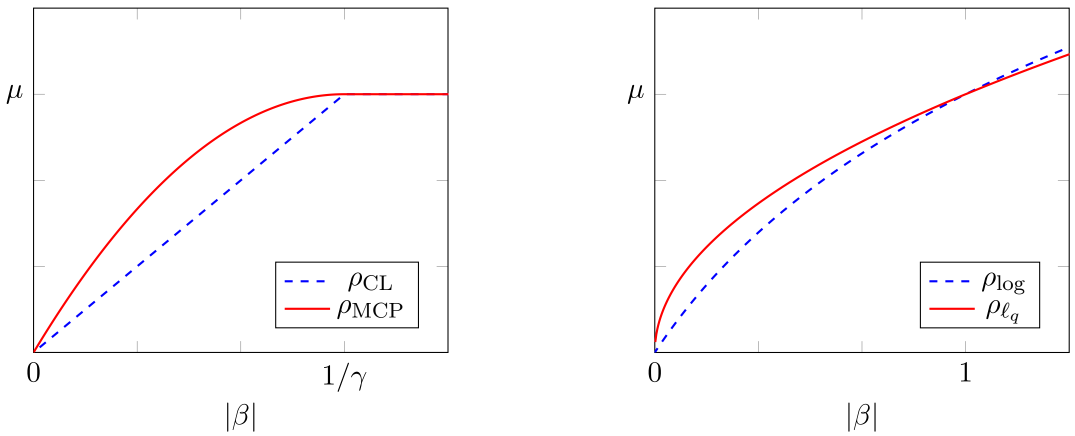
MCP

The MCP penalty takes the component-wise form

$$\rho(|\beta|;\mu,\gamma) = \mu\min\{g(|\beta|),1\}$$

where g is any function with $g(|\beta|) = 2\gamma|\beta| - \gamma^2\beta^2$ whenever $|\beta| \leq 1/\gamma$ and $g(|\beta|) \geq 1$ whenever $|\beta| > 1/\gamma$. An example of one such g is shown in Table 1. A plot of ρ is shown in Figure 1a. Another valid choice of g is $g(|\beta|) = \max\{2\gamma|\beta| - \gamma^2\beta^2, \gamma|\beta|\}$. In this case, the trimmed counterpart is

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \mu\gamma\sum_{i>\ell} \max\left\{2|\beta_{(i)}| - \gamma\beta_{(i)}^2, |\beta_{(i)}|\right\}.$$



(a) Clipped Lasso and MCP

(b) Log and ℓ_q

Figure 1: Plots of $\rho(|\beta|;\mu,\gamma)$ for some of the penalty functions in Table 1.

Note that this problem is amenable to the same class of techniques as applied to the trimmed Lasso problem in the form (31) because of the increasing nature of g, although the subproblems with respect to $\boldsymbol{\beta}$ are no longer convex (although it is a usual MCP estimation problem which is well-suited to convex optimization approaches; see [55]). Also observe that we can separate the penalty function into a trimmed Lasso component and another component:

$$\sum_{i>\ell}|\beta_{(i)}| \quad \text{and} \quad \sum_{i>\ell}\left(|\beta_{(i)}|-\gamma\beta_{(i)}^2\right)_+ .$$

Observe that the second component is uniformly bounded above by $(p-\ell)/(4\gamma)$, and so as $\gamma\to\infty$, the trimmed Lasso penalty dominates.

4.4 The generality of trimmed estimation

We now turn our focus to more closely studying the relationship between the separable-penalty estimation problem (26) and its trimmed estimation counterpart. The central problems of interest are the clipped Lasso and its trimmed counterpart, viz., the trimmed Lasso:[9]

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\mu\sum_i \min\{\gamma|\beta_i|,1\} \tag{$\text{CL}_{\mu,\gamma}$}$$

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda T_\ell\left(\boldsymbol{\beta}\right). \tag{$\text{TL}_{\lambda,\ell}$}$$

As per Theorem 4.1, if $\boldsymbol{\beta}^*$ is an optimal solution to ($\text{CL}_{\mu,\gamma}$), then $\boldsymbol{\beta}^*$ is an optimal solution to ($\text{TL}_{\lambda,\ell}$), where $\lambda=\mu\gamma$ and $\ell=|\{i:|\beta_i^*|\geq 1/\gamma\}|$. We now consider the converse: given some $\lambda>0$ and $\ell\in\{0,1,\ldots,p\}$ and a solution $\boldsymbol{\beta}^*$ to ($\text{TL}_{\lambda,\ell}$), when does there exist some $\mu,\gamma>0$ so that $\boldsymbol{\beta}^*$

[9] One may be concerned about the well-definedness of such problems (e.g. as guaranteed vis-à-vis coercivity of the objective, *c.f.* [60]). In all the results of Section 4.4, it is possible to add a regularizer $\eta\|\boldsymbol{\beta}\|_1$ for some fixed $\eta>0$ to both ($\text{CL}_{\mu,\gamma}$) and ($\text{TL}_{\lambda,\ell}$) and the results remain valid, *mutatis mutandis*. The addition of this regularizer implies coercivity of the objective functions and, consequently, that the minimum is indeed well-defined. For completeness, we note a technical reason for a choice of $\eta\|\boldsymbol{\beta}\|_1$ is its positive homogeneity; thus, the proof technique of Lemma 4.3 easily adapts to this modification.

is an optimal solution to $(\mathrm{CL}_{\mu,\gamma})$? As the following theorem suggests, the existence of such a γ is closely connected to an underlying discrete form of "convexity" of the sequence of problems $(\mathrm{TL}_{\lambda,k})$ for $k \in \{0, 1, \dots, p\}$. We will focus on the case when $\lambda = \mu\gamma$, as this is the natural correspondence of parameters in light of Theorem 4.1.

Theorem 4.2. *If $\lambda > 0$, $\ell \in \{0, \dots, p\}$, and $\boldsymbol{\beta}^*$ is an optimal solution to $(\mathrm{TL}_{\lambda,\ell})$, then there exist $\mu, \gamma > 0$ with $\mu\gamma = \lambda$ and so that $\boldsymbol{\beta}^*$ is an optimal solution to $(\mathrm{CL}_{\mu,\gamma})$ if and only if*

$$Z(\mathrm{TL}_{\lambda,\ell_e}) < \frac{j-\ell_e}{j-i} Z(\mathrm{TL}_{\lambda,i}) + \frac{\ell_e - i}{j-i} Z(\mathrm{TL}_{\lambda,j}) \tag{32}$$

for all $0 \le i < \ell_e < j \le p$, where $Z(\mathrm{P})$ denotes the optimal objective value to optimization problem (P) and $\ell_e = \min\{\ell, \|\boldsymbol{\beta}^*\|_0\}$.

Let us note why we refer to the condition in (32) as a discrete analogue of convexity of the sequence $\{z_k := Z(\mathrm{TL}_{\lambda,k}),\ k = 0, \dots, p\}$. In particular, observe that this sequence satisfies the condition of Theorem 4.2 if and only if the function defined as the linear interpolation between the points $(0, z_0)$, $(1, z_1)$, $\dots$, and (p, z_p) is strictly convex about the point (ℓ, z_ℓ).[10]

Before proceeding with the proof of the theorem, we state and prove a technical lemma about the structure of $(\mathrm{TL}_{\lambda,\ell})$.

Lemma 4.3. *Fix $\lambda > 0$ and suppose that $\boldsymbol{\beta}^*$ is optimal to $(\mathrm{TL}_{\lambda,\ell})$.*

(a) The optimal objective value of $(\mathrm{TL}_{\lambda,\ell})$ is $Z(\mathrm{TL}_{\lambda,\ell}) = (\|\mathbf{y}\|_2^2 - \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2)/2$.

(b) If $\boldsymbol{\beta}^*$ is also optimal to $(\mathrm{TL}_{\lambda,\ell'})$, where $\ell < \ell'$, then $\|\boldsymbol{\beta}^*\|_0 \le \ell$ and $\boldsymbol{\beta}^*$ is optimal to $(\mathrm{TL}_{\lambda,j})$ for all integral j with $\ell < j < \ell'$.

(c) If $\kappa := \|\boldsymbol{\beta}^*\|_0 < \ell$, then $\boldsymbol{\beta}^*$ is also optimal to $(\mathrm{TL}_{\lambda,\kappa})$, $(\mathrm{TL}_{\lambda,\kappa+1})$, $\dots$, and $(\mathrm{TL}_{\lambda,\ell-1})$. Further, $\boldsymbol{\beta}^*$ is not *optimal to $(\mathrm{TL}_{\lambda,0})$, $(\mathrm{TL}_{\lambda,1})$, $\dots$, nor $(\mathrm{TL}_{\lambda,\kappa-1})$.*

Proof. Suppose $\boldsymbol{\beta}^*$ is optimal to $(\mathrm{TL}_{\lambda,\ell})$. Define

$$a(\epsilon) := \|\mathbf{y} - \epsilon\mathbf{X}\boldsymbol{\beta}^*\|_2^2/2 + \epsilon\lambda T_\ell(\boldsymbol{\beta}^*).$$

By the optimality of $\boldsymbol{\beta}^*$, $a(\epsilon) \ge a(1)$ for all $\epsilon \ge 0$. As a is a polynomial with degree at most two, one must have that $a'(1) = 0$. This implies that

$$a'(1) = -\langle \mathbf{y}, \mathbf{X}\boldsymbol{\beta}^* \rangle + \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \lambda T_\ell(\boldsymbol{\beta}^*) = 0.$$

Adding $(\|\mathbf{y}\|_2^2 - \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2)/2$ to both sides, the desired result of part (a) follows.

Now suppose that $\boldsymbol{\beta}^*$ is also optimal to $(\mathrm{TL}_{\lambda,\ell'})$, where $\ell' > \ell$. By part (a), one must necessarily have that $Z(\mathrm{TL}_{\lambda,\ell}) = Z(\mathrm{TL}_{\lambda,\ell'}) = (\|\mathbf{y}\|_2^2 - \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2)/2$. Inspecting $Z(\mathrm{TL}_{\lambda,\ell}) - Z(\mathrm{TL}_{\lambda,\ell'})$, we see that

$$0 = Z(\mathrm{TL}_{\lambda,\ell}) - Z(\mathrm{TL}_{\lambda,\ell'}) = \lambda \sum_{i=\ell+1}^{\ell'} |\beta^*_{(i)}|.$$

Hence, $|\beta^*_{(\ell+1)}| = 0$ and therefore $\|\boldsymbol{\beta}^*\|_0 \le \ell$.

Finally, for any integral j with $\ell \le j \le \ell'$, one always has that $Z(\mathrm{TL}_{\lambda,\ell}) \ge Z(\mathrm{TL}_{\lambda,j}) \ge Z(\mathrm{TL}_{\lambda,\ell'})$. As per the preceding argument, $Z(\mathrm{TL}_{\lambda,\ell}) = Z(\mathrm{TL}_{\lambda,\ell})$ and so $Z(\mathrm{TL}_{\lambda,\ell}) = Z(\mathrm{TL}_{\lambda,j})$, and therefore $\boldsymbol{\beta}^*$ must also be optimal to $(\mathrm{TL}_{\lambda,j})$ by applying part (a). This completes part (b).

Part (c) follows from a straightforward inspection of objective functions and using the fact that $Z(\mathrm{TL}_{\lambda,j}) \ge Z(\mathrm{TL}_{\lambda,\ell})$ whenever $j \le \ell$. □

[10] To be precise, we mean that the real-valued function that is a linear interpolation of the points has a subdifferential at the point (ℓ, z_ℓ) which is an interval of strictly positive width.

Using this lemma, we can now proceed with the proof of the theorem.

Proof of Theorem 4.2. Let $z_k = Z(\mathrm{TL}_{\lambda,k})$ for $k \in \{0, 1, \ldots, p\}$. Suppose that $\mu, \gamma > 0$ is so that $\lambda = \mu\gamma$ and $\boldsymbol{\beta}^*$ is an optimal solution to $(\mathrm{CL}_{\mu,\gamma})$. Let $\ell_e = \min\{\ell, \|\boldsymbol{\beta}^*\|_0\}$. Per equation (29), $\boldsymbol{\beta}^*$ must be optimal to

$$\min_{\boldsymbol{\beta}} \min_{k\in\{0,\ldots,p\}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mu k + \mu\gamma T_k(\boldsymbol{\beta}). \tag{33}$$

Observe that this implies that if k is such that k is a minimizer of $\min_k \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*)$, then $\boldsymbol{\beta}^*$ must be optimal to $(\mathrm{TL}_{\lambda,k})$.

We claim that this observation, combined with Lemma 4.3, implies that

$$\ell_e = \operatorname*{arg\,min}_{k\in\{0,\ldots,p\}} \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*).$$

This can be shown as follows:

(a) Suppose $\ell \le \|\boldsymbol{\beta}^*\|_0$ and so $\ell_e = \min\{\ell, \|\boldsymbol{\beta}^*\|_0\} = \ell$. Therefore, by Lemma 4.3(b), $\boldsymbol{\beta}^*$ is not optimal to $(\mathrm{TL}_{\lambda,j})$ for any $j < \ell$, and thus

$$\min_{k\in\{0,\ldots,p\}} \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*) = \min_{k\in\{\ell,\ldots,p\}} \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*).$$

If $k > \ell$ is such that k is a minimizer of $\min_k \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*)$, then $\boldsymbol{\beta}^*$ must be optimal to $(\mathrm{TL}_{\lambda,k})$ (using the observation), and hence by Lemma 4.3(b), $\|\boldsymbol{\beta}^*\|_0 \le \ell$. Combined with $\ell \le \|\boldsymbol{\beta}^*\|_0$, this implies that $\|\boldsymbol{\beta}^*\|_0 = \ell$. Yet then, $\mu\ell = \mu\ell + \mu\gamma T_\ell(\boldsymbol{\beta}^*) < \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*)$, contradicting the optimality of k. Therefore, we conclude that $\ell_e = \ell$ is the *only* minimizer of $\min_k \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*)$.

