Les Cahiers du GERAD ISSN: 0711–2440

Robust Regression and Lasso

H. Xu, C. Caramanis,
S. Mannor

G–2008–77

November 2008

La publication de ces rapports de recherche bénéficie d'une subvention du Fonds québécois de la recherche sur la nature et les technologies.

Robust Regression and Lasso

Huan Xu

GERAD and
Department of Electrical and Computer Engineering
McGill University
xuhuan@cim.mcgill.ca

Constantine Caramanis

Department of Electrical and Computer Engineering
The University of Texas at Austin
cmcaram@ece.utexas.edu

Shie Mannor

GERAD and
Department of Electrical and Computer Engineering
McGill University
shie.mannor@mcgill.ca

November 2008

Les Cahiers du GERAD

G–2008–77

Abstract

Lasso, or ℓ^1 regularized least squares has been explored extensively for its remarkable sparsity properties. The first result of this paper, is that the solution to Lasso, in addition to its sparsity, has robustness properties: it is the solution to a robust optimization problem. This has two important consequences. First, robustness provides a connection of the regularizer to a physical property, namely, protection from noise. This allows principled selection of the regularizer, and in particular, by considering different uncertainty sets, we construct generalizations of Lasso that also yield convex optimization problems.

Secondly, robustness can itself be used as an avenue to exploring different properties of the solution. In particular, we show that robustness of the solution itself explains why the solution is sparse. The analysis as well as the specific results we obtain differ from standard sparsity results, providing different geometric intuition. We next show that the robust optimization formulation is related to kernel density estimation, and following this approach, we use robustness directly to reprove that Lasso is consistent.

Résumé

L'algorithme Lasso, ou recherche de moindres carrés régularisés avec contraintes ℓ^1, a été exploré intensivement vu le caractère minimaliste des solutions qu'il engendre. Le premier résultat de cet article est que la solution Lasso non seulement minimise le nombre de paramètres, mais elle est également robuste; c'est en fait la solution à un problème d'optimisation robuste. Il en résulte deux conséquences importantes : d'abord, vu la robustesse de la solution, l'estimateur hérite d'une propriété physique associée à la régularisation, en l'occurrence la protection vis-à-vis du bruit. Ceci permet une sélection éclairé du type de régularisation requis, et mène à partir d'une considération de différents ensembles incertains, à des généralisations de Lasso correspondant à une classe de problèmes d'optimisation convexe.

Ensuite, différentes propriétés de la solution peuvent être explorées sous l'angle de la robustesse. Nous montrons notamment que la robustesse de la solution explique pourquoi la solution est minimaliste. L'analyse ainsi que les résultats que nous obtenons se distinguent des raisonnements et résultats habituels concernant le caractère minimaliste des solutions, et mènent à des intuitions géométriques nouvelles. Nous démontrons enfin que la formulation de l'optimisation robuste est reliée à l'estimation de la densité du noyau, et ce point de vue est utilisé pour développer une démonstration nouvelle du caractère convergent de Lasso.

1 Introduction

In this paper we consider linear regression problems with least-square error. The problem is to find a vector $\mathbf{x}$ so that the ℓ_2 norm of the residual $\mathbf{b} - A\mathbf{x}$ is minimized, for a given matrix $A \in \mathbb{R}^{n\times m}$ and vector $\mathbf{b} \in \mathbb{R}^n$. From a learning/regression perspective, each row of A can be regarded as a training sample, and the corresponding element of b as the target value of this observed sample. Each column of A corresponds to a feature, and the objective is to find a set of weights so that the weighted sum of the feature values approximates the target value.

It is well known that minimizing the least squared error can lead to sensitive solutions [1–4]. Many regularization methods have been proposed to decrease this sensitivity. Among them, Tikhonov regularization [5] and Lasso [6,7] are two widely known and cited algorithms. These methods minimize a weighted sum of the residual norm and a certain regularization term, $\|\mathbf{x}\|_2$ for Tikhonov regularization and $\|\mathbf{x}\|_1$ for Lasso. In addition to providing regularity, Lasso is also known for the tendency to select sparse solutions. Recently this has attracted much attention for its ability to reconstruct sparse solutions when sampling occurs far below the Nyquist rate, and also for its ability to recover the sparsity pattern exactly with probability one, asymptotically as the number of observations increases (there is an extensive literature on this subject, and we refer the reader to [8–12] and references therein).

The first result of this paper, is that the solution to Lasso, in addition to its sparsity, has robustness properties: it is the solution to a robust optimization problem. In itself, this interpretation of Lasso as the solution to a robust least squares problem is a development in line with the results of [13]. There, the authors propose an alternative approach of reducing sensitivity of linear regression by considering a robust version of the regression problem, i.e., minimizing the worst-case residual for the observations under some unknown but bounded disturbance. Most of the research in this area considers either the case where the disturbance is row-wise uncorrelated [14], or the Frobenius norm of the disturbance matrix is bounded [13].

None of these robust optimization approaches produces a solution that has sparsity properties (in particular, the solution to Lasso does not solve any of these previously formulated robust optimization problems). In contrast, we investigate the robust regression problem where the uncertainty set is defined by feature-wise constraints. Such a noise model is of interest when values of features are obtained with some noisy pre-processing steps, and the magnitudes of such noises are known or bounded. Another situation of interest is where features are meaningfully correlated. We define *correlated* and *uncorrelated* disturbances and uncertainty sets precisely in Section 2.1 below. Intuitively, a disturbance is feature-wise correlated if the variation or disturbance across features satisfy joint constraints, and uncorrelated otherwise.

Considering the solution to Lasso as the solution of a robust least squares problem has two important consequences. First, robustness provides a connection of the regularizer to a physical property, namely, protection from noise. This allows more principled selection of the regularizer, and in particular, considering different uncertainty sets, we construct generalizations of Lasso that also yield convex optimization problems.

Secondly, and perhaps most significantly, robustness is a strong property that can itself be used as an avenue to investigating different properties of the solution. We show that robustness of the solution can explain why the solution is sparse. The analysis as well as the specific results we obtain differ from standard sparsity results, providing different geometric intuition, and extending beyond the least-squares setting. Sparsity results obtained for Lasso ultimately depend on the fact that introducing additional features incurs larger ℓ^1-penalty than the least squares error reduction. In contrast, we exploit the fact that a robust solution is, by definition, the optimal solution under a worst-case perturbation. Our results show that, essentially, a coefficient of the solution is nonzero if the corresponding feature is relevant under all allowable perturbations. In addition to sparsity, we also use robustness directly to prove consistency of Lasso.

We briefly list the main contributions of this paper.

- We formulate the robust regression problem with feature-wise independent disturbances, and show that this formulation is equivalent to a least-square problem with a weighted ℓ_1 norm regularization term. Hence, we provide an interpretation for Lasso from a robustness perspective.
- We generalize the robust regression formulation to loss functions of arbitrary norm, which we use below to extend our sparsity results to this domain as well. We also consider uncertainty sets that require disturbances of different features to satisfy joint conditions. This can be used to mitigate the conservativeness of the robust solution, and also obtain solutions with additional properties. We call these features "correlated". We further consider uncertainty sets with both column-wise and feature-wise disturbances. In particular, we formulate a class of robust-regression problems which smoothly interpolate between Lasso and a (possibly non-sparse) ℓ_∞-norm regularizer.
- We present new sparsity results for the robust regression problem with feature-wise independent disturbances. This provides a new robustness-based explanation for why Lasso produces sparse solutions. Our approach gives new analysis and also geometric intuition, and furthermore allows one to obtain sparsity results for more general loss functions, beyond the squared loss.
- Next, we relate Lasso to kernel density estimation. This allows us to re-prove consistency in a statistical learning setup, using the new robustness tools and formulation we introduce. Along with our results on sparsity, this illustrates the power of robustness in explaining and also exploring different properties of the solution.
- Finally, we prove a "no-free-lunch" theorem, stating that an algorithm that encourages sparsity fails to have a non-trivial stability bound.

This paper is organized as follows. In Section 2 we formulate and solve the robust regression setup with uncorrelated disturbance, which we show to be equivalent to Lasso. The robust regression for general uncertainty sets is considered in Section 3. We investigate the sparsity of the robust regression in Section 4. In Section 5 we relate robust regression problems to kernel density estimation. We provide the "no-free-lunch" result in Section 6.

Notation. We use capital letters to represent matrices, and boldface letters to represent column vectors. Row vectors are represents as the transpose of column vectors. For a vector $\mathbf{z}$, z_i denotes its i^{th} element. Throughout the paper, $\mathbf{a}_i$ and $\mathbf{r}_j^\top$ are used to denote the i^{th} column and the j^{th} row of the observation matrix A, respectively. We use a_{ij} to denote the ij element of A, hence it is the j^{th} element of $\mathbf{r}_i$, and i^{th} element of $\mathbf{a}_j$. For a convex function $f(\cdot)$, $\partial f(\mathbf{z})$ represents any of its sub-gradients evaluated at $\mathbf{z}$.

2 Robust Regression with Feature-wise Disturbance

In this section, we show that our robust regression formulation recovers Lasso as a special case. We also derive probabilistic bounds that guide in the construction of the uncertainty set.

The regression formulation we consider differs from the standard Lasso formulation, as we minimize the norm of the error, rather than the squared norm. It is known that these two coincide up to a change of the regularization coefficient. Yet as we discuss above, our results amount to more than a representation or equivalence theorem. In addition to more flexible and potentially powerful robust formulations, we prove new results, and give new insight into known results.

2.1 Formulation

Robust linear regression considers the case where the observed matrix is corrupted by some potentially malicious disturbance. The objective is to find the optimal solution in the worst case sense. This is usually formulated as the following min-max problem,

Robust Linear Regression:

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\mathcal{U}}\|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_2\right\}, \tag{1}$$

where $\mathcal{U}$ is called the *uncertainty set*, or the set of admissible disturbances of the matrix A. In this section, we consider the class of uncertainty sets that bound the norm of the disturbance to each feature, without placing any joint requirements across feature disturbances. That is, we consider the class of uncertainty sets:

$$\mathcal{U}\triangleq\left\{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\Big|\|\boldsymbol{\delta}_i\|_2\le c_i,\ \ i=1,\cdots,m\right\}, \tag{2}$$

for given $c_i\ge 0$. We call these uncertainty sets *feature-wise uncorrelated*, in contrast to *correlated* uncertainty sets that require disturbances of different features to satisfy some joint constraints (we discuss these extensively below, and their significance). While the inner maximization problem of (1) is nonconvex, we show in the next theorem that uncorrelated norm-bounded uncertainty sets lead to an easily solvable optimization problem.

Theorem 1 *The robust regression problem (1) with uncertainty set of the form (2) is equivalent to the following ℓ^1 regularized regression problem:*

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\|\mathbf{b}-A\mathbf{x}\|_2+\sum_{i=1}^m c_i|x_i|\right\}. \tag{3}$$

Proof. Fix $\mathbf{x}^*$. We prove that $\max_{\Delta A\in\mathcal{U}}\|\mathbf{b}-(A+\Delta A)\mathbf{x}^*\|_2=\|\mathbf{b}-A\mathbf{x}^*\|_2+\sum_{i=1}^m c_i|x_i^*|$.

The left hand side can be written as

$$\begin{aligned}
&\max_{\Delta A\in\mathcal{U}}\|\mathbf{b}-(A+\Delta A)\mathbf{x}^*\|_2\\
=&\max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)|\|\boldsymbol{\delta}_i\|_2\le c_i}\Big\|\mathbf{b}-\big(A+(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\big)\mathbf{x}^*\Big\|_2\\
=&\max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)|\|\boldsymbol{\delta}_i\|_2\le c_i}\|\mathbf{b}-A\mathbf{x}^*-\sum_{i=1}^m x_i^*\boldsymbol{\delta}_i\|_2\\
\le&\max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)|\|\boldsymbol{\delta}_i\|_2\le c_i}\Big\|\mathbf{b}-A\mathbf{x}^*\Big\|_2+\sum_{i=1}^m\|x_i^*\boldsymbol{\delta}_i\|_2\\
\le&\|\mathbf{b}-A\mathbf{x}^*\|_2+\sum_{i=1}^m|x_i^*|c_i.
\end{aligned} \tag{4}$$

Now, let

$$\mathbf{u} \triangleq \begin{cases} \frac{\mathbf{b}-A\mathbf{x}^*}{\|\mathbf{b}-A\mathbf{x}^*\|_2} & \text{if } A\mathbf{x}^* \neq \mathbf{b}, \\ \text{any vector with unit } \ell^2 \text{ norm} & \text{otherwise;} \end{cases}$$

and let

$$\boldsymbol{\delta}_i^* \triangleq -c_i \text{sgn}(x_i^*)\mathbf{u}.$$

Observe that $\|\boldsymbol{\delta}_i^*\|_2 \leq c_i$, hence $\Delta A^* \triangleq (\boldsymbol{\delta}_1^*, \cdots, \boldsymbol{\delta}_m^*) \in \mathcal{U}$. Notice that

$$\begin{aligned} &\max_{\Delta A \in \mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}^*\|_2 \\ \geq &\|\mathbf{b} - (A + \Delta A^*)\mathbf{x}^*\|_2 \\ = &\Big\|\mathbf{b} - \big(A + (\boldsymbol{\delta}_1^*, \cdots, \boldsymbol{\delta}_m^*)\big)\mathbf{x}^*\Big\|_2 \\ = &\Big\|(\mathbf{b} - A\mathbf{x}^*) - \sum_{i=1}^m \big(-x_i^* c_i \text{sgn}(x_i^*)\mathbf{u}\big)\Big\|_2 \\ = &\Big\|(\mathbf{b} - A\mathbf{x}^*) + (\sum_{i=1}^m c_i|x_i^*|)\mathbf{u}\Big\|_2 \\ = &\|\mathbf{b} - A\mathbf{x}^*\|_2 + \sum_{i=1}^m c_i|x_i^*|. \end{aligned} \tag{5}$$

The last equation holds from the definition of $\mathbf{u}$.