(b) Now instead suppose that $\ell_e = \|\boldsymbol{\beta}^*\|_0 < \ell$. Lemma 4.3(c) implies that any optimal solution k to $\min_k \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*)$ must satisfy $k \ge \|\boldsymbol{\beta}^*\|_0$ (by the second part combined with the observation). As before, if $k > \|\boldsymbol{\beta}^*\|_0 = \ell_e$, then $\mu k > \mu\ell_e$, and so k cannot be optimal. As a result, $k = \ell_e = \|\boldsymbol{\beta}^*\|_0$ is the unique minimum.

In either case, we have that ℓ_e is the unique minimizer to $\min_k \mu k + \mu\gamma T_k(\boldsymbol{\beta}^*)$.

It then follows that $Z(\text{problem (33)}) = z_{\ell_e} + \mu\ell_e$. Further, by optimality of $\boldsymbol{\beta}^*$, $z_{\ell_e} + \mu\ell_e < z_i + \mu i$ for all $0 \le i \le p$ with $i \ne \ell_e$. For $0 \le i < \ell_e$, this implies $\mu < (z_i - z_{\ell_e})/(\ell_e - i)$ and for $j > \ell_e$, $\mu > (z_{\ell_e} - z_j)/(j - \ell_e)$. In other words, for $0 \le i < \ell_e < j \le p$,

$$\frac{z_{\ell_e} - z_j}{j - \ell_e} < \frac{z_i - z_{\ell_e}}{\ell_e - i}, \quad \text{i.e.,} \quad z_{\ell_e} < \frac{j - \ell_e}{j - i} z_i + \frac{\ell_e - i}{j - i} z_j.$$

This completes the forward direction. The reverse follows in the same way by taking any μ with

$$\mu \in \left(\max_{j>\ell_e} \frac{z_{\ell_e} - z_j}{j - \ell_e}, \min_{i<\ell_e} \frac{z_i - z_{\ell_e}}{\ell_e - i} \right).$$

□

We briefly remark upon one implication of the proof of Theorem 4.2. In particular, if $\boldsymbol{\beta}^*$ is a solution to $(\mathrm{TL}_{\lambda,\ell})$ and $\ell < \|\boldsymbol{\beta}^*\|_0$, then $\boldsymbol{\beta}^*$ is not the solution to $(\mathrm{TL}_{\lambda,k})$ for any $k \ne \ell$.

An immediate question is whether the convexity condition (32) of Theorem 4.2 always holds. While the sequence $\{Z(\mathrm{TL}_{\lambda,k}) : k = 0, 1, \ldots, p\}$ is always non-increasing, the following example shows that the convexity condition need not hold in general; as a result, there exist instances of the trimmed Lasso problem whose solutions *cannot* be found by solving a clipped Lasso problem.

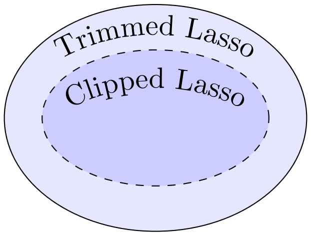


Figure 2: Stylized relation of clipped Lasso and trimmed Lasso models. Every clipped Lasso model can be written as a trimmed Lasso model, but the reverse does not hold in general.

Example 4.4. *Consider the case when $p = n = 2$ with*

$$\mathbf{y} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

Let $\lambda = 1/2$ and $\ell = 1$, and consider $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + |\beta_{(2)}|/2 = \min_{\beta_1,\beta_2}(1-\beta_1+\beta_2)^2/2 + (1+\beta_1-2\beta_2)^2/2 + |\beta_{(2)}|/2$. This has unique optimal solution $\boldsymbol{\beta}^* = (3/2, 1)$ with corresponding objective value $z_1 = 3/4$. One can also compute $z_0 = Z(\mathrm{TL}_{1/2,0}) = 39/40$ and $z_2 = Z(\mathrm{TL}_{1/2,2}) = 0$. Note that $z_1 = 3/4 > (39/40)/2 + (0)/2 = z_0/2 + z_2/2$, and so there do not exist any $\mu, \gamma > 0$ with $\mu\gamma = 1/2$ so that $\boldsymbol{\beta}^*$ is an optimal solution to $(\mathrm{CL}_{\mu,\gamma})$ by Theorem 4.2. Further, it is possible to show that $\boldsymbol{\beta}^*$ is not an optimal solution to $(\mathrm{CL}_{\mu,\gamma})$ for any *choice of $\mu, \gamma \geq 0$. (See Appendix B.)*

An immediate corollary of this example, combined with Theorem 4.1, is that the class of trimmed Lasso models contains the class of clipped Lasso models as a *proper* subset, regardless of whether we restrict our attention to $\lambda = \mu\gamma$. In this sense, the trimmed Lasso models comprise a richer set of models. The relationship is depicted in stylized form in Figure 2.

Limit analysis

It is important to contextualize the results of this section as $\lambda \to \infty$. This corresponds to $\gamma \to \infty$ for the clipped Lasso problem, in which case $(\mathrm{CL}_{\mu,\gamma})$ converges to the penalized form of subset selection:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_0. \tag{$\mathrm{CL}_{\mu,\infty}$}$$

Note that penalized problems for all of the penalties listed in Table 1 have this as their limit as $\gamma \to \infty$. On the other hand, $(\mathrm{TL}_{\lambda,\ell})$ converges to constrained best subset selection:

$$\min_{\|\boldsymbol{\beta}\|_0 \leq \ell} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{$\mathrm{TL}_{\infty,k}$}$$

Indeed, from this comparison it now becomes clear why a convexity condition of the form in Theorem 4.2 appears in describing when the clipped Lasso solves the trimmed Lasso problem. In particular, the conditions under $(\mathrm{CL}_{\mu,\infty})$ solves the constrained best subset selection problem $(\mathrm{TL}_{\infty,k})$ are precisely those in Theorem 4.2.

4.5 Unbounded penalty functions

We close this section by now considering nonconvex penalty functions which are unbounded and therefore do not take the form $\mu \min\{g(|\beta|), 1\}$. Two such examples are the ℓ_q penalty $(0 < q < 1)$

and the log family of penalties as shown in Table 1 and depicted in Figure 1b. Estimation problems with these penalties can be cast in the form

$$\min_{\boldsymbol{\phi}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\phi}\|_2^2 + \mu \sum_{i=1}^{p} g(|\phi_i|;\gamma) \tag{34}$$

where $\mu, \gamma > 0$ are parameters, g is an unbounded and strictly increasing function, and $g(|\phi_i|;\gamma) \xrightarrow{\gamma\to\infty} I\{|\phi_i| > 0\}$. The change of variables in (34) is intentional and its purpose will become clear shortly.

Observe that because g is now unbounded, there exists some $\overline{\lambda} = \overline{\lambda}(\mathbf{y}, \mathbf{X}, \mu, \gamma) > 0$ so that for all $\lambda > \overline{\lambda}$ any optimal solution $(\boldsymbol{\phi}^*, \boldsymbol{\epsilon}^*)$ to the problem

$$\min_{\boldsymbol{\phi},\boldsymbol{\epsilon}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}(\boldsymbol{\phi}+\boldsymbol{\epsilon})\|_2^2 + \lambda\|\boldsymbol{\epsilon}\|_1 + \mu \sum_{i=1}^{p} g(|\phi_i|;\gamma) \tag{35}$$

has $\boldsymbol{\epsilon}^* = \mathbf{0}$.[11] Therefore, (34) is a special case of (35). We claim that in the limit as $\gamma \to \infty$ (all else fixed), that (35) can be written exactly as a trimmed Lasso problem $(\mathrm{TL}_{\lambda,k})$ for some choice of k and with the identification of variables $\boldsymbol{\beta} = \boldsymbol{\phi} + \boldsymbol{\epsilon}$.

We summarize this as follows:

Proposition 4.5. *As $\gamma \to \infty$, the penalized estimation problem (34) is a special case of the trimmed Lasso problem.*

Proof. This can be shown in a straightforward manner: namely, as $\gamma \to \infty$, (35) becomes

$$\min_{\boldsymbol{\phi},\boldsymbol{\epsilon}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}(\boldsymbol{\phi}+\boldsymbol{\epsilon})\|_2^2 + \lambda\|\boldsymbol{\epsilon}\|_1 + \mu\|\boldsymbol{\phi}\|_0$$

which can be in turn written as

$$\min_{\substack{\boldsymbol{\phi},\boldsymbol{\epsilon}:\\ \|\boldsymbol{\phi}\|_0 \le k}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}(\boldsymbol{\phi}+\boldsymbol{\epsilon})\|_2^2 + \lambda\|\boldsymbol{\epsilon}\|_1$$

for some $k \in \{0, 1, \ldots, p\}$. But as per the observations of Section 2.3, this is exactly $(\mathrm{TL}_{\lambda,k})$ using a change of variables $\boldsymbol{\beta} = \boldsymbol{\phi} + \boldsymbol{\epsilon}$. In the case when λ is sufficiently large, we necessarily have $\boldsymbol{\beta} = \boldsymbol{\phi}$ at optimality. □

While this result is not surprising (given that as $\gamma \to \infty$ the problem is (34) is precisely penalized best subset selection), it is useful for illustrating the connection between (34) and the trimmed Lasso problem even when the trimmed Lasso parameter λ is not necessarily large: in particular, $(\mathrm{TL}_{\lambda,k})$ can be viewed as estimating $\boldsymbol{\beta}$ as the sum of two components—a sparse component $\boldsymbol{\phi}$ and small-norm ("noise") component $\boldsymbol{\epsilon}$. Indeed, in this setup, λ precisely controls the desirable level of allowed "noise" in $\boldsymbol{\beta}$. From this intuitive perspective, it becomes clearer why the trimmed Lasso type approach represents a continuous connection between best subset selection (λ large) and ordinary least squares (λ small).

We close this section by making the following observation regarding problem (35). In particular, observe that regardless of λ, we can rewrite this as

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{p} \widetilde{\rho}(|\beta_i|)$$

[11] The proof involves a straightforward modification of an argument along the lines of that given in Theorem 2.3. Also note that we can choose $\overline{\lambda}$ so that it is decreasing in γ, *ceteris paribus*.

where $\widetilde{\rho}(|\beta_i|)$ is the new penalty function defined as

$$\widetilde{\rho}(|\beta_i|) = \min_{\phi+\epsilon=\beta_i} \lambda|\epsilon| + \mu g(|\phi|;\gamma).$$

For the unbounded and concave penalty functions shown in Table 1, this new penalty function is quasi-concave and can be rewritten easily in closed form. For example, for the ℓ_q penalty $\rho(|\beta_i|) = \mu|\beta_i|^{1/\gamma}$ (where $\gamma > 1$), the new penalty function is

$$\widetilde{\rho}(|\beta_i|) = \min\{\mu|\beta_i|^{1/\gamma}, \lambda|\beta_i|\}.$$

5 Algorithmic Approaches

We now turn our attention to algorithms for estimation with the trimmed Lasso penalty. Our principle focus throughout will be the same problem considered in Theorem 2.3, namely

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k(\boldsymbol{\beta}) + \eta\|\boldsymbol{\beta}\|_1 \tag{36}$$

We present three possible approaches to finding potential solutions to (36): a first-order-based alternating minimization scheme that has accompanying local optimality guarantees and was first studied in [39, 72]; an augmented Lagrangian approach that appears to perform noticeably better, despite lacking optimality guarantees; and a convex envelope approach. We contrast these methods with approaches for certifying global optimality of solutions to (36) (described in [69]) and include an illustrative computational example. Implementations of the various algorithms presented can be found at

https://github.com/copenhaver/trimmedlasso.

5.1 Upper bounds via convex methods

We start by focusing on the application of convex optimization methods to finding to finding potential solutions to (36). Technical details are contained in Appendix C.

Alternating minimization scheme

We begin with a first-order-based approach for obtaining a locally optimal solution of (36) as described in [39, 72]. The key tool in this approach is the theory of difference of convex optimization ("DCO") [1, 2, 66]. Set the following notation:

$$\begin{array}{rcl} f(\boldsymbol{\beta}) & = & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + \lambda T_k(\boldsymbol{\beta}) + \eta\|\boldsymbol{\beta}\|_1, \\ f_1(\boldsymbol{\beta}) & = & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + (\eta+\lambda)\|\boldsymbol{\beta}\|_1, \\ f_2(\boldsymbol{\beta}) & = & \lambda \sum_{i=1}^k |\beta_{(i)}|. \end{array}$$

Let us make a few simple observations:

(a) Problem (36) can be written as $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$.

(b) For all $\boldsymbol{\beta}$, $f(\boldsymbol{\beta}) = f_1(\boldsymbol{\beta}) - f_2(\boldsymbol{\beta})$.

(c) The functions f_1 and f_2 are convex.