Combining Inequalities (4) and (5), establishes the equality $\max_{\Delta A \in \mathcal{U}} \|\mathbf{b}-(A+\Delta A)\mathbf{x}^*\|_2 = \|\mathbf{b}-A\mathbf{x}^*\|_2 + \sum_{i=1}^m c_i|x_i^*|$ for any $\mathbf{x}^*$. Minimizing over $\mathbf{x}$ on both sides proves the theorem. □

Taking $c_i = c$ and normalizing $\mathbf{a}_i$ for all i, Problem (3) recovers the well-known Lasso [6,7].

2.2 Uncertainty Set Construction

The selection of an uncertainty set $\mathcal{U}$ in Robust Optimization is of fundamental importance. One way this can be done is as an approximation of so-called *chance constraints*, where a deterministic constraint is replaced by the requirement that a constraint is satisfied with at least some probability. These can be formulated when we know the distribution exactly, or when we have only partial information of the uncertainty, such as, e.g., first and second moments. This chance-constraint formulation is particularly important when the distribution has large support, rendering the naive robust optimization formulation overly pessimistic.

For confidence level η, the chance constraint formulation becomes:

$$\begin{aligned} \text{minimize:} \quad & t \\ \text{Subject to:} \quad & \Pr(\|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_2 \leq t) \geq 1 - \eta. \end{aligned}$$

Here, $\mathbf{x}$ and t are the decision variables.

Constructing the uncertainty set for feature i can be done quickly via line search and bisection, as long as we can evaluate $\Pr(\|\mathbf{a}_i\|_2 \geq c)$. If we know the distribution exactly (i.e., if we have complete probabilistic information), this can be quickly done via sampling. Another setting of interest is when we have access only to some moments of the distribution of the uncertainty, e.g., the mean and variance. In this setting, the uncertainty sets are constructed via a bisection procedure which evaluates the worst-case probability over all distributions with given mean and variance. We do this using a tight bound on the probability of an event, given the first two moments.

In the scalar case, the Markov Inequality provides such a bound. The next theorem is a generalization of the Markov inequality to $\mathbb{R}^n$, which bounds the probability where the disturbance on a given feature is more than c_i, if only the first and second moment of the random variable are known. We postpone the proof to the Appendix, and refer the reader to [15] for similar results using semi-definite optimization.

Theorem 2 *Consider a random vector $\mathbf{v} \in \mathbb{R}^n$, such that $\mathbb{E}(\mathbf{v}) = \mathbf{a}$, and $\mathbb{E}(\mathbf{v}\mathbf{v}^\top) = \Sigma$, $\Sigma \succeq 0$. Then we have*

$$\Pr\{\|\mathbf{v}\|_2 \geq c_i\} \leq \left\{ \begin{array}{ll} \min_{P,\mathbf{q},r,\lambda} & \mathrm{Trace}(\Sigma P) + 2\mathbf{q}^\top \mathbf{a} + r \\ \textit{subject to:} & \begin{pmatrix} P & \mathbf{q} \\ mathbf{q}^\top & r \end{pmatrix} \succeq 0 \\ & \begin{pmatrix} I(m) & \mathbf{0} \\ \mathbf{0}^\top & -c_i^2 \end{pmatrix} \preceq \lambda \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^\top & r-1 \end{pmatrix} \\ & \lambda \geq 0. \end{array} \right. \tag{6}$$

The optimization problem (13) is a semi-definite programming, which is known be solved efficiently. Furthermore, if we replace $\mathbb{E}(\mathbf{v}\mathbf{v}^\top) = \Sigma$ by an inequality $\mathbb{E}(\mathbf{v}\mathbf{v}^\top) \leq \Sigma$, the uniform bound still holds. Thus, even if our estimation to the variance is not precise, we are still able to bound the probability of having "large" disturbance.

3 General Uncertainty Sets

One reason the robust optimization formulation is powerful, is that having provided the connection to Lasso, it then allows the opportunity to generalize to efficient "Lasso-like" regularization algorithms.

In this subsection, we make several generalizations of the robust formulation (1) and derive counterparts of Theorem 1. We generalize the robust formulation in two ways: (a) to the case of arbitrary norm; and (b) to the case of correlated uncertainty sets. In Section 3.2 we investigate a class of uncertainty sets inspired by [16], that control the *cardinality* of perturbed features. The uncertainty sets are non-convex, but nevertheless we show that the resulting robust regression problem is still tractable. In the last subsection, we consider a disturbance model where both column-wise disturbance and row-wise disturbance exist simultaneously.

3.1 Arbitrary norm and Correlated disturbance

We first consider the case of an arbitrary norm $\|\cdot\|_a$ of $\mathbb{R}^n$ as a cost function rather than the squared loss. Recall that a norm must satisfy

$$\begin{array}{ll} (1). & \|\mathbf{x}\|_a \geq 0,\ \forall \mathbf{x} \in \mathbb{R}^n; \quad \|\mathbf{x}\|_a = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}. \\ (2). & \|c\mathbf{x}\|_a = c\|\mathbf{x}\|_a,\ \forall c \geq 0, \forall \mathbf{x} \in \mathbb{R}^n. \\ (3). & \|\mathbf{x}+\mathbf{y}\|_a \leq \|\mathbf{x}\|_a + \|\mathbf{y}\|_a,\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \end{array}$$

The proof of the next theorem is identical to that of Theorem 1, with only the ℓ^2 norm changed to $\|\cdot\|_a$.

Theorem 3 *The robust regression problem*

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\mathcal{U}_a}\|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_a\right\};\ \ \mathcal{U}_a\triangleq\left\{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\Big|\|\boldsymbol{\delta}_i\|_a\le c_i,\ \ i=1,\cdots,m\right\};$$

is equivalent to the following regularized regression problem

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\|\mathbf{b}-A\mathbf{x}\|_a+\sum_{i=1}^m c_i|x_i|\right\}.$$

We next remove the assumption that the disturbances are feature-wise uncorrelated. Allowing correlated uncertainty sets is useful when we have some additional information about potential noise in the problem, and we want to limit the conservativeness of the worst-case formulation. Consider the following uncertainty set:

$$\mathcal{U}'\triangleq\left\{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\big|f_j(\|\boldsymbol{\delta}_1\|_a,\cdots,\|\boldsymbol{\delta}_m\|_a)\le 0;\ j=1,\cdots,k\right\},$$

where $f_j(\cdot)$ are convex functions. Notice that, both k and f_j can be arbitrary, hence this is a very general formulation, and provides us with significant flexibility in designing uncertainty sets and equivalently new regression algorithms (see for example Corollary 1 and 2). The following theorem converts this formulation to tractable optimization problems.