While simple, these observations enable one to apply the theory of DCO, which focuses precisely on problems of the form

$$\min_{\boldsymbol{\beta}} f_1(\boldsymbol{\beta}) - f_2(\boldsymbol{\beta}),$$

where f_1 and f_2 are convex. In particular, the optimality conditions for such a problem have been studied extensively [2]. Let us note that while it may appear that the representation of the objective f as $f_1 - f_2$ might otherwise seem like an artificial algebraic manipulation, the min-min representation in Theorem 3.1 shows how such a difference-of-convex representation can arise naturally.

We now discuss an associated alternating minimization scheme (or equivalently, a sequential linearization scheme), shown in Algorithm 1, for finding local optima of (36). The convergence properties of Algorithm 1 can be summarized as follows:[12]

Theorem 5.1 ([39], Convergence of Algorithm 1)**.** *(a) The sequence $\{f(\boldsymbol{\beta}^\ell) : \ell = 0, 1, \ldots\}$, where $\boldsymbol{\beta}^\ell$ are as found in Algorithm 1, is non-increasing.*

(b) The set $\{\boldsymbol{\gamma}^\ell : \ell = 0, 1, \ldots\}$ is finite and eventually periodic.

(c) Algorithm 1 converges in a finite number of iterations to local minimum of (36).

(d) The rate of convergence of $f(\boldsymbol{\beta}^\ell)$ is linear.

Algorithm 1 An alternating scheme for computing a local optimum to (36)

1. Initialize with any $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ ($\ell = 0$); for $\ell \geq 0$, repeat Steps 2-3 until $f(\boldsymbol{\beta}^\ell) = f(\boldsymbol{\beta}^{\ell+1})$.

2. Compute $\boldsymbol{\gamma}^\ell$ as

$$\boldsymbol{\gamma}^\ell \in \begin{array}{ll} \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} & \langle \boldsymbol{\gamma}, \boldsymbol{\beta}^\ell \rangle \\ \text{s. t.} & \sum_i |\gamma_i| \leq \lambda k \\ & |\gamma_i| \leq \lambda \ \forall i. \end{array} \tag{37}$$

3. Compute $\boldsymbol{\beta}^{\ell+1}$ as

$$\boldsymbol{\beta}^{\ell+1} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + (\eta + \lambda)\|\boldsymbol{\beta}\|_1 - \langle \boldsymbol{\beta}, \boldsymbol{\gamma}^\ell \rangle. \tag{38}$$

Observation 5.2. *Let us return to a remark that preceded Algorithm 1. In particular, we noted that Algorithm 1 can also be viewed as a sequential linearization approach to solving (36). Namely, this corresponds to sequentially performing a linearization of f_2 (and leaving f_1 as is), and then solving the new convex linearized problem.*

Further, let us note why we refer to Algorithm 1 as an alternating minimization scheme. In particular, in light of the reformulation (43) of (36), we can rewrite (36) exactly as

$$(36) = \begin{array}{ll} \underset{\boldsymbol{\beta}, \boldsymbol{\gamma}}{\min} & f_1(\boldsymbol{\beta}) - \langle \boldsymbol{\gamma}, \boldsymbol{\beta} \rangle \\ \text{s. t.} & \sum_i |\gamma_i| \leq \lambda k \\ & |\gamma_i| \leq \lambda \ \forall i. \end{array}$$

[12] To be entirely correct, this result holds for Algorithm 1 with a minor technical modification—see details in Appendix C.

In this sense, if one takes care in performing alternating minimization in $\boldsymbol{\beta}$ (with $\boldsymbol{\gamma}$ fixed) and in $\boldsymbol{\gamma}$ (with $\boldsymbol{\beta}$ fixed) (as in Algorithm 1), then a locally optimal solution is guaranteed.

We now turn to how to actually apply Algorithm 1. Observe that the algorithm is quite simple; in particular, it only requires solving two types of well-structured convex optimization problems. The first such problem, for a fixed $\boldsymbol{\beta}$, is shown in (37). This can be solved in closed form by simply sorting the entries of $|\boldsymbol{\beta}|$, i.e., by finding $|\beta_{(1)}|, \ldots, |\beta_{(p)}|$. The second subproblem, shown in (38) for a fixed $\boldsymbol{\gamma}$, is precisely the usual Lasso problem and is amenable to any of the possible algorithms for the Lasso [31, 42, 70].

Augmented Lagrangian approach

We briefly mention another technique for finding potential solutions to (36) using an Alternating Directions Method of Multiplers (ADMM) [20] approach. To our knowledge, the application of ADMM to the trimmed Lasso problem is novel, although it appears closely related to [68]. We begin by observing that (36) can be written exactly as

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\boldsymbol{\gamma}} & \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\gamma}) \\ \text{s. t.} & \boldsymbol{\beta} = \boldsymbol{\gamma}, \end{array}$$

which makes use of the canonical variable splitting. Introducing dual variable $\mathbf{q} \in \mathbb{R}^p$ and parameter $\sigma > 0$, this becomes in augmented Lagrangian form

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \max_{\mathbf{q}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\gamma}) + \\ \langle \mathbf{q}, \boldsymbol{\beta}-\boldsymbol{\gamma}\rangle + \frac{\sigma}{2}\|\boldsymbol{\beta}-\boldsymbol{\gamma}\|_2^2 . \end{aligned} \tag{39}$$

The utility of such a reformulation is that it is directly amenable to ADMM, as detailed in Algorithm 2. While the problem is nonconvex and therefore the ADMM is not guaranteed to converge, numerical experiments suggest that this approach has superior performance to the DCO-inspired method considered in Algorithm 1.

We close by commenting on the subproblems that must be solved in Algorithm 2. Step 2 can be carried out using "hot" starts. Step 3 is the solution of the trimmed Lasso in the orthogonal design case and can be solved by performed by sorting p numbers; see Appendix C.

Convexification approach

We briefly consider the convex relaxation of the problem (36). We begin by computing the convex envelope [21, 60] of T_k on $[-1, 1]^p$ (here the choice of $[-1, 1]^p$ is standard, such as in the convexification of ℓ_0 over this set which leads to ℓ_1). The proof follows standard techniques (e.g. computing the biconjugate [60]) and is omitted.

Lemma 5.3. *The convex envelope of T_k on $[-1, 1]^p$ is the function $\overline{T_k}$ defined as*

$$\overline{T_k}(\boldsymbol{\beta}) = (\|\boldsymbol{\beta}\|_1 - k)_+ .$$

In words, the convex envelope of T_k is a "soft thresholded" version of the Lasso penalty (thresholded at level k). This can be thought of as an alternative way of interpreting the name "trimmed Lasso."

Algorithm 2 ADMM algorithm for (39)

```
1. Initialize with any $\boldsymbol{\beta}^0, \boldsymbol{\gamma}^0, \mathbf{q}^0 \in \mathbb{R}^p$ and $\sigma > 0$. Repeat, for $\ell \geq 0$, Steps 2, 3, and 4 until a desired numerical convergence tolerance is satisfied.
2. Set
   $$\boldsymbol{\beta}^{\ell+1} \in \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \langle \mathbf{q}^\ell, \boldsymbol{\beta}\rangle + \frac{\sigma}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}^\ell\|_2^2.$$
3. Set
   $$\boldsymbol{\gamma}^{\ell+1} \in \operatorname*{argmin}_{\boldsymbol{\gamma}} \lambda T_k(\boldsymbol{\gamma}) + \frac{\sigma}{2}\|\boldsymbol{\beta}^{\ell+1} - \boldsymbol{\gamma}\|_2^2 - \langle \mathbf{q}^\ell, \boldsymbol{\gamma}\rangle.$$
4. Set $\mathbf{q}^{\ell+1} = \mathbf{q}^\ell + \sigma\left(\boldsymbol{\beta}^{\ell+1} - \boldsymbol{\gamma}^{\ell+1}\right)$.
```

As a result of Lemma 5.3, it follows that the convex analogue of (36), as taken over $[-1,1]^p$, is precisely

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \lambda\left(\|\boldsymbol{\beta}\|_1 - k\right)_+. \tag{40}$$

Problem (40) is amenable to a variety of convex optimization techniques such as subgradient descent [21].

5.2 Certificates of optimality for (36)

We close our discussion of the algorithmic implications of the trimmed Lasso by discussing techniques for finding certifiably optimal solutions to (36). All approaches presented in the preceding section find potential candidates for solutions to (36), but none is necessarily globally optimal. Let us return to a representation of (36) that makes use Lemma 2.1:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\mathbf{z}} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \lambda\langle \mathbf{z}, |\boldsymbol{\beta}|\rangle \\ \text{s.t.} & \sum\limits_i z_i = p - k \\ & \mathbf{z} \in \{0,1\}^p. \end{array}$$

As noted in [39], this representation of (36) is amenable to mixed integer optimization ("MIO") methods [19] for finding globally optimal solutions to (36), in the same spirit as other MIO-based approaches to statistical problems [14, 16].

One approach, as described in [69], uses the notion of "big M." In particular, for $M > 0$ sufficiently large, problem (36) can be written exactly as the following linear MIO problem:

$$\begin{aligned} \min_{\boldsymbol{\beta},\mathbf{z},\mathbf{a}} \quad & \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \lambda\sum_i a_i \\ \text{s. t.} \quad & \sum_i z_i = p-k \\ & \mathbf{z}\in\{0,1\}^p \\ & \mathbf{a} \geq \boldsymbol{\beta} + M\mathbf{z} - M\mathbf{1} \\ & \mathbf{a} \geq -\boldsymbol{\beta} + M\mathbf{z} - M\mathbf{1} \\ & \mathbf{a} \geq \mathbf{0}. \end{aligned} \tag{41}$$

This representation as a linear MIO problem enables the direct application of numerous existing MIO algorithms (such as [40]).[13] Also, let us note that the linear relaxation of (41), i.e., problem (41) with the constraint $\mathbf{z}\in\{0,1\}^p$ replaced with $\mathbf{z}\in[0,1]^p$, is the problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 + \lambda\left(\|\boldsymbol{\beta}\|_1 - Mk\right)_+,$$

where we see the convex envelope penalty appear directly. As such, when M is large, the linear relaxation of (41) is the ordinary Lasso problem $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1$.

5.3 Computational example

Because a rigorous computational comparison is not the primary focus of this paper, we provide a limited demonstration that describes the behavior of solutions to (36) as computed via the different approaches. Precise computational details are contained in Appendix C.4. We will focus on two different aspects: sparsity and approximation quality.

Sparsity properties

As the motivation for the trimmed Lasso is ostensibly sparse modeling, its sparsity properties are of particular interest. We consider a problem instance with $p = 20$, $n = 100$, $k = 2$, and signal-to-noise ratio 10 (the sparsity of the ground truth model $\boldsymbol{\beta}_{\text{true}}$ is 10). The relevant coefficient profiles as a function of λ are shown in Figure 3. In this example none of the convex approaches finds the optimal two variable solution computed using mixed integer optimization. Further, as one would expect *a priori*, the optimal coefficient profiles (as well as the ADMM profiles) are not continuous in λ. Finally, note that by design of the algorithms, the alternating minimization and ADMM approaches yield solutions with sparsity at most k for λ sufficiently large.

Optimality gap

Another critical question is the degree of suboptimality of solutions found via the convex approaches. We average optimality gaps across 100 problem instances with $p = 20$, $n = 100$, and $k = 2$; the relevant results are shown in Figure 4. The results are entirely as one might expect. When λ is small and the problem is convex or nearly convex, the heuristics perform well. However, this breaks down as λ increases and the sparsity-inducing nature of the trimmed Lasso penalty comes into play. Further, we see that the convex envelope approach tends to perform the worst, with the ADMM

[13] There are certainly other possible representations of (43), such as using special ordered set (SOS) constraints, see e.g. [14]. Without more sophisticated tuning of M as in [14], the SOS formulations appear to be vastly superior in terms of time required to prove optimality. The precise formulation essentially takes the form of problem (10). An SOS-based implementation is provided in the supplementary code as the default method of certifying optimality.

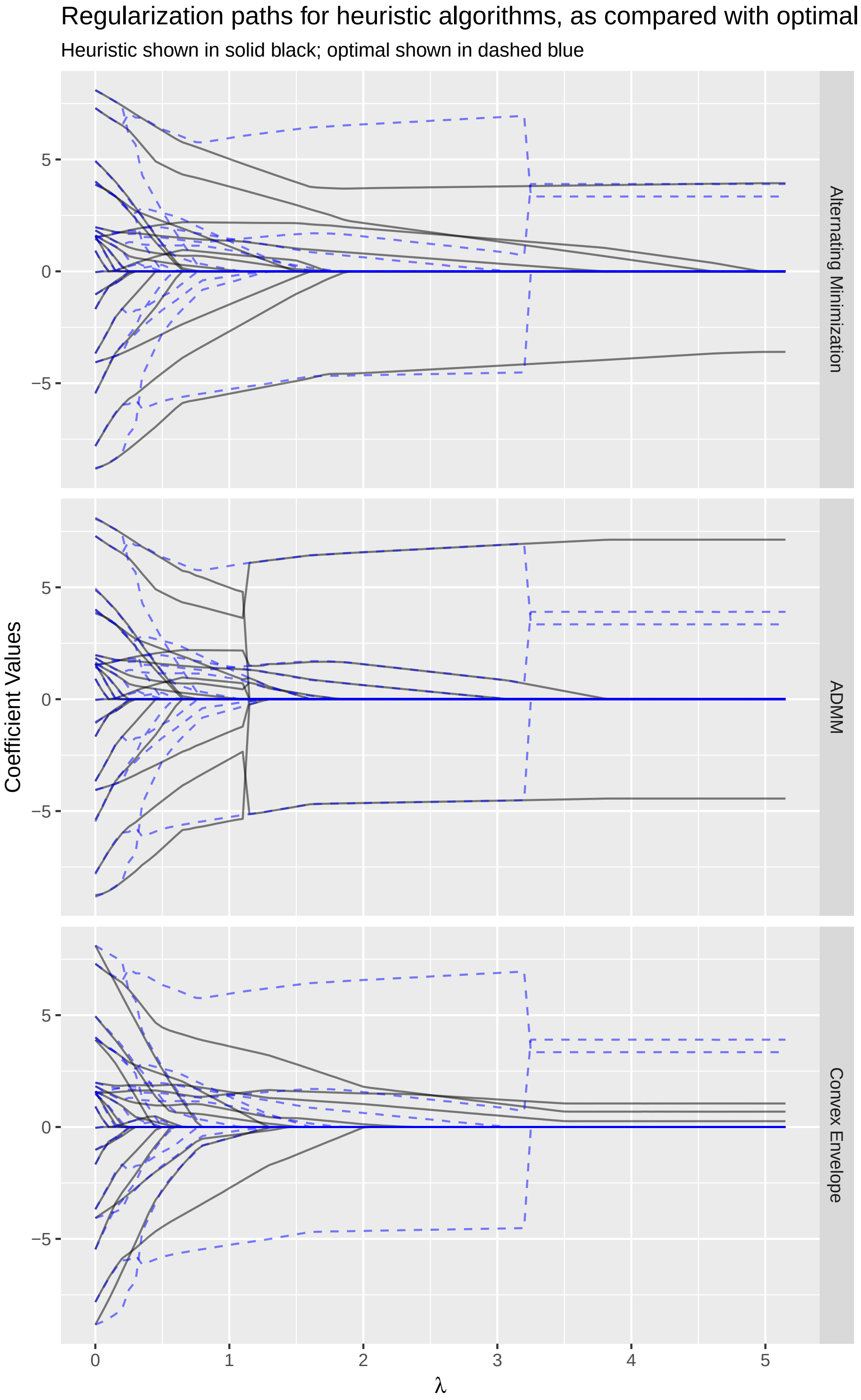

Regularization paths for heuristic algorithms, as compared with optimal
Heuristic shown in solid black; optimal shown in dashed blue
Coefficient Values
5
0
−5
Alternating Minimization
ADMM
Convex Envelope
0
1
2
3
4
5
λ

Figure 3

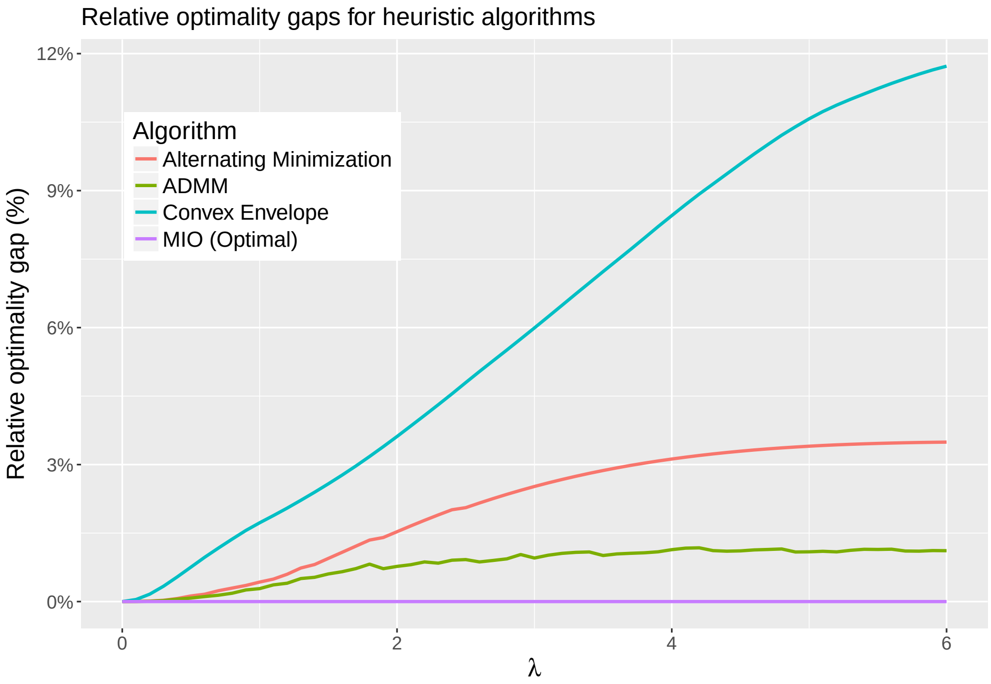


Figure 4

performing the best of the three heuristics. This is perhaps not surprising, as any solution found via the ADMM can be guaranteed to be locally optimal by subsequently applying the alternating minimization scheme of Algorithm 1 to any solution found via Algorithm 2.

Computational burden

Loosely speaking, the heuristic approaches all carry a similar computational cost per iteration, namely, solving a Lasso-like problem. In contrast, the MIO approach can take significantly more computational resources. However, by design, the MIO approach maintains a suboptimality gap throughout computation and can therefore be terminated, before optimality is certified, with a certificate of suboptimality. We do not consider any empirical analysis of runtime here.

Other considerations

There are other additional computational considerations that are potentially of interest as well, but they are primarily beyond the scope of the present work. For example, instead of considering optimality purely in terms of objective values in (36), there are other critical notions from a statistical perspective (e.g. ability to recover true sparse models and performance on out-of-sample data) that would also be necessary to consider across the multiple approaches.

6 Conclusions

In this work, we have studied the trimmed Lasso, a nonconvex adaptation of Lasso that acts as an exact penalty method for best subset selection. Unlike some other approaches to exact penalization which use coordinate-wise separable functions, the trimmed Lasso offers direct control of the desired sparsity k. Further, we emphasized the interpretation of the trimmed Lasso from the perspective of robustness. In doing so, we provided contrasts with the SLOPE penalty as well as comparisons with estimators from the robust statistics and total least squares literature.

We have also taken care to contextualize the trimmed Lasso within the literature on nonconvex penalized estimation approaches to sparse modeling, showing that penalties like the trimmed Lasso can be viewed as a generalization of such approaches in the case when the penalty function is bounded. In doing so, we also highlighted how precisely the problems were related, with a complete characterization given in the case of the clipped Lasso.

Finally, we have shown how modern developments in optimization can be brought to bear for the trimmed Lasso to create convex optimization optimization algorithms that can take advantage of the significant developments in algorithms for Lasso-like problems in recent years.

Our work here raises many interesting questions about further properties of the trimmed Lasso and the application of similar ideas in other settings. We see two particularly noteworthy directions of focus: algorithms and statistical properties. For the former, we anticipate that an approach like trimmed Lasso, which leads to relatively straightforward algorithms that use close analogues from convex optimization, is simple to interpret and to implement. At the same time, the heuristic approaches to the trimmed Lasso presented herein carry no more of a computational burden than solving convex, Lasso-like problems. On the latter front, we anticipate that a deeper analysis of the statistical properties of estimators attained using the trimmed Lasso would help to illuminate it in its own right while also further connecting it to existing approaches in the statistical estimation literature.

Appendix A General min-max representation of SLOPE

For completeness, in this appendix we include the more general representation of the SLOPE penalty $R_{\text{SLOPE}(\mathbf{w})}$ in the same spirit of Proposition 1.2. Here we work with SLOPE in its most general form, namely,

$$R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\beta}) = \sum_{i=1}^{p} w_i |\beta_{(i)}|,$$

where $\mathbf{w}$ is a (fixed) vector of weights with $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and $w_1 > 0$.

To describe the general min-max representation, we first set some notation. For a matrix $\boldsymbol{\Delta} \in \mathbb{R}^{n \times p}$, we let $\boldsymbol{\nu}(\boldsymbol{\Delta}) \in \mathbb{R}^p$ be the vector $(\|\boldsymbol{\Delta}_1\|_2, \ldots, \|\boldsymbol{\Delta}_p\|_2)$ with entries sorted so that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_p$. As usual, for two vectors $\mathbf{x}$ and $\mathbf{y}$, we use $\mathbf{x} \leq \mathbf{y}$ to denote that coordinate-wise inequality holds. With this notation, we have the following:

Proposition A.1. *Problem (9) with uncertainty set*

$$\mathcal{U}_{\mathbf{w}} = \{\boldsymbol{\Delta} : \boldsymbol{\nu}(\boldsymbol{\Delta}) \leq \mathbf{w}\}$$

is equivalent to problem (3) with $R(\boldsymbol{\beta}) = R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\beta})$. Further, problem (9) with uncertainty set

$$\mathcal{U}_{\mathbf{w}} = \left\{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\phi}\|_2 \leq R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\phi})\ \forall \boldsymbol{\phi}\right\}$$

is equivalent to problem (3) with $R(\boldsymbol{\beta}) = R_{\text{SLOPE}(\mathbf{w})}(\boldsymbol{\beta})$.

The proof, like the proof of Proposition 1.2, follows basic techniques described in [9] and is therefore omitted.

Appendix B Additional proofs

This appendix section contains supplemental proofs not contained in the main text.

Proof of Theorem 2.3. Let $\overline{\lambda} = \|\mathbf{y}\|_2 \cdot (\max_j \|\mathbf{x}_j\|_2)$, where $\mathbf{x}_j$ denotes the jth row of $\mathbf{X}$. We fix $\lambda > \overline{\lambda}$, k, and $\eta > 0$ throughout the entire proof. We begin by observing that it suffices to show that any solution $\boldsymbol{\beta}$ to

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda T_k\left(\boldsymbol{\beta}\right) + \eta\|\boldsymbol{\beta}\|_1 \tag{42}$$

satisfies $T_k\left(\boldsymbol{\beta}\right) = 0$, or equivalently, $\|\boldsymbol{\beta}\|_0 \leq k$. As per Lemma 2.1, problem (42) can be rewritten exactly as

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\mathbf{z}} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\langle \mathbf{z}, |\boldsymbol{\beta}|\rangle + \eta\|\boldsymbol{\beta}\|_1 \\ \text{s.t.} & \sum\limits_i z_i = p - k \\ & \mathbf{z} \in \{0,1\}^p. \end{array} \tag{43}$$

Let $(\boldsymbol{\beta}^*, \mathbf{z}^*)$ be any solution to (43). Observe that necessarily $\boldsymbol{\beta}^*$ is also a solution to the problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\langle \mathbf{z}^*, |\boldsymbol{\beta}|\rangle + \eta\|\boldsymbol{\beta}\|_1. \tag{44}$$

Note that, unlike (42), the problem in (44) is readily amenable to an analysis using the theory of proximal gradient methods [7, 28]. In particular, we must have for any $\gamma > 0$ that

$$\boldsymbol{\beta}^* = \operatorname{prox}_{\gamma R}\left(\boldsymbol{\beta}^* - \gamma(\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}'\mathbf{y})\right), \tag{45}$$

where $R(\boldsymbol{\beta}) = \eta\|\boldsymbol{\beta}\|_1 + \lambda \sum\limits_{i\,:\,z_i^*=1} |\beta_i|$. Suppose that $T_k\left(\boldsymbol{\beta}^*\right) > 0$. In particular, for some $j \in \{1, \ldots, p\}$, we have $\beta_j^* \neq 0$ and $z_j^* = 1$. Yet, as per (45),[14]

$$\left|\beta_j^* - \gamma\langle \mathbf{x}_j, \mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}\rangle\right| > \gamma(\eta + \lambda) \qquad \text{for all } \gamma > 0,$$

where $\mathbf{x}_j$ denotes the jth row of $\mathbf{X}$. This implies that

$$|\langle \mathbf{x}_j, \mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}\rangle| \geq \eta + \lambda.$$

Now, using the definition of $\overline{\lambda}$, observe that

$$\begin{aligned} \eta + \lambda \leq |\langle \mathbf{x}_j, \mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}\rangle| &\leq \|\mathbf{x}_j\|_2\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}\|_2 \\ &\leq \|\mathbf{x}_j\|_2\|\mathbf{y}\| \leq \overline{\lambda} < \lambda, \end{aligned}$$

which is a contradiction since $\eta > 0$. Hence, $T_k\left(\boldsymbol{\beta}^*\right) = 0$, completing the proof. □

[14] This is valid for the following reason: since $\beta_j^* \neq 0$ and β_j^* satisfies (45), it must be the case that $\left|\beta_j^* - \gamma\mathbf{x}_j'(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y})\right| > \gamma(\eta+\lambda)$, for otherwise the soft-thresholding operator at level $\gamma(\eta+\lambda)$ would set this quantity to zero.

Extended statement of Proposition 2.5

We now include a precise version of the convergence claim in Proposition 2.5. Let us set a standard notion: we say that $\boldsymbol{\beta}$ is ϵ-optimal (for $\epsilon > 0$) to an optimization problem (P) if the optimal objective value of (P) is within ϵ of the objective value of $\boldsymbol{\beta}$. We add an additional regularizer $\eta\|\boldsymbol{\beta}\|_1$, for $\eta > 0$ fixed, to the objective in order to ensure coercivity of the objective functions.