Theorem 4 *Assume that the set*

$$\mathcal{Z}\triangleq\{\mathbf{z}\in\mathbb{R}^m|f_j(\mathbf{z})\le 0,\ j=1,\cdots,k;\ \mathbf{z}\ge\mathbf{0}\}$$

has non-empty relative interior. Then the robust regression problem

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\mathcal{U}'}\|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_a\right\}$$

is equivalent to the following regularized regression problem

$$\begin{aligned}&\min_{\boldsymbol{\lambda}\in\mathbb{R}^k_+,\boldsymbol{\kappa}\in\mathbb{R}^m_+,\mathbf{x}\in\mathbb{R}^m}\left\{\|\mathbf{b}-A\mathbf{x}\|_a+v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})\right\};\\ &\textit{where: } v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})\triangleq\max_{\mathbf{c}\in\mathbf{R}^m}\left[(\boldsymbol{\kappa}+|\mathbf{x}|)^\top\mathbf{c}-\sum_{j=1}^k\lambda_j f_j(\mathbf{c})\right]\end{aligned}\tag{7}$$

We postpone the proof to the Appendix.

Remark 1 Problem (14) is efficiently solvable. Denote $z^{\mathbf{c}}(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})\triangleq\left[(\boldsymbol{\kappa}+|\mathbf{x}|)^\top\mathbf{c}-\sum_{j=1}^k\lambda_j f_j(\mathbf{c})\right]$. This is a convex function of $(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})$, and the sub-gradient of $z^{\mathbf{c}}(\cdot)$ can be computed easily for any $\mathbf{c}$. The function $v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})$ is the maximum of a set of convex functions, $z^{\mathbf{c}}(\cdot)$, hence is convex, and satisfies

$$\partial v(\boldsymbol{\lambda}^*,\boldsymbol{\kappa}^*,\mathbf{x}^*)=\partial z^{\mathbf{c}_0}(\boldsymbol{\lambda}^*,\boldsymbol{\kappa}^*,\mathbf{x}^*),$$

where $\mathbf{c}_0$ maximizes $\left[(\boldsymbol{\kappa}^*+|\mathbf{x}|^*)^\top\mathbf{c}-\sum_{j=1}^k\lambda_j^* f_j(\mathbf{c})\right]$. We can efficiently evaluate $\mathbf{c}_0$ due to convexity of $f_j(\cdot)$, and hence we can efficiently evaluate the sub-gradient of $v(\cdot)$.

The next two corollaries are a direct application of Theorem 4.

Corollary 1 *Suppose* $\mathcal{U}' = \Big\{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m)\Big| \big\| \|\boldsymbol{\delta}_1\|_a, \cdots, \|\boldsymbol{\delta}_m\|_a \big\|_s \le l; \Big\}$ *for a symmetric norm* $\|\cdot\|_s$*, then the resulting regularized regression problem is*

$$\min_{\mathbf{x}\in\mathbb{R}^m} \Big\{\|\mathbf{b} - A\mathbf{x}\|_a + l\|\mathbf{x}\|_s^*\Big\}; \quad \text{where } \|\cdot\|_s^* \text{ is the dual norm of } \|\cdot\|_s.$$

This corollary interprets *arbitrary* norm-based regularizers from a robust regression perspective. For example, it is straightforward to show that if we take both $\|\cdot\|_\alpha$ and $\|\cdot\|_s$ as the Euclidean norm, then $\mathcal{U}'$ is the set of matrices with their Frobenious norms bounded, and Corollary 1 reduces to the robust formulation introduced by [13].

Corollary 2 *Suppose* $\mathcal{U}' = \Big\{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m)\Big| \exists \mathbf{c} \ge \mathbf{0} : T\mathbf{c} \le \mathbf{s};\ \|\boldsymbol{\delta}_j\|_a \le c_j; \Big\}$*, then the resulting regularized regression problem is*

$$\begin{aligned} \textit{Minimize:} \quad & \|\mathbf{b} - A\mathbf{x}\|_a + \mathbf{s}^\top \boldsymbol{\lambda} \\ \textit{Subject to:} \quad & \mathbf{x} \le T^\top \boldsymbol{\lambda} \\ & -\mathbf{x} \le T^\top \boldsymbol{\lambda} \\ & \boldsymbol{\lambda} \ge \mathbf{0}. \end{aligned}$$

Unlike previous results, this corollary considers general polytope uncertainty sets. Advantages of such sets include the linearity of the final formulation. Moreover, the modeling power is considerable, as many interesting disturbances can be modeled in this way. One such disturbance model is the so-called *cardinality constrained* uncertainty set, where a only a fixed number of features are corrupted. We turn to this in the next section.

3.2 A class of non-convex uncertainty sets

Theorem 4 deals with convex uncertainty sets. Next we consider a class of non-convex but still solvable uncertainty sets, which can be regarded as interpolations between the uncorrelated case and the fully correlated case. To be specific, we consider the case that no more than a given number of features are disturbed. This formulation is inspired by [16] in which a similar uncertainty set for robust LP is considered. Let

$$\begin{aligned} \mathcal{Z}_t \triangleq \{\mathbf{z} \in \mathbb{R}^m \big| & \exists S \subseteq \{1, \cdots, m\}, |S| = \lfloor t \rfloor, \forall i \in S, 0 \le z_i \le c_i; \\ & \exists j \in \{1, \cdots, m\} \backslash S; 0 \le z_j \le (t - \lfloor t \rfloor) c_j;\ \forall k \notin S \cup \{j\}, z_k = 0\}. \end{aligned}$$

$$\mathcal{U}_t \triangleq \big\{\big(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m\big) \big| \exists \mathbf{z} \in \mathcal{Z}_t,\ \|\boldsymbol{\delta}_i\|_a = z_i.\big\}$$

Here, $\lfloor t \rfloor$ stands for the largest integer not larger than t. $\mathcal{U}_t$ represents an uncertainty set, such that the deviation of each feature is bounded by c_i and only t features are allowed to deviate. For t being a non-integer, it is interpreted as to allow $\lfloor t \rfloor$ features to completely deviate, and one other feature to partially deviate. Neither $\mathcal{Z}_t$ nor $\mathcal{U}_t$ is a convex set. Nevertheless, the robust regression problem with $\mathcal{U}_t$ as the uncertainty set is still tractable because it is equivalent to a robust regression problem with the following polyhedral uncertainty set:

$$\tilde{\mathcal{Z}}_t \triangleq \left\{\mathbf{z} \in \mathbb{R}^m \big| 0 \le z_i \le c_i;\ \sum_{i=1}^m z_i/c_i \le t\right\};$$

$$\tilde{\mathcal{U}}_t \triangleq \Big\{\big(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m\big) \big| \exists \mathbf{z} \in \tilde{\mathcal{Z}}_t,\ \|\boldsymbol{\delta}_i\|_a = z_i.\Big\}.$$

Notice that, $\tilde{\mathcal{U}}_t$ itself has an intuitive appealing interpretation as the set of disturbances such that besides the norm bound for disturbance on each feature, there exists an extra constraint which bounds the (weighted) total disturbance.