Proposition B.1 (Extended form of Proposition 2.5)**.** *Let $g : \mathbb{R}_+ \to \mathbb{R}_+$ be an unbounded, continuous, and strictly increasing function with $g(0) = 0$. Consider the problems*

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\pi_k^g(\boldsymbol{\beta}) + \eta\|\boldsymbol{\beta}\|_1 \tag{46}$$

and

$$\min_{\|\boldsymbol{\beta}\|_0 \le k} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1. \tag{47}$$

For every $\epsilon > 0$, there exists some $\underline{\lambda} = \underline{\lambda}(\epsilon) > 0$ so that for all $\lambda > \underline{\lambda}$,

1. *For every optimal $\boldsymbol{\beta}^*$ to (46), there is some $\widehat{\boldsymbol{\beta}}$ so that $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \le \epsilon$, $\widehat{\boldsymbol{\beta}}$ is feasible to (47), and $\widehat{\boldsymbol{\beta}}$ is ϵ-optimal to (47).*
2. *Every optimal $\boldsymbol{\beta}^*$ to (47) is ϵ-optimal to (46).*

Proof. The proof follows a basic continuity argument that is simpler than the one presented below in Theorem B.2. For that reason, we do not include a full proof. Observe that the assumptions on g imply that g^{-1} is well-defined on, say, $g([0,1])$. If we let $\epsilon > 0$ and suppose that $\boldsymbol{\beta}^*$ is optimal to (46), where $\lambda > \underline{\lambda} := \|\mathbf{y}\|_2^2/(2g(\epsilon/p))$, and if we define $\widehat{\boldsymbol{\beta}}$ to be $\boldsymbol{\beta}^*$ with all but the k largest magnitude entries truncated to zero (ties broken arbitrarily), then $\pi_k^g(\boldsymbol{\beta}^*) \le \|\mathbf{y}\|_2^2/(2\lambda)$ and $\pi_k^g(\boldsymbol{\beta}^*) = \sum_{i=1}^p g(|\beta_i^* - \widehat{\beta}_i|)$ so that $|\beta_i^* - \widehat{\beta}_i| \le g^{-1}(\|\mathbf{y}\|_2^2/(2\lambda)) \le \epsilon/p$ by definition of $\underline{\lambda}$. Hence, $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 \le \epsilon$, and all the other claims essentially follow from this. □

Proof of Theorem 3.1. We begin by showing that for any $\boldsymbol{\beta}$,

$$\min_{\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 - \lambda\sum_{i=1}^k |\beta_{(i)}|\right)_+$$

where $(a)_+ := \max\{0, a\}$. Fix $\boldsymbol{\beta}$ and set $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. We assume without loss of generality that $\mathbf{r} \ne \mathbf{0}$ and that $\boldsymbol{\beta} \ne \mathbf{0}$. For any $\boldsymbol{\Delta}$, note that $\|\mathbf{r} - \boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \ge 0$ and $\|\mathbf{r} - \boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \ge \|\mathbf{r}\|_2 - \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2$ by the reverse triangle inequality. Now observe that for $\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda$,

$$\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \le \sum_i |\beta_i| \|\boldsymbol{\Delta}_i\|_2 \le \sum_{i=1}^k \lambda|\beta_{(i)}|.$$

Therefore, $\|\mathbf{r} - \boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \ge \left(\|\mathbf{r}\|_2 - \lambda\sum_{i=1}^k |\beta_{(i)}|\right)_+$. Let $I \subseteq \{1, \ldots, p\}$ be a set of k indices which correspond to the k largest entries of $\boldsymbol{\beta}$ (if $|\beta_{(k)}| = |\beta_{(k+1)}|$, break ties arbitrarily). Define $\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda$ as the matrix whose ith column is

$$\begin{cases} \underline{\lambda}\,\mathrm{sgn}(\beta_i)\mathbf{r}/\|\mathbf{r}\|_2, & i \in I \\ 0, & i \notin I, \end{cases}$$

where $\underline{\lambda} = \min\left\{\lambda, \|\mathbf{r}\|_2/\left(\sum_{i=1}^k |\beta_{(i)}|\right)\right\}$. It is easy to verify that $\boldsymbol{\Delta} \in \mathcal{U}_k^\lambda$ and $\|\mathbf{r} - \boldsymbol{\Delta}\boldsymbol{\beta}\|_2 = \left(\|\mathbf{r}\|_2 - \lambda\sum_{i=1}^k |\beta_{(i)}|\right)_+$. Combined with the lower bound, we have

$$\min_{\boldsymbol{\Delta}\in\mathcal{U}_k^\lambda} \|\mathbf{y} - (\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 - \lambda\sum_{i=1}^k |\beta_{(i)}|\right)_+$$

which completes the first claim.

It follows that the problem (20) can be rewritten exactly as

$$\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 - \lambda\sum_{i=1}^k |\beta_{(i)}|\right)_+ + r(\boldsymbol{\beta}). \tag{48}$$

To finish the proof of the theorem, it suffices to show that if $\boldsymbol{\beta}^*$ is a solution to (48), then

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 - \lambda\sum_{i=1}^k |\beta^*_{(i)}| \geq 0.$$

If this is not true, then $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 - \lambda\sum_{i=1}^k |\beta^*_{(i)}| < 0$ and so $\boldsymbol{\beta}^* \neq \mathbf{0}$. However, this implies that for $1 > \epsilon > 0$ sufficiently small, $\boldsymbol{\beta}_\epsilon := (1-\epsilon)\boldsymbol{\beta}^*$ satisfies $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\epsilon\|_2 - \lambda\sum_{i=1}^k |(\beta_\epsilon)_{(i)}| < 0$. This in turn implies that

$$\begin{aligned} &\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\epsilon\|_2 - \lambda\textstyle\sum_{i=1}^k |(\beta_\epsilon)_{(i)}|\right)_+ + r(\boldsymbol{\beta}_\epsilon) \\ &< \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 - \lambda\textstyle\sum_{i=1}^k |\beta^*_{(i)}|\right)_+ + r(\boldsymbol{\beta}^*), \end{aligned}$$

which contradicts the optimality of $\boldsymbol{\beta}^*$. (We have used the absolute homogeneity of the norm r and that $\boldsymbol{\beta}^* \neq \mathbf{0}$.) Hence, any optimal $\boldsymbol{\beta}^*$ to (48) necessarily satisfies $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 - \lambda\sum_{i=1}^k |\beta^*_{(i)}| \geq 0$ and so the desired results follows. □

N.B. The assumption that r is a norm can be relaxed somewhat (as is clear in the proof), although the full generality is not necessary for our purposes.

Corollary 3.2 and related discussions

Here we include a precise statement of the "approximate" claim in Corollary 3.2. After the proof, we include a discussion of related technical issues.

Theorem B.2 (Precise statement of Corollary 3.2)**.** *For $\tau > \lambda > 0$, consider the problems*

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + (\tau-\lambda)\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\beta}) \\ \text{s. t.} \quad & \lambda\sum_{i=1}^k |\beta_{(i)}| \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2. \end{aligned} \tag{49}$$

and

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + (\tau-\lambda)\|\boldsymbol{\beta}\|_1 + \lambda T_k(\boldsymbol{\beta}). \tag{50}$$

For all $\epsilon > 0$, there exists $\overline{\lambda} = \overline{\lambda}(\epsilon) > 0$ so that whenever $\lambda \in (0, \overline{\lambda})$,

1. *Every optimal $\boldsymbol{\beta}^*$ to (49) is ϵ-optimal to (50).*

2. *For every optimal $\boldsymbol{\beta}^*$ to (50), there is some $\widehat{\boldsymbol{\beta}}$ so that $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \leq \epsilon$, $\widehat{\boldsymbol{\beta}}$ is feasible to (49), and $\widehat{\boldsymbol{\beta}}$ is ϵ-optimal to (49).*

Proof. Fix $\tau > 0$ throughout. We assume without loss of generality that $\mathbf{y} \neq \mathbf{0}$, as otherwise the claim is obvious. We will prove the second claim first, as it essentially implies the first.

Let us consider two situations. In particular, we consider whether there exists a nonzero optimal solution to

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \tau\|\boldsymbol{\beta}\|_1. \tag{51}$$

Case 1—existence of nonzero optimal solution to (51)

We first consider the case when there exists a nonzero solution to problem (51). We show a few lemmata:

1. We first show that the norm of solutions to (50) are uniformly bounded away from zero, independent of λ. To proceed, let $\widehat{\boldsymbol{\beta}}$ be any nonzero optimal solution to (51). Observe that if $\boldsymbol{\beta}^*$ is optimal to (50), then

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)\|\boldsymbol{\beta}^*\|_1 + \lambda T_k(\boldsymbol{\beta}^*) &\leq \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2 + (\tau - \lambda)\|\widehat{\boldsymbol{\beta}}\|_1 + \lambda T_k(\widehat{\boldsymbol{\beta}}) \\ &\leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + \tau\|\boldsymbol{\beta}^*\|_1 - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 + \lambda T_k(\widehat{\boldsymbol{\beta}}),\end{aligned}$$

   implying that $\|\widehat{\boldsymbol{\beta}}\|_1 - T_k(\widehat{\boldsymbol{\beta}}) \leq \|\boldsymbol{\beta}^*\|_1 - T_k(\boldsymbol{\beta}^*)$. In other words, $\sum_{i=1}^k |\widehat{\beta}_{(i)}| \leq \sum_{i=1}^k |\beta^*_{(i)}| \leq \|\boldsymbol{\beta}^*\|_1$. Using the fact that $\widehat{\boldsymbol{\beta}} \neq \mathbf{0}$, we have that any solution $\boldsymbol{\beta}^*$ to (50) has strictly positive norm:

$$\|\boldsymbol{\beta}^*\|_1 \geq C > 0,$$

   where $C := \sum_{i=1}^k |\widehat{\beta}_{(i)}|$ is a universal constant depending only on τ (and not λ).

2. We now upper bound the norm of solutions to (50). In particular, if $\boldsymbol{\beta}^*$ is optimal to (50), then

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)\|\boldsymbol{\beta}^*\|_1 + \lambda T_k(\boldsymbol{\beta}^*) \leq \|\mathbf{y}\|_2 + 0 + 0 = \|\mathbf{y}\|_2,$$

   and so $\|\boldsymbol{\beta}^*\|_1 \leq \|\mathbf{y}\|_2/(\tau - \lambda)$. (This bound is not uniform in λ, but if we restrict our attention to, say $\lambda \leq \tau/2$, it is.)

3. We now lower bound the loss for scaled version of optimal solutions. In particular, if $\sigma \in [0, 1]$ and $\boldsymbol{\beta}^*$ is optimal to (50), then by optimality we have that

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)\|\boldsymbol{\beta}^*\|_1 + \lambda T_k(\boldsymbol{\beta}^*) \leq \|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)\sigma\|\boldsymbol{\beta}^*\|_1 + \lambda\sigma T_k(\boldsymbol{\beta}^*),$$

   which in turn implies that

$$\begin{aligned}\|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 &\geq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)(1 - \sigma)\|\boldsymbol{\beta}^*\|_1 + \lambda(1 - \sigma)T_k(\boldsymbol{\beta}^*) \\ &\geq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)(1 - \sigma)C \geq (\tau - \lambda)(1 - \sigma)C\end{aligned}$$

   by combining with the first observation.

Using these, we are now ready to proceed. Let $\epsilon > 0$; we assume without loss of generality that $\epsilon < 2\|\mathbf{y}\|_2/\tau$. Let

$$\bar{\lambda} := \min\left\{\frac{\epsilon\tau^3 C}{4\|\mathbf{y}\|_2(2\|\mathbf{y}\|_2 - \epsilon\tau)}, \frac{\tau}{2}\right\}.$$

Fix $\lambda \in (0, \overline{\lambda})$ and let $\boldsymbol{\beta}^*$ be any optimal solution to (50). Define

$$\sigma := \left(1 - \frac{\epsilon\tau}{2\|\mathbf{y}\|_2}\right) \quad \text{and} \quad \widehat{\boldsymbol{\beta}} := \sigma\boldsymbol{\beta}^*.$$

We claim that $\widehat{\boldsymbol{\beta}}$ satisfies the desired requirements of the theorem:

1. We first argue that $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \leq \epsilon$. Observe that

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 = \epsilon\tau\|\boldsymbol{\beta}^*\|_2/(2\|\mathbf{y}\|_2) \leq \epsilon\tau\|\boldsymbol{\beta}^*\|_1/(2\|\mathbf{y}\|_2) \leq \epsilon\tau\|\mathbf{y}\|_2/(2\|\mathbf{y}\|_2(\tau - \lambda)) \leq \epsilon.$$

2. We now show that $\widehat{\boldsymbol{\beta}}$ is feasible to (49). This requires us to argue that $\lambda \sum_{i=1}^k |\widehat{\beta}_{(i)}| \leq \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2$. Yet,

$$\begin{aligned} \lambda \sum_{i=1}^k |\widehat{\beta}_{(i)}| \leq \lambda\|\widehat{\boldsymbol{\beta}}\|_1 = \lambda\sigma\|\boldsymbol{\beta}^*\|_1 \leq 2\lambda\sigma\|\mathbf{y}\|_2/\tau \leq \frac{\tau}{2}(1-\sigma)C \\ \leq (\tau - \lambda)(1-\sigma)C \leq \|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2, \end{aligned}$$

as desired. The only non-obvious step is the inequality $2\lambda\sigma\|\mathbf{y}\|_2/\tau \leq \tau(1-\sigma)C/2$, which follows from algebraic manipulations using the definitions of σ and $\overline{\lambda}$.