Proposition 1 *For any $\mathbf{x}^*$, and $1 \leq t \leq m$, the following holds*

$$\max_{\Delta A \in \mathcal{U}_t} \|\mathbf{b} - (A + \Delta A)\mathbf{x}^*\|_a \equiv \max_{\Delta A \in \tilde{\mathcal{U}}_t} \|\mathbf{b} - (A + \Delta A)\mathbf{x}^*\|_a$$

Combining Proposition 1 and Theorem 4 leads to the following corollary.

Corollary 3 *The robust regression problem*

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}_t} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_a \right\};$$

is equivalent to the following regularized regression problem

$$\begin{aligned}
\textit{Minimize:} \quad & \|\mathbf{b} - A\mathbf{x}\|_a + \sum_{i=1}^m c_i \lambda_i + t\xi \\
\textit{Subject to:} \quad & x_i - \lambda_i - \xi/c_i \leq 0, \quad i = 1, \cdots, m \\
& -x_i - \lambda_i - \xi/c_i \leq 0, \quad i = 1, \cdots, m \\
& \lambda_i \geq 0, \quad i = 1, \cdots, m \\
& \xi \geq 0.
\end{aligned}$$

If all the c_i are same, the robust regression with $\mathcal{U}_m$ (a non-correlated set) is Lasso, while the robust regression with $\mathcal{U}_1$ (a fully correlated set) leads to a ℓ^∞ norm regularization, which is known to be non-sparse.

3.3 Row and Column Uncertainty Case

Next we consider a case where we have both row-wise uncertainty and column-wise uncertainty. One motivation to consider this is the well-known elastic net method ([17]) known to sometimes outperform Lasso, in addition to possessing other properties of interest.

Combing row-wise and column-wise uncertainty leads to the following robust optimization problem

$$\begin{aligned}
& \min_{\mathbf{x}} \max_{\Delta A_1 \in \mathcal{U}_1, \Delta A_2 \in \mathcal{U}_2} \|\mathbf{b} - (A + \Delta A_1 + \Delta A_2)\mathbf{x}\|_2, \\
\text{where:} \quad & \mathcal{U}_1 = \left\{ (\mathbf{l}_1, \cdots, \mathbf{l}_n)^\top \Big| \mathbf{l}_j^\top \Sigma_j^{-1} \mathbf{l}_j \leq 1, \quad i = 1, \cdots, n \right\}; \\
& \mathcal{U}_2 = \left\{ (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) \Big| \|\boldsymbol{\delta}_i\|_2 \leq c_i, \ i = 1, \cdots, m \right\};
\end{aligned} \tag{8}$$

for positive definite matrices Σ_j and positive scalars c_i.

Theorem 5 *Denote the j^{th} row of A as $\mathbf{r}_j^\top$. Then given $\mathbf{x}$, the following holds*

$$\max_{\Delta A_1 \in \mathcal{U}_1, \Delta A_2 \in \mathcal{U}_2} \|\mathbf{b} - (A + \Delta A_1 + \Delta A_2)\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^n \left(|b_j - \mathbf{r}_j^\top \mathbf{x}| + \|\Sigma_j^{1/2} \mathbf{x}\|_2 \right)^2} + \sum_{i=1}^m c_i |x_i|,$$

and moreover, the robust regression problem (8) is equivalent to the following Second Order Cone Program on $(\mathbf{x}, \mathbf{z}, \mathbf{t}, w)$*:*

$$\begin{aligned}
\textit{Minimize:}\quad & w + \sum_{i=1}^{m} c_i z_i \\
\textit{Subject to:}\quad & \mathbf{x} \le \mathbf{z}; \\
& -\mathbf{x} \le \mathbf{z} \\
& \|\Sigma_j^{1/2}\mathbf{x}\|_2 \le t_j - b_j + \mathbf{r}_j^\top \mathbf{x}; \quad j = 1, \cdots, n. \\
& \|\Sigma_j^{1/2}\mathbf{x}\|_2 \le t_j + b_j - \mathbf{r}_j^\top \mathbf{x}; \quad j = 1, \cdots, n. \\
& \|\mathbf{t}\|_2 \le w.
\end{aligned}$$

The proof is postponed to the Appendix.

4 Sparsity

In this section, we investigate the sparsity properties of robust regression (1), and equivalently Lasso. Lasso's ability to recover sparse solutions has been extensively studied and discussed (cf. [8–11]), and this work generally takes one of two approaches. The first approach investigates the problem from a statistical perspective. That is, it assumes that the observations are generated by a (sparse) linear combination of the features, and investigates the asymptotic or probabilistic conditions required for Lasso to correctly recover the generative model. The second approach treats the problem from an optimization perspective, and studies under what conditions a pair $(A,\mathbf{b})$ defines a problem with sparse solutions (e.g., [18]).

We follow the second approach and do not assume a generative model. Instead, we consider the conditions that lead to a feature receiving zero weight. Our first result paves the way for the remainder of this section. We show in Theorem 6 that, essentially, a feature receives no weight (namely, $x_i^* = 0$) if there exists an allowable perturbation of that feature which makes it irrelevant. This result holds for general norm loss functions, but in the ℓ^2 case, we obtain further geometric results. Using Theorem 6, we show, among other results, that (a) "nearly" orthogonal features get zero weight (Theorem 7); and (b) "nearly" linearly dependent features get zero weight (Theorem 9).

Substantial research regarding sparsity properties of Lasso can be found in the literature (cf. [8–11, 19–22] and many others). In particular, similar results as in point (a), that rely on an *incoherence* property, have been established in, e.g., [18], and are used as standard tools in investigating sparsity of Lasso from the statistical perspective. However, a proof exploiting robustness and properties of the uncertainty is novel. Indeed, such a proof shows a fundamental connection between robustness and sparsity, and implies that robustifying w.r.t. a feature-wise independent uncertainty set might be a plausible way to achieve sparsity for other problems.

To state the main theorem of this section, from which the other results derive, we introduce some notation to facilitate the discussion. Given a feature-wise uncorrelated uncertainty set, $\mathcal{U}$, an index subset $I \subseteq \{1, \ldots, n\}$, and any $\Delta A \in \mathcal{U}$, let ΔA^I denote the element of $\mathcal{U}$ that equals ΔA on each feature indexed by $i \in I$, and is zero elsewhere. Then, we can write any element $\Delta A \in \mathcal{U}$ as $\Delta A^I + \Delta A^{I^c}$ (where $I^c = \{1, \ldots, n\} \setminus I$). Then we have the following theorem. We note that the result holds for any norm loss function, but we state and prove it for the ℓ^2 norm, since the proof for other norms is identical.

Theorem 6 *The robust regression problem*

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\mathcal{U}}\|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_2\right\},$$

has a solution supported on an index set I, if there exists some perturbation $\Delta A^{I^c}\in\mathcal{U}$ of the features in I^c, such that the robust regression problem

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A^I\in\mathcal{U}^I}\|\mathbf{b}-(A+\Delta A^I+\Delta A^{I^c})\mathbf{x}\|_2\right\},$$

has a solution supported on the set I.