3. Finally, we show that $\widehat{\boldsymbol{\beta}}$ is $(\epsilon\|\mathbf{X}\|_2)$-optimal to (49). Indeed, because $\boldsymbol{\beta}^*$ is optimal to (50) which necessarily lowers bound problem (49), we have that the objective value gap between $\widehat{\boldsymbol{\beta}}$ and an optimal solution to (49) is at most

$$\begin{aligned} &\|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + (\tau - \lambda)(\sigma - 1)\|\boldsymbol{\beta}^*\|_1 + \lambda(\sigma - 1)T_k(\boldsymbol{\beta}^*) \\ &\leq (1-\sigma)\|\mathbf{X}\boldsymbol{\beta}^*\|_2 + 0 + 0 \leq (1-\sigma)\|\mathbf{X}\|_2\|\boldsymbol{\beta}^*\|_2 \leq 2(1-\sigma)\|\mathbf{X}\|_2\|\mathbf{y}\|_2/\tau \\ &= 2\epsilon\tau/(2\|\mathbf{y}\|_2)\|\mathbf{X}\|_2\|\mathbf{y}\|_2/\tau = \epsilon\|\mathbf{X}\|_2. \end{aligned}$$

As the choice of $\epsilon > 0$ was arbitrary, this completes the proof of claim 2 in the theorem in the case when $\mathbf{0}$ is not a solution to (51).

Case 2—no nonzero optimal solution to (51)

In the case when there is no nonzero optimal solution to (51), $\mathbf{0}$ is optimal and it is the only optimal point. Our analysis will be similar to the previous approach, with the key difference being in how we lower bound the quantity $\|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2$ where $\boldsymbol{\beta}^*$ is optimal to (50). Again, we have several lemmata:

1. As before, if $\boldsymbol{\beta}^*$ is optimal to (50), then $\|\boldsymbol{\beta}^*\|_1 \leq \|\mathbf{y}\|_2/(\tau - \lambda)$.

2. We now lower bound the quantity $\|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2$, where $\boldsymbol{\beta}^*$ is optimal to (50) and $\sigma \in [0, 1]$. As such, consider the function

$$f(\sigma) := \|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 + \sigma\tau\|\boldsymbol{\beta}^*\|_1.$$

Because f is convex in σ and the unique optimal solution to (51) is $\mathbf{0}$, we have that

$$f(\sigma) \geq f(0) + \sigma f'(0) \quad \forall\sigma \in [0,1] \quad \text{and} \quad f'(0) \geq 0$$

(It is not difficult to argue that f is differentiable at 0.) An elementary computation shows that $f'(0) = \tau\|\boldsymbol{\beta}^*\|_1 - \langle \mathbf{y}, \mathbf{X}\boldsymbol{\beta}^*\rangle/\|\mathbf{y}\|_2$. Therefore, we have that

$$\|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 + \sigma\tau\|\boldsymbol{\beta}^*\|_1 \geq \|\mathbf{y}\|_2 + \sigma\left(\tau\|\boldsymbol{\beta}^*\|_1 - \langle \mathbf{y}, \mathbf{X}\boldsymbol{\beta}^*\rangle/\|\mathbf{y}\|_2\right),$$

implying that

$$\|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 \geq \|\mathbf{y}\|_2 - \sigma\langle \mathbf{y}, \mathbf{X}\boldsymbol{\beta}^*\rangle/\|\mathbf{y}\|_2 \geq \|\mathbf{y}\|_2 - \sigma\tau\|\boldsymbol{\beta}^*\|_1 \geq \|\mathbf{y}\|_2 - \sigma\tau\|\mathbf{y}\|_2/(\tau - \lambda),$$

with the final step following by an application of the previous lemma.

We are now ready to proceed. Let $\epsilon > 0$; we assume without loss of generality that $\epsilon < 2\|\mathbf{y}\|_2/\tau$. Let

$$\overline{\lambda} := \min\left\{\frac{\epsilon\tau^2}{4\|\mathbf{y}\|_2 - \epsilon\tau}, \frac{\tau}{2}\right\}.$$

Fix $\lambda \in (0, \overline{\lambda})$ and let $\boldsymbol{\beta}^*$ be any optimal solution to (50). Define

$$\sigma := \left(1 - \frac{\epsilon\tau}{2\|\mathbf{y}\|_2}\right) \quad \text{and} \quad \widehat{\boldsymbol{\beta}} := \sigma\boldsymbol{\beta}^*.$$

We claim that $\widehat{\boldsymbol{\beta}}$ satisfies the desired requirements:

1. The proof of the claim that $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \leq \epsilon$ is exactly as before.
2. We now show that $\widehat{\boldsymbol{\beta}}$ is feasible to (49), which requires a different proof. Again this requires us to argue that $\lambda\sum_{i=1}^k |\widehat{\beta}_{(i)}| \leq \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2$. Yet,

$$\begin{aligned}\lambda\sum_{i=1}^k |\widehat{\beta}_{(i)}| \leq \lambda\|\widehat{\boldsymbol{\beta}}\|_1 = \lambda\sigma\|\boldsymbol{\beta}^*\|_1 \leq \lambda\sigma\|\mathbf{y}\|_2/(\tau - \lambda) &\leq \|\mathbf{y}\|_2 - \sigma\tau\|\mathbf{y}\|_2/(\tau - \lambda)\\ &\leq \|\mathbf{y} - \sigma\mathbf{X}\boldsymbol{\beta}^*\|_2 = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2,\end{aligned}$$

as desired. The only non-obvious step is the inequality $\lambda\sigma\|\mathbf{y}\|_2/(\tau-\lambda) \leq \|\mathbf{y}\|_2 - \sigma\tau\|\mathbf{y}\|_2/(\tau - \lambda)$, which follows from algebraic manipulations using the definitions of σ and $\overline{\lambda}$.
3. Finally, the proof that $\widehat{\boldsymbol{\beta}}$ is $(\epsilon\|\mathbf{X}\|_2)$-optimal to (49) follows in the same way as before.

Therefore, we conclude that in the case when $\mathbf{0}$ is the unique optimal solution to (51), then again we have that the claim 2 of the theorem holds.

Finally, we show that claim 1 holds: any solution $\boldsymbol{\beta}^*$ to (49) is ϵ-optimal to (50). This follows by letting $\overline{\boldsymbol{\beta}}$ be any optimal solution to (50). By applying the entire argument above, we know that the objective value of some $\widehat{\boldsymbol{\beta}}$, feasible to (49) and close to $\overline{\boldsymbol{\beta}}$, is within ϵ of the optimal objective value of (49), i.e., the objective value of $\boldsymbol{\beta}^*$, and within ϵ of the objective value of (50), i.e., the objective value of $\overline{\boldsymbol{\beta}}$. This completes the proof. □

In short, the key complication is that the quantity $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2$ does not need to be uniformly bounded away from zero for solutions $\boldsymbol{\beta}^*$ to problem (50). This is part of the complication of working with the homogeneous form of the trimmed Lasso problem. For a concrete example, if one considers the homogeneous Lasso problem with $p = n = 1$, $\mathbf{y} = (1)$, and $\mathbf{X} = (1)$, then the homogeneous Lasso problem $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \eta\|\boldsymbol{\beta}\|_1$ is

$$\min_{\beta} |1 - \beta| + \eta|\beta|.$$

For $\eta \in [0,1]$, $\beta^* = 1$ is an optimal solution to this problem with corresponding error $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\| = 0$. If we make an assumption about the behavior of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|$, then we do not need the setup as shown above.

Proof of Proposition 3.3. The proof is entirely analogous to that of Theorems 3.1 and B.2 and is therefore omitted. □

Proof of validity of Example 4.4. Let us consider the problem instance where $p = n = 2$ with

$$\mathbf{y} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

Let $\lambda = 1/2$ and $\ell = 1$, and consider the problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + |\beta_{(2)}| = \min_{\beta_1,\beta_2} (1 - \beta_1 + \beta_2)^2 + (1 + \beta_1 - 2\beta_2)^2 + |\beta_{(2)}|. \tag{52}$$

We have omitted the factor of $1/2$ as shown in the actual example in the main text in order to avoid unnecessary complications.

Solving problem (52) and its related counterparts (for $\ell \in \{0,2\}$) can rely on convex analysis because we can simply enumerate all possible scenarios. In particular, the solution to (52) is $\boldsymbol{\beta}^* = (3/2, 1)$ based on an analysis of two related problems:

$$\min_{\beta_1,\beta_2} (1 - \beta_1 + \beta_2)^2 + (1 + \beta_1 - 2\beta_2)^2 + |\beta_1|.$$
$$\min_{\beta_1,\beta_2} (1 - \beta_1 + \beta_2)^2 + (1 + \beta_1 - 2\beta_2)^2 + |\beta_2|.$$

(We should be careful to impose the additional constraints $|\beta_1| \le |\beta_2|$ and $|\beta_1| \ge |\beta_2|$, respectively, although a simple argument shows that these constraints are not required in this example.) A standard convex analysis using the Lasso (e.g. by directly using subdifferentials) shows that the problems have respective solutions $(1/2, 1/2)$ and $(3/2, 1)$, with the latter having the better objective value in (52). As such, $\boldsymbol{\beta}^*$ is indeed optimal. The solution in the cases of $\ell \in \{0,2\}$ follows a similarly standard analysis.

It is perhaps more interesting to study the general case where $\mu, \gamma \ge 0$. In particular, we will show that $\boldsymbol{\beta}^* = (3/2, 1)$ is not an optimal solution to the clipped Lasso problem

$$\min_{\beta_1,\beta_2} (1 - \beta_1 + \beta_2)^2 + (1 + \beta_1 - 2\beta_2)^2 + \mu \min\{\gamma|\beta_1|, 1\} + \mu \min\{\gamma|\beta_2|, 1\} \tag{53}$$

for any choices of μ and γ. While in general such a problem may be difficult to fully analyze, we can again rely on localized analysis using convex analysis. To proceed, let

$$f(\beta_1, \beta_2) = (1 - \beta_1 + \beta_2)^2 + (1 + \beta_1 - 2\beta_2)^2 + \mu \min\{\gamma|\beta_1|, 1\} + \mu \min\{\gamma|\beta_2|, 1\},$$

with the parameters μ and γ implicit. We consider the following exhaustive cases:

1. $\boxed{\gamma > 1}$: In this case, f is convex and differentiable in a neighborhood of $\boldsymbol{\beta}^*$. Its gradient at $\boldsymbol{\beta}^*$ is $\nabla f(\boldsymbol{\beta}^*) = (0, -1)$, and therefore $\boldsymbol{\beta}^*$ is neither locally optimal nor globally optimal to problem (53).

2. $\boxed{\gamma < 2/3}$: In this case, f is again convex and differentiable in a neighborhood of $\boldsymbol{\beta}^*$. Its gradient at $\boldsymbol{\beta}^*$ is $\nabla f(\boldsymbol{\beta}^*) = (\mu\gamma, \mu\gamma - 1)$. Again, this cannot equal $(0,0)$ and therefore $\boldsymbol{\beta}^*$ is neither locally nor globally optimal to problem (53).

3. $\boxed{2/3 < \gamma < 1}$: In this case, f is again convex and differentiable in a neighborhood of $\boldsymbol{\beta}^*$. Its gradient at $\boldsymbol{\beta}^*$ is $\nabla f(\boldsymbol{\beta}^*) = (0, \mu\gamma - 1)$. As a necessary condition for local optimality, we must have that $\mu\gamma = 1$, implying that $\mu > 1$. Further, if $\boldsymbol{\beta}^*$ is optimal to (53), then $f(\boldsymbol{\beta}^*) \leq f(0,0)$. Yet,

$$\begin{aligned} f(\boldsymbol{\beta}^*) &= 1/2 + \mu + \mu\gamma = 3/2 + \mu \\ f(0,0) &= 2, \end{aligned}$$

implying that $\mu \leq 1/2$, in contradiction of $\mu > 1$. Hence, $\boldsymbol{\beta}^*$ cannot be optimal to (53).

4. $\boxed{\gamma = 2/3}$: In this case, we make two comparisons, using the points $\boldsymbol{\beta}^*$, $(0,0)$, and $(3,2)$:

$$\begin{aligned} f(\boldsymbol{\beta}^*) &= 1/2 + \mu + 2\mu/3 = 1/2 + 5\mu/3 \\ f(0,0) &= 2 \\ f(3,2) &= 2\mu. \end{aligned}$$

Assuming optimality of $\boldsymbol{\beta}^*$, we have that $f(\boldsymbol{\beta}^*) \leq f(0,0)$, i.e., $\mu \leq 9/10$; similarly, $f(\boldsymbol{\beta}^*) \leq f(3,2)$, i.e., $\mu \geq 3/2$. Clearly both cannot hold, and so therefore $\boldsymbol{\beta}^*$ cannot be optimal.

5. $\boxed{\gamma = 1}$: Finally, we see that $f(\boldsymbol{\beta}^*) \leq f(3,2)$ would imply that $1/2 + 2\mu \leq 2\mu$, which is impossible; hence, $\boldsymbol{\beta}^*$ is not optimal to (53). (This argument can clearly also be used in the case when $\gamma > 1$, although it is instructive to see the argument given above in that case.)

In any case, we have that $\boldsymbol{\beta}^*$ cannot be a solution to the clipped Lasso problem (53). This completes the proof of validity of Example 4.4. □

Appendix C Supplementary details for Algorithms

This appendix contains further details on algorithms as discussed in Section 5. The presentation here is primarily self-contained. Note that the alternating minimization scheme based on difference-of-convex optimization can be found in [39].

C.1 Alternating minimization scheme

Let us set the following notation:

$$\begin{array}{rcl} f(\boldsymbol{\beta}) & = & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + \lambda T_k(\boldsymbol{\beta}) + \eta\|\boldsymbol{\beta}\|_1, \\ f_1(\boldsymbol{\beta}) & = & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + (\eta + \lambda)\|\boldsymbol{\beta}\|_1, \\ f_2(\boldsymbol{\beta}) & = & \lambda \sum_{i=1}^k |\beta_{(i)}|. \end{array}$$

Definition C.1. *For any function $F : \mathbb{R}^p \to \mathbb{R}$ and $\epsilon \geq 0$, we define the ϵ-subdifferential of F at $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ to be the set $\partial_\epsilon F(\boldsymbol{\beta}_0)$ defined as*

$$\{\boldsymbol{\gamma} \in \mathbb{R}^p \ : \ F(\boldsymbol{\beta}) - F(\boldsymbol{\beta}_0) \geq \langle \boldsymbol{\gamma}, \boldsymbol{\beta} - \boldsymbol{\beta}_0 \rangle - \epsilon \ \forall \ \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

In particular, when $\epsilon = 0$, we refer to $\partial_0 F(\boldsymbol{\beta}_0)$ as the subdifferential of F at $\boldsymbol{\beta}_0$, and we will denote this as $\partial F(\boldsymbol{\beta}_0)$.

Using this definition, we have the following result precisely characterizing local and global optima of (36).

Theorem C.2. *(a) A point $\boldsymbol{\beta}^*$ is a local minimum of f if and only if $\partial f_2(\boldsymbol{\beta}^*) \subseteq \partial f_1(\boldsymbol{\beta}^*)$.*

(b) A point $\boldsymbol{\beta}^*$ is a global minimum of f if and only if $\partial_\epsilon f_2(\boldsymbol{\beta}^*) \subseteq \partial_\epsilon f_1(\boldsymbol{\beta}^*)$ for all $\epsilon \geq 0$.

Proof. This is a direct application of results in [66, Thm. 1]. Part (b) is immediate. The forward implication of part (a) is immediate as well; the converse implication follows by observing that f_2 is a *polyhedral* convex function [2, Thm. 1(ii)] (see definition therein). □

Let us note that ∂f_1 and ∂f_2 are both easily computable, and hence, local optimality can be verified given some candidate $\boldsymbol{\beta}^*$ per Theorem C.2.[15] We now discuss the associated alternating minimization scheme (or equivalently, as a sequential linearization scheme), shown in Algorithm 1 for finding local optima of (36) by making use of Theorem C.2. Through what follows, we make use of the standard notion of a conjugate function, defined as follows:

Definition C.3. *For any function $F : \mathbb{R}^p \to \mathbb{R}$, we define its conjugate function $F^* : \mathbb{R}^p \to \mathbb{R}$ to be the function*

$$F^*(\boldsymbol{\gamma}) = \sup_{\boldsymbol{\beta}} \langle \boldsymbol{\gamma}, \boldsymbol{\beta} - F(\boldsymbol{\beta}) \rangle.$$

We will make the following minor technical assumption: in step 2) of Algorithm 1, we assume without loss of generality that the $\boldsymbol{\gamma}^\ell$ so computed satisfies the additional criteria:

1. it is an extreme point of the relevant feasible region,
2. and that if $\partial f_2(\boldsymbol{\beta}^\ell) \not\subseteq \partial f_1(\boldsymbol{\beta}^\ell)$, then $\boldsymbol{\gamma}^\ell$ is chosen such that $\boldsymbol{\gamma}^\ell \in \partial f_2(\boldsymbol{\beta}^\ell) \setminus \partial f_1(\boldsymbol{\beta}^\ell)$.

Solving (37) with these additional assumptions can nearly be solved in closed form by simply sorting the entries of $|\boldsymbol{\beta}|$, i.e., by finding $|\beta_{(1)}|, \ldots, |\beta_{(p)}|$.We must take some care to ensure that the second without loss of generality condition on $\boldsymbol{\gamma}$ is satisfied. This is straightforward but tedious; the details are shown in Appendix C.2.

Using this modification, the convergence properties of Algorithm 1 can be proven as follows:

Proof of Theorem 5.1. This is an application of [66, Thms. 3-5]. The only modification is in requiring that $\boldsymbol{\gamma}^\ell$ is chosen so that $\boldsymbol{\gamma}^\ell \in \partial f_2(\boldsymbol{\beta}^*) \setminus \partial f_1(\boldsymbol{\beta}^*)$ if $\boldsymbol{\beta}^\ell$ is not a local minimum of f—see [66, §3.3] for a motivation and justification for such a modification. Finally, the correspondence between $\boldsymbol{\gamma}^\ell \in \partial f_2(\boldsymbol{\beta}^\ell)$ and (37), and between $\boldsymbol{\beta}^{\ell+1} \in \partial f_1^*(\boldsymbol{\gamma}^\ell)$ and (38), is clear from an elementary argument applied to subdifferentials of variational formulations of functions. □

C.2 Algorithm 1, Step 2

Here we present the details of solving (37) in Algorithm 1 in a way that ensures that the associated without loss of generality claims hold. In doing so, we also implicitly study how to verify the conditions for local optimality (*c.f.* Theorem C.2). Throughout, we use the sgn function defined as

$$\operatorname{sgn}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0. \end{cases}$$

[15] For the specific functions of interest, verifying local optimality of a candidate $\boldsymbol{\beta}^*$ can be performed in $O(p\min\{n,p\} + p\log p)$ operations; the first component relates to the computation of $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^*$, while the second captures the sorting of the entries of $\boldsymbol{\beta}^*$. See Appendix C.2 for details.

For fixed $\boldsymbol{\beta}$, the problem of interest is

$$\begin{array}{ll}\max\limits_{\boldsymbol{\gamma}} & \langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle \\ \text{s. t.} & \sum\limits_i |\gamma_i| \leq \lambda k \\ & |\gamma_i| \leq \lambda \ \forall i. \end{array}$$

We wish to find a maximizer $\boldsymbol{\gamma}$ for which the following hold:

1. $\boldsymbol{\gamma}$ is an extreme point of the relevant feasible region,
2. and that if $\partial f_2(\boldsymbol{\beta}) \not\subseteq \partial f_1(\boldsymbol{\beta})$, then $\boldsymbol{\gamma}$ is such that $\boldsymbol{\gamma} \in \partial f_2(\boldsymbol{\beta}) \setminus \partial f_1(\boldsymbol{\beta})$.

As the problem on its own can be solved by sorting the entries of $\boldsymbol{\beta}$, the crux of the problem is ensuring that 2) holds.

Given the highly structured nature of f_1 and f_2 in our setup, it is simple, albeit tedious, to ensure that such a condition is satisfied. Let $I = \{i : |\beta_i| = |\beta_{(k)}|\}$. If $|I| = 1$, the optimal solution is unique, and there is nothing to show. Therefore, we will assume that $|I| \geq 2$. We will construct an optimal solution $\boldsymbol{\gamma}$ which satisfies the desired conditions. First observe that we necessarily must have that 1) $\gamma_i = \lambda \operatorname{sgn}(\beta_i)$ if $|\beta_i| > |\beta_{(k)}|$ and 2) $\gamma_i = 0$ if $|\beta_i| < |\beta_{(k)}|$. We now proceed to define the rest of the entries of $\boldsymbol{\gamma}$. We consider two cases:

1. First consider the case when $|\beta_{(k)}| > 0$. We claim that $\partial f_2(\boldsymbol{\beta}) \not\subseteq \partial f_1(\boldsymbol{\beta})$. To do so, we will inspect the ith entries of $\partial f_1(\boldsymbol{\beta})$ for $i \in I$; as such, let $P_i^j = \{\delta_i : \boldsymbol{\delta} \in \partial f_j(\boldsymbol{\beta})\}$ for $j \in \{1, 2\}$ and $i \in I$ (a projection). For each $i \in I$, we have using basic convex analysis that P_i^1 is a singelton: $P_i^1 = \{\langle \mathbf{X}_i, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle + (\eta + \lambda) \operatorname{sgn}(\beta_i)\}$, where $\mathbf{X}_i$ is the ith column of $\mathbf{X}$. In contrast, because $|I| \geq 2$, the set P_i^2 is an interval with strictly positive length for each $i \in I$ (it is either $[-\lambda, 0]$ or $[0, \lambda]$, depending on whether $\beta_i < 0$ or $\beta_i > 0$, respectively). Therefore, $\partial f_2(\boldsymbol{\beta}) \not\subseteq \partial f_1(\boldsymbol{\beta})$, as claimed.

   Fix an arbitrary $j \in I$. Per the above argument, we must have that $\langle \mathbf{X}_j, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle + (\eta + \lambda) \operatorname{sgn}(\beta_j) \neq 0$ or $\langle \mathbf{X}_j, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle + (\eta + \lambda) \operatorname{sgn}(\beta_j) \neq \lambda \operatorname{sgn}(\beta_j)$. In the former case, set $\gamma_i = 0$, while in the latter case we define $\gamma_i = \lambda \operatorname{sgn}(\beta_i)$ (if both are true, either choice suffices). It is clear that it is possible to fill in the remaining entries of γ_i for $i \in I \setminus \{j\}$ in a straightforward manner so that $\boldsymbol{\gamma} \in \partial f_2(\boldsymbol{\beta})$. Further, by construction, $\boldsymbol{\gamma} \notin \partial f_1(\boldsymbol{\beta})$, as desired.

2. Now consider the case when $|\beta_{(k)}| = 0$. Using the preceding argument, we see that P_i^1 is the interval $[\langle \mathbf{X}_i, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle - (\eta + \lambda), \langle \mathbf{X}_i, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle + \eta + \lambda]$ for $i \in I$. In contrast, P_i^2 is the interval $[-\lambda, \lambda]$ for $i \in I$. If for all $i \in I$ one has that $P_i^2 \subseteq P_i^1$, then the choice of γ_i for $i \in I$ is obvious: any optimal extreme point $\boldsymbol{\gamma}$ of the problem will suffice. (Note here that it may or may not be that $\partial f_2(\boldsymbol{\beta}) \subseteq \partial f_1(\boldsymbol{\beta})$. This entirely depends on β_i for $i \notin I$.)

   Therefore, we may assume that there exists some $j \in I$ so that $P_j^2 \not\subseteq P_j^1$. (It follows immediately that $\partial f_2(\boldsymbol{\beta}) \not\subseteq \partial f_1(\boldsymbol{\beta})$.) We must have that $\langle \mathbf{X}_j, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle - (\eta + \lambda) > -\lambda$ or $\langle \mathbf{X}_j, \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \rangle + (\eta + \lambda) < \lambda$. In the former case, set $\gamma_i = -\lambda$, while in the latter case we define $\gamma_i = \lambda$ (if both are true, either choice suffices). It is clear that it is possible to fill in the remaining entries of γ_i for $i \in I \setminus \{j\}$ in a straightforward manner so that $\boldsymbol{\gamma} \in \partial f_2(\boldsymbol{\beta})$. By construction, $\boldsymbol{\gamma} \notin \partial f_1(\boldsymbol{\beta})$, as desired.

In either case, we have that one can choose $\boldsymbol{\gamma} \in \partial f_2(\boldsymbol{\beta})$ so that 1) $\boldsymbol{\gamma}$ is an extreme point of the feasible region $\{\boldsymbol{\gamma} : \sum_i |\gamma_i| \leq \lambda k, \ |\gamma_i| \leq \lambda \ \forall i\}$ and that 2) $\boldsymbol{\gamma} \in \partial f_2(\boldsymbol{\beta}) \setminus \partial f_1(\boldsymbol{\beta})$ whenever

$\partial f_2(\boldsymbol{\beta}) \not\subseteq \partial f_1(\boldsymbol{\beta})$. This concludes the analysis; thus, we have shown the validity (and computational feasibility) of the without loss of generality claim present in Algorithm 1. Indeed, per our analysis, Step 2 in Algorithm 1 can be solved in $O(p \min\{n,p\} + p \log p)$ operations (sorting of $\boldsymbol{\beta}$ in $O(p \log p)$ followed by $O(p)$ conditionals and gradient evaluation in $O(np)$). In reality, if we keep track of gradients in Step 3, there is no need to recompute gradients in Step 2, and therefore in practice Step 2 is of the same complexity of sorting a list of p numbers. (We assume that $\mathbf{X}'\mathbf{y}$ has been computed offline and store throughout for simplicity.)