Thus, a robust regression has an optimal solution supported on a set I, if *any* perturbation of the features corresponding to the complement of I makes them irrelevant. An equivalent statement of the theorem is:

Theorem 6' *Let $\mathbf{x}^*$ be an optimal solution of the robust regression problem:*

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\mathcal{U}}\|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_2\right\},$$

and let $I\subseteq\{1,\cdots,m\}$ be such that $x_j^*=0\ \forall j\notin I$. Let

$$\tilde{\mathcal{U}}\triangleq\left\{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\Big|\|\boldsymbol{\delta}_i\|_2\le c_i,\ \ i\in I;\ \ \|\boldsymbol{\delta}_j\|_2\le c_j+l_j,\ j\notin I\right\}.$$

Then, $\mathbf{x}^*$ is an optimal solution of

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\tilde{\mathcal{U}}}\|\mathbf{b}-(\tilde{A}+\Delta A)\mathbf{x}\|_2\right\},$$

for any $\tilde{A}$ that satisfies $\|\tilde{\mathbf{a}}_j-\mathbf{a}_j\|\le l_j$ for $j\notin I$, and $\tilde{\mathbf{a}}_i=\mathbf{a}_i$ for $i\in I$.

Proof. Notice that

$$\begin{aligned}&\max_{\Delta A\in\tilde{\mathcal{U}}}\Big\|\mathbf{b}-(A+\Delta A)\mathbf{x}^*\Big\|_2\\=&\max_{\Delta A\in\mathcal{U}}\Big\|\mathbf{b}-(A+\Delta A)\mathbf{x}^*\Big\|_2\\=&\max_{\Delta A\in\mathcal{U}}\Big\|\mathbf{b}-(\tilde{A}+\Delta A)\mathbf{x}^*\Big\|_2.\end{aligned}$$

These equalities hold because for $j\notin I$, $x_j^*=0$, hence the j^{th} column of both $\tilde{A}$ and ΔA has no effect on the residual.

For an arbitrary $\mathbf{x}'$, we have

$$\begin{aligned}&\max_{\Delta A\in\tilde{\mathcal{U}}}\Big\|\mathbf{b}-(A+\Delta A)\mathbf{x}'\Big\|_2\\\ge&\max_{\Delta A\in\mathcal{U}}\Big\|\mathbf{b}-(\tilde{A}+\Delta A)\mathbf{x}'\Big\|_2.\end{aligned}$$

This is because, $\|\mathbf{a}_j - \tilde{\mathbf{a}}_j\| \leq l_j$ for $j \notin I$, and $\mathbf{a}_i = \tilde{\mathbf{a}}_i$ for $i \in I$. Hence, we have

$$\big\{A + \Delta A \big| \Delta A \in \mathcal{U}\big\} \subseteq \big\{\tilde{A} + \Delta A \big| \Delta A \in \tilde{\mathcal{U}}\big\}.$$

Finally, notice that

$$\max_{\Delta A \in \mathcal{U}} \Big\|\mathbf{b} - (A + \Delta A)\mathbf{x}^*\Big\|_2 \leq \max_{\Delta A \in \mathcal{U}} \Big\|\mathbf{b} - (A + \Delta A)\mathbf{x}'\Big\|_2.$$

Therefore we have

$$\max_{\Delta A \in \tilde{\mathcal{U}}} \Big\|\mathbf{b} - (\tilde{A} + \Delta A)\mathbf{x}^*\Big\|_2 \leq \max_{\Delta A \in \tilde{\mathcal{U}}} \Big\|\mathbf{b} - (\tilde{A} + \Delta A)\mathbf{x}'\Big\|_2.$$

Since this holds for arbitrary $\mathbf{x}'$, we establish the theorem. □

We can understand the result of this theorem by considering a generative model[1] $b = \sum_{i \in I} w_i a_i + \tilde{\xi}$ where $I \subseteq \{1 \cdots, m\}$ and $\tilde{\xi}$ is a random variable, i.e., b is generated by features belonging to I. In this case, for a feature $j \notin I$, Lasso would assign zero weight as long as there exists a perturbed value of this feature, such that the optimal regression assigned it zero weight.

When we have ℓ^2 loss, we can translate the condition of a feature being "irrelevant" into a geometric condition, namely, orthogonality. We now use the result of Theorem 6 to show that robust regression has a sparse solution as long as an incoherence-type property is satisfied. This result is more in line with the traditional sparsity results, but we note that the geometric reasoning is different, and ours is based on robustness. Indeed, we show that a feature receives zero weight, if it is "nearly" (i.e., within an allowable perturbation) orthogonal to the signal, and all relevant features.

Theorem 7 *Let $c_i = c$ for all i. If there exists $I \subset \{1, \cdots, m\}$ such that for all $\mathbf{v} \in \mathrm{span}\big(\{\mathbf{a}_\mathrm{i}, \mathrm{i} \in \mathrm{I}\} \bigcup \{\mathbf{b}\}\big)$, $\|\mathbf{v}\| = 1$, we have $\mathbf{v}^\top \mathbf{a}_j \leq c \ \forall j \notin I$, then any optimal solution $\mathbf{x}^*$ satisfies $x_j^* = 0$, $\forall j \notin I$.*

Proof. For $j \notin I$, let $\mathbf{a}_j^=$ denote the projection of $\mathbf{a}_j$ onto the span of $\{\mathbf{a}_i,\ i \in I\} \bigcup \{\mathbf{b}\}$, and let $\mathbf{a}_j^+ \triangleq \mathbf{a}_j - \mathbf{a}_j^=$. Thus, we have $\|\mathbf{a}_j^=\| \leq c$. Let $\hat{A}$ be such that

$$\hat{\mathbf{a}}_i = \begin{cases} \mathbf{a}_i & i \in I; \\ \mathbf{a}_i^+ & i \notin I. \end{cases}$$

Now let

$$\hat{\mathcal{U}} \triangleq \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c,\ i \in I; \|\boldsymbol{\delta}_j\|_2 = 0,\ j \notin I\}.$$

Consider the robust regression problem $\min_{\hat{\mathbf{x}}} \Big\{ \max_{\Delta A \in \hat{\mathcal{U}}} \big\|\mathbf{b} - (\hat{A} + \Delta A)\hat{\mathbf{x}}\big\|_2 \Big\}$, which is equivalent to $\min_{\hat{\mathbf{x}}} \Big\{ \big\|\mathbf{b} - \hat{A}\hat{\mathbf{x}}\big\|_2 + \sum_{i \in I} c|\hat{x}_i| \Big\}$. Note that the $\hat{\mathbf{a}}_j$ are orthogonal to the span of $\{\hat{\mathbf{a}}_i,\ i \in I\} \bigcup \{\mathbf{b}\}$. Hence for any given $\hat{\mathbf{x}}$, by changing $\hat{x}_j$ to zero for all $j \notin I$, the minimizing objective does not increase.