C.3 Algorithm 2, Step 3

Here we show how to solve Step 3 in Algorithm 2, namely, solving the orthogonal design trimmed Lasso problem

$$\min_{\boldsymbol{\gamma}} \lambda T_k(\boldsymbol{\gamma}) + \frac{\sigma}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 - \langle \mathbf{q}, \boldsymbol{\gamma} \rangle, \tag{54}$$

where $\boldsymbol{\beta}$ and $\mathbf{q}$ are fixed. This is solvable in closed form. Let $\boldsymbol{\alpha} = \boldsymbol{\beta} - \mathbf{q}/\sigma$. First observe that we can rewrite (54) as

$$\begin{aligned}
(54) &= \min_{\boldsymbol{\gamma}} \lambda T_k(\boldsymbol{\gamma}) + \sigma\|\boldsymbol{\gamma} - \boldsymbol{\alpha}\|_2^2/2 \\
&= \min_{\substack{\boldsymbol{\gamma},\mathbf{z}: \\ \sum_i z_i = p-k \\ \mathbf{z}\in\{0,1\}^p}} \lambda\langle \mathbf{z}, |\boldsymbol{\gamma}| \rangle + \sigma\|\boldsymbol{\gamma} - \boldsymbol{\alpha}\|_2^2/2 \\
&= \min_{\substack{\boldsymbol{\gamma},\mathbf{z}: \\ \sum_i z_i = p-k \\ \mathbf{z}\in\{0,1\}^p}} \sum_i \left(\lambda z_i |\gamma| + \sigma(\gamma_i - \alpha_i)^2/2 \right).
\end{aligned}$$

The penultimate step follows via Lemma 2.1. Per this final representation, the solution becomes clear. In particular, let I be a set of k indices of $\boldsymbol{\alpha}$ corresponding to $\alpha_{(1)}, \alpha_{(2)}, \ldots, \alpha_{(k)}$. (If $|\alpha_{(k)}| = |\alpha_{(k+1)}|$, we break ties arbitrarily.) Then a solution $\boldsymbol{\gamma}^*$ to (54) is

$$\gamma_i^* = \begin{cases} \alpha_i, & i \in I \\ \operatorname{soft}_{\lambda/\sigma}(\alpha_i), & i \notin I, \end{cases}$$

where $\operatorname{soft}_{\lambda/\sigma}(\alpha_i) = \operatorname{sgn}(\alpha_i)\,|\alpha_i - \lambda/\sigma|$.

C.4 Computational details

For completeness and reproducibility, we also include all computational details. For Figure 3, the following parameters were used to generate the test instance: $n = 100$, $p = 20$, $\text{SNR} = 10$, `julia` seed $= 1$, $\eta = 0.01$, $k = 2$. The example was generated from the following true model:

1. $\boldsymbol{\beta}_{\text{true}}$ is a vector with ten entries equal to 1 and all others equal to zero. (So $\|\boldsymbol{\beta}_{\text{true}}\|_0 = 10$.)

2. covariance matrix $\boldsymbol{\Sigma}$ is generated with $\Sigma_{ij} = .8^{|i-j|}$.

3. $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

4. $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \boldsymbol{\beta}_0' \boldsymbol{\Sigma} \boldsymbol{\beta}_0/\text{SNR})$

5. $\mathbf{y}$ is then defined as $\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$

The 100 examples generated for Figure 4 were using the following parameters: $n = 100$, $p = 20$, SNR = 10, `julia` seed $\in \{1, \ldots, 100\}$, $\eta = 0.01$, $k = 2$, bigM = 20. MIO using Gurobi solver. Max iterations: alternating minimization—1000; ADMM (inner)—2000; ADMM (outer)—10000. ADMM parameters: $\sigma = 1$, $\tau = 0.9$. The examples themselves had the same structure as the previous example. The optimal gaps shown are relative to the objective in (36). The averages are computed as geometric means (relative to optimal 100%) across the 100 instances, and then displayed relative to the optimal 100%.

Acknowledgments

Copenhaver was partially supported by the Department of Defense, Office of Naval Research, through the National Defense Science and Engineering Graduate (NDSEG) Fellowship. Mazumder was partially supported by ONR Grant N000141512342.

References

[1] L. T. H. An, “Analyse numérique des algorithmes de l’optimisation DC. Approches locale et globale. Codes et simulations numériques en grande dimension. Applications,” Ph.D. dissertation, Université de Rouen, 1994.

[2] L. T. H. An and P. D. Tao, “The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems,” *Annals of Operations Research*, vol. 133, pp. 23–46, 2005.

[3] T. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, New York, 2003.

[4] R. Andreani, L. Secchin, and P. Silva, “Convergence properties of a second order augmented Lagrangian method for mathematical programs with complementarity constraints,” 2017.

[5] A. Bandeira, E. Dobriban, D. Mixon, and W. Sawin, “Certifying the Restricted Isometry Property is hard,” *IEEE Transactions in Information Theory*, vol. 59, pp. 3448–3450, 2013.

[6] D. Bartholomew, M. Knott, and I. Moustaki, *Latent variable models and factor analysis: a unified approach.* Wiley, 2011.

[7] H. Bauschke and P. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces.* Springer, 2011.

[8] A. Beck and A. Ben-Tal, “Duality in robust optimization: primal worst equals dual best,” *Operations Research Letters*, vol. 37, no. 1, pp. 1–6, 2009.

[9] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization.* Princeton University Press, 2009.

[10] D. Bertsekas, “Multiplier methods: a survey,” *Automatica*, vol. 12, no. 2, pp. 133–145, 1976.

[11] ——, *Constrained optimization and Lagrange multiplier methods.* Academic Press, 2014.

[12] D. Bertsimas, D. Brown, and C. Caramanis, “Theory and applications of robust optimization,” *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[13] D. Bertsimas, M. S. Copenhaver, and R. Mazumder, "Certifiably optimal low rank factor analysis," *Journal of Machine Learning Research*, vol. 18, no. 29, pp. 1–53, 2017.

[14] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *The Annals of Statistics*, vol. 44, no. 2, pp. 813–852, 2016.

[15] D. Bertsimas and M. S. Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, 2017.

[16] D. Bertsimas and R. Mazumder, "Least quantile regression via modern optimization," *The Annals of Statistics*, vol. 42, no. 6, pp. 2494–2525, 2014.

[17] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, pp. 1705–1732, 2009.

[18] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. Candès, "SLOPE: Adaptive variable selection via convex optimization," *Annals of Applied Statistics*, vol. 9, pp. 1103–1140, 2015.

[19] P. Bonami, M. Kilinc, and J. Linderoth, *Mixed integer nonlinear programming*. Springer, 2012, ch. Algorithms and software for convex mixed integer nonlinear programs.

[20] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, pp. 1–122, 2011.

[21] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[22] M. Branda, M. Bucher, M. Červinka, and A. Schwartz, "Convergence of a scholtes-type regularization method for cardinality-constrained optimization problems with an application in sparse robust portfolio optimization," *arXiv preprint arXiv:1703.10637*, 2017.

[23] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

[24] O. P. Burdakov, C. Kanzow, and A. Schwartz, "Mathematical programs with cardinality constraints: Reformulation by complementarity-type conditions and a regularization method," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 397–425, 2016.

[25] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications in Pure and Applied Mathematics*, vol. 59, pp. 1207–1223, 2005.

[26] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–37, 2011.

[27] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of Operations Research*, vol. 153, no. 1, pp. 235–256, 2007.

[28] P. Combettes and V. Wajs, "Signal recovering by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–200, 2005.

[29] H. Dong, M. Ahn, and J.-S. Pang, "Structural properties of affine sparsity constraints," *Optimization Online*, 2017.

[30] D. Donoho, “Compressed sensing,” *IEEE Transactions in Information Theory*, vol. 52, pp. 1289–1306, 2006.

[31] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *The Annals of Statistics*, vol. 32, no. 2, pp. 407–99, 2004.

[32] Y. Eldar and G. Kutyniok, Eds., *Compressed sensing: theory and applications.* Cambridge University Press, 2012.

[33] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[34] M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter, “Complementarity formulations of ℓ_0-norm optimization problems,” *Industrial Engineering and Management Sciences. Technical Report. Northwestern University, Evanston, IL, USA*, 2013.

[35] M. Figueiredo and R. Nowak, “Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization,” *arXiv preprint arXiv:1409.4005*, 2014.

[36] I. Frank and J. Friedman, “A statistical view of some chemometrics regression tools,” *Technometrics*, vol. 35, pp. 109–148, 1993.

[37] J. Friedman, “Fast sparse regression and classification,” 2008, technical report, Department of Statistics, Stanford University.

[38] G. Golub and C. V. Loan, “An analysis of the total least squares problem,” *SIAM Journal of Numerical Analysis*, vol. 17, no. 6, pp. 883–893, 1980.

[39] J.-Y. Gotoh, A. Takeda, and K. Tono, “DC formulations and algorithms for sparse optimization problems,” *Preprint, METR*, vol. 27, 2015.

[40] Gurobi Optimization, Inc., “Gurobi optimizer reference manual,” 2016. [Online]. Available: http://www.gurobi.com

[41] L. Guttman, “To what extent can communalities reduce rank?” *Psychometrika*, vol. 23, no. 4, pp. 297–308, 1958.

[42] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2009.

[43] A. B. Hempel and P. J. Goulart, “A novel method for modelling cardinality and rank constraints,” in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 2014, pp. 4322–4327.

[44] T. Hoheisel, C. Kanzow, and A. Schwartz, “Theoretical and numerical comparison of relaxation methods for mathematical programs with complementarity constraints,” *Mathematical Programming*, pp. 1–32, 2013.

[45] P. Huber and E. Ronchetti, *Robust statistics*, 2nd ed. Wiley, 2009.

[46] G. M. James, P. Radchenko, and J. Lv, “DASSO: connections between the dantzig selector and lasso,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 1, pp. 127–142, 2009.

[47] C. Kanzow and A. Schwartz, "A new regularization method for mathematical programs with complementarity constraints with strong convergence properties," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 770–798, 2013.

[48] ——, "The price of inexactness: convergence properties of relaxation methods for mathematical programs with complementarity constraints revisited," *Mathematics of Operations Research*, vol. 40, no. 2, pp. 253–275, 2014.

[49] K. Klamroth, E. Köbis, A. Schöbel, and C. Tammer, "A unified approach to uncertain optimization," *European Journal of Operational Research*, vol. 260, no. 2, pp. 403–420, 2017.

[50] G.-H. Lin and M. Fukushima, "A modified relaxation scheme for mathematical programs with complementarity constraints," *Annals of Operations Research*, vol. 133, no. 1, pp. 63–84, 2005.

[51] H. Liu, T. Yao, and R. Li, "Global solutions to folded concave penalized nonconvex learning," *Annals of Statistics*, vol. 44, no. 2, pp. 629–659, 2016.

[52] H. Liu, T. Yao, R. Li, and Y. Ye, "Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions," *Mathematical Programming*, pp. 1–34, 2016.

[53] K. Mardia, J. Kent, and J. Bibby, *Multivariate analysis.* Academic Press, 1979.

[54] I. Markovsky and S. V. Huffel, "Overview of total least-squares methods," *Signal Processing*, vol. 87, pp. 2283–2302, 2007.

[55] R. Mazumder, J. Friedman, and T. Hastie, "SparseNet: Coordinate descent with nonconvex penalties," *Journal of the American Statistical Association*, vol. 106, pp. 1125–1138, 2011.

[56] R. Mazumder and P. Radchenko, "The discrete Dantzig selector: Estimating sparse linear models via mixed integer linear optimization," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3053–3075, 2017.

[57] A. Miller, *Subset selection in regression.* CRC Press, 2002.

[58] S. Morgenthaler, "A survey of robust statistics," *Statistical Methods and Applications*, vol. 15, pp. 271–293, 2007.

[59] M. Osborne, B. Presnell, and B. Turlach, "On the lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, 2000.

[60] R. Rockafeller, *Convex analysis.* Princeton University Press, 1970.

[61] P. Rousseeuw and A. Leroy, *Robust regression and outlier detection.* Wiley, 1987.

[62] S. Scholtes and M. Stöhr, "Exact penalization of mathematical programs with equilibrium constraints," *SIAM Journal on Control and Optimization*, vol. 37, no. 2, pp. 617–652, 1999.

[63] A. Shapiro, "Rank-reducability of a symmetric matrix and sampling theory of minimum trace factor analysis," *Psychometrika*, vol. 47, pp. 187–199, 1982.

[64] X. Shen, W. Pan, and Y. Zhu, "Likelihood-based selection and sharp parameter estimation," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 223–232, 2012.

[65] X. Shen, W. Pan, Y. Zhu, and H. Zhou, "On constrained and regularized high-dimensional regression," *Annals of the Institute of Statistical Mathematics*, vol. 65, no. 5, pp. 807–832, 2013.

[66] P. D. Tao and L. T. H. An, "Convex analysis approach to DC programming: theory, algorithms, and applications," *Acta Mathematica Vietnamica*, vol. 22, pp. 287–355, 1997.

[67] J. Ten-Berge, "Some recent developments in factor analysis and the search for proper communalities," in *Advances in data science and classification.* Springer, 1998, pp. 325–334.

[68] Y. Teng, L. Yang, B. Yu, and X. Song, "An augmented Lagrangian proximal alternating method for sparse discrete optimization problems," *Optimization Online*, 2017.

[69] M. Thiao, P. D. Tao, and L. An, "A DC programming approach for sparse eigenvalue problem," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 1063–1070.

[70] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[71] A. Tillman and M. Pfetsch, "The computational complexity of the Restricted Isometry Property, the nullspace property, and related concepts in compressed sensing," *IEEE Transactions in Information Theory*, vol. 60, pp. 1248–1259, 2014.

[72] K. Tono, A. Takeda, and J.-Y. Gotoh, "Efficient DC algorithm for constrained sparse optimization," *arXiv preprint arXiv:1701.08498*, 2017.

[73] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and Lasso," *IEEE Transactions in Information Theory*, vol. 56, no. 7, pp. 3561–74, 2010.

[74] C. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, pp. 894–942, 2010.

[75] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, pp. 576–593, 2012.

[76] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *Journal of Machine Learning Research*, vol. 11, pp. 1081–1107, 2010.

[77] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.