Since $\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_j\| = \|\mathbf{a}_j^=\| \leq c \ \forall j \notin I$, (and recall that $\mathcal{U} = \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c, \forall i\}$) applying Theorem 6 concludes the proof. □

[1] While we are not assuming generative models to establish the results, it is still interesting to see how these results can help in a generative model setup.

The next theorem gives conditions for when an optimal solution may have support on indices outside a given index subset I. Let $\mathbf{x}^*$ be an optimal solution that puts weight on features with indices in $I \cup J$ (take I and J disjoint). Then the residual with respect to I, $(\mathbf{b} - \sum_{i \in I'} x_i^* \mathbf{a}_i)$, and the contribution of the elements in J, namely, $\sum_{i \in J}$, must subtend a small angle.

Theorem 8 *Given a candidate solution $\mathbf{x}^*$, if there exists $I \subset \{1, \cdots, m\}$ such that for $j \notin I$, x_j^* are not all zero, and*

$$(\mathbf{b} - \sum_{i \in I} x_i^* \mathbf{a}_i)^\top (\sum_{j \notin I} x_j^* \mathbf{a}_j) < \frac{\|\mathbf{b} - \sum_{i \in I} x_i^* \mathbf{a}_i\|_2 \| \sum_{j \notin I} x_j^* \mathbf{a}_j\|_2}{\max_{t \notin I}(\|\mathbf{a}_t\|_2 / c_t)},$$

then $\mathbf{x}^*$ is not an optimal solution to the robust regression problem

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_2 \right\}.$$

We postpone the proof to the Appendix, but state two corollaries easily derived from the theorem.

Corollary 4 *Let $c_i = c$ and $\|\mathbf{a}_i\|_2 = 1$ for all i. If there exists $I \subset \{1, \cdots, m\}$ such that*

$$\begin{aligned} \mathbf{a}_i^\top \mathbf{a}_j &< c, \quad \forall i \in I, \forall j \notin I; \\ \mathbf{b}^\top \mathbf{a}_j &< c; \quad \forall j \notin I; \end{aligned}$$

then any optimal solution $\mathbf{x}^*$ satisfies

$$x_j^* = 0, \quad \forall j \notin I.$$

Corollary 5 *Suppose there exists $I \subseteq \{1, \cdots, m\}$, such that for all $j \notin I$, $\|\mathbf{a}_j\| < c_j$. Then, any optimal solution $\mathbf{x}^*$ satisfies $x_j^* = 0$, for $j \notin I$.*

The next theorem shows that sparsity is achieved when a set of features are "almost" linearly dependent. We postpone the proof to the Appendix.

Theorem 9 *Given $I \subseteq \{1, \cdots, m\}$, if there exists a non-zero vector $(w_i)_{i \in I}$ satisfying*

$$\| \sum_{i \in I} w_i \mathbf{a}_i \|_2 \leq \min_{\sigma_i \in \{-1, +1\}} | \sum_{i \in I} \sigma_i c_i w_i |,$$

then there exists an optimal solution $\mathbf{x}^*$ such that $\exists i \in I : x_i^* = 0$.

We provide an example illustrating the usefulness of Theorem 9 in establishing sparsity results.

Example 1 *Let $c_i = c$ for all i. Given any two-dimensional subspace $\mathcal{H}$, we can decompose each feature $\mathbf{a}_i = \mathbf{a}_i^{\mathcal{H}} + \mathbf{a}_i^{\perp}$ where $\mathbf{a}_i^{\mathcal{H}} \in \mathcal{H}$ and $\mathbf{a}_i^{\perp} \perp \mathcal{H}$. If for some $\alpha > 0$, k features (say, $\mathbf{a}_1, \ldots, \mathbf{a}_k$) satisfy the three properties below, then there exists an optimal solution that assigns nonzero weight to at most 2 of the k features.*

(i) $\|\mathbf{a}_i^{\mathcal{H}}\|_2 = 1; \quad i = 1, \cdots, k.$

(ii) $\alpha \leq \angle(\mathbf{a}_i^{\mathcal{H}}, \mathbf{a}_j^{\mathcal{H}}) \leq \pi - \alpha; \quad i \neq j, \ i, j = 1, \cdots, k.$

(iii) $\|\mathbf{a}_i^{\perp}\|_2 \leq c(\sec(\alpha) - 1)/3, \quad i = 1, \cdots, k.$

Proof. Consider an optimal solution which is sparsest. To derive a contradiction, assume that no fewer than three features (say, $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ w.o.l.g) are assigned non-zero weights. Since $\mathbf{a}_1^{\mathcal{H}}$, $\mathbf{a}_2^{\mathcal{H}}$ and $\mathbf{a}_3^{\mathcal{H}}$ belong to a 2-dimensional space $\mathcal{H}$, there exists a non-zero $\mathbf{w}$ such that $\sum_{i=1}^3 w_i \mathbf{a}_i^{\mathcal{H}} = \mathbf{0}$. Without loss of generality we let $|w_2|, |w_3| \leq |w_1| = 1$. Thus, we have

$$\min_{\sigma_i \in \{-1,+1\}} |\sum_{i=1}^3 \sigma_i c_i w_i| = c \min_{\sigma_i \in \{-1,+1\}} |\sum_{i=1}^3 \sigma_i w_i| = c\big||w_2| + |w_3| - 1\big|.$$

Next we bound $|w_2| + |w_3|$. We use superscript $\triangle$ to represent the projection of a vector on the direction of $\mathbf{a}_1$. Since $\sum_{i=1}^3 w_i \mathbf{a}_i = \mathbf{0}$ we have $(w_2\mathbf{a}_2 + w_2\mathbf{a}_3)^{\triangle} = -\mathbf{a}_1$. This leads to

$$\|(w_2\mathbf{a}_2 + w_2\mathbf{a}_3)^{\triangle}\|_2 = 1.$$

Further notice that by assumption (i) and (ii) we have

$$\|(w_i\mathbf{a}_i)^{\triangle}\|_2 \leq |w_i| \cos(\alpha), \quad i = 2, 3.$$

Thus, we have

$$|w_2| + |w_3| \geq 1/\cos(\alpha) = \sec(\alpha),$$

which leads to

$$\min_{\sigma_i \in \{-1,+1\}} |\sum_{i=1}^3 \sigma_i c_i w_i| \geq c(\sec(\alpha) - 1).$$

By (iii), we have

$$\begin{aligned} \|\sum_{i=1}^3 w_i \mathbf{a}_i\|_2 = \|\sum_{i=1}^3 w_i \mathbf{a}_i^{\mathcal{H}} + \sum_{i=1}^3 w_i \mathbf{a}_i^{\perp}\| = \|\sum_{i=1}^3 w_i \mathbf{a}_i^{\perp}\| \\ \leq (|w_1| + |w_2| + |w_3|)c(\sec(\alpha) - 1)/3 \leq c(\sec(\alpha) - 1). \end{aligned}$$

Hence the condition of Theorem 9 is satisfied, and we can construct a solution which is optimal and more sparse. This leads to the desired contradiction. □

Notice that for linearly dependent features, there exists non-zero $(w_i)_{i \in I}$ such that $\|\sum_{i \in I} w_i \mathbf{a}_i\|_2 = 0$, which leads to the following corollary.

Corollary 6 *Consider the robust regression problem*

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_2 \right\}.$$

Given $I \subseteq \{1, \cdots, m\}$, *denote* $A_I \triangleq \Big(\mathbf{a}_i\Big)_{i \in I}$, *and let* $t \triangleq \mathrm{rank}(A_I)$. *There exists an optimal solution* $\mathbf{x}^*$ *such that* $\mathbf{x}_I^* \triangleq (x_i)_{i \in I}^{\top}$ *has at most* t *non-zero coefficients.*

Setting $I = \{1, \cdots, m\}$, we immediately get the following corollary.

Corollary 7 *Consider the robust regression problem*

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \max_{\Delta A\in\mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_2 \right\}.$$

If $n < m$, then there exists an optimal solution with no more than n non-zero coefficients.

5 Density Estimation and Consistency

In this section, we investigate the robust linear regression formulation from a statistical perspective and rederive *using only robustness properties* that Lasso is asymptotically consistent. We note that our result applies to a considerably more general framework than Lasso. In [23] we use some intermediate results used to prove consistency to show that regularization can be identified with the so-called maxmin expected utility (MMEU) framework, thus tying regularization to a fundamental tenet of decision-theory.

We show that the robust optimization formulation can be seen to be the maximum error w.r.t. a class of probability measures. This class includes a kernel density estimator, and using this, we show that Lasso is consistent.

We restrict our discussion to the case where the magnitude of the allowable uncertainty for all features equals c, (i.e., the standard Lasso) and establish the statistical consistency of Lasso from a distributional robustness argument. Generalization to the non-uniform case is straightforward. Throughout, we use c_n to represent c where there are n samples (we take c_n to zero).

Recall the standard generative model in statistical learning: let $\mathbb{P}$ be a probability measure with bounded support that generates i.i.d samples $(b_i, \mathbf{r}_i)$, and has a density $f^*(\cdot)$. Denote the set of the first n samples by $\mathcal{S}_n$. Define

$$\mathbf{x}(c_n, \mathcal{S}_n) \triangleq \arg\min_{\mathbf{x}} \left\{ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - \mathbf{r}_i^\top \mathbf{x})^2} + c_n\|x\|_1 \right\}$$
$$= \arg\min_{\mathbf{x}} \left\{ \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^{n}(b_i - \mathbf{r}_i^\top \mathbf{x})^2} + c_n\|x\|_1 \right\};$$
$$\mathbf{x}(\mathbb{P}) \triangleq \arg\min_{\mathbf{x}} \left\{ \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x})^2 d\mathbb{P}(b, \mathbf{r})} \right\}.$$

In words, $\mathbf{x}(c_n, \mathcal{S}_n)$ is the solution to Lasso with the tradeoff parameter set to $c_n\sqrt{n}$, and $\mathbf{x}(\mathbb{P})$ is the "true" optimal solution. We have the following consistency result. The theorem itself is a well-known result. However, the proof technique is novel. This technique is of interest because the standard techniques to establish consistency in statistical learning including VC dimension and algorithm stability often work for a limited range of algorithms, e.g., SVMs are known to have infinite VC dimension, and we show in Section 6 that *Lasso is not stable.* In contrast, a much wider range of algorithms have robustness interpretations, allowing a unified approach to prove their consistency.

Theorem 10 *Let $\{c_n\}$ be such that $c_n \downarrow 0$ and $\lim_{n\to\infty} n(c_n)^{m+1} = \infty$. Suppose there exists a constant H such that $\|\mathbf{x}(c_n, \mathcal{S}_n)\|_2 \leq H$ almost surely. Then,*

$$\lim_{n\to\infty} \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mathbb{P}(b, \mathbf{r})} = \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(\mathbb{P}))^2 d\mathbb{P}(b, \mathbf{r})},$$

almost surely.

We provide the main ideas and outline here, after which we give the proof. The proof of intermediate results outlined in the steps below are postponed to the Appendix. The key to the proof is establishing a connection between robustness and kernel density estimation.

Step 1: For a given $\mathbf{x}$, we show that the robust regression loss over the training data is equal to the worst-case expected *generalization error*. To show this we establish a more general result:

Proposition 2 *Given a function $g : \mathbb{R}^{m+1} \to \mathbb{R}$ and Borel sets $\mathcal{Z}_1, \cdots, \mathcal{Z}_n \subseteq \mathbb{R}^{m+1}$, let*

$$\mathcal{P}_n \triangleq \{\mu \in \mathcal{P} | \forall S \subseteq \{1, \cdots, n\} : \mu(\bigcup_{i\in S} \mathcal{Z}_i) \geq |S|/n\}.$$

The following holds

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{(\mathbf{r}_i, b_i)\in\mathcal{Z}_i} h(\mathbf{r}_i, b_i) = \sup_{\mu\in\mathcal{P}_n} \int_{\mathbb{R}^{m+1}} h(\mathbf{r}, b) d\mu(\mathbf{r}, b).$$

We also have the following corollary, which we use below to interpret Lasso from a density estimation perspective, and to prove Theorem 10.

Corollary 8 *Given $\mathbf{b} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$, the following equation holds for any $\mathbf{x} \in \mathbb{R}^m$,*

$$\|\mathbf{b} - A\mathbf{x}\|_2 + \sqrt{n}c + \sqrt{n} \sum_{i=1}^{m} c_i |x_i| = \sup_{\mu\in\hat{\mathcal{P}}(n)} \sqrt{n \int_{\mathbb{R}^{m+1}} (b' - \mathbf{r}'^\top \mathbf{x})^2 d\mu(\mathbf{r}', b')}. \tag{9}$$

Here,[2]

$$\hat{\mathcal{P}}(n) \triangleq \bigcup_{\Delta | \forall j, \sum_j \delta_{ij}^2 = nc_j^2} \mathcal{P}_n(A, \Delta, \mathbf{b}, c);$$

$$\mathcal{P}_n(A, \Delta, \mathbf{b}, c) \triangleq \{\mu \in \mathcal{P} | \mathcal{Z}_i = [b_i - c, b_i + c] \times \prod_{j=1}^{m} [a_{ij} - \delta_{ij}, a_{ij} + \delta_{ij}];$$
$$\forall S \subseteq \{1, \cdots, n\} : \mu(\bigcup_{i\in S} \mathcal{Z}_i) \geq |S|/n\},$$

Remark 2 We briefly explain Corollary 8 to avoid possible confusions before we proceed to the proof. Equation (9) is a non-probabilistic equality. That is, it holds without any assumption (e.g., i.i.d. or generated by certain distributions) on $\mathbf{b}$ and A. And it does not

[2] Recall that a_{ij} is the j^{th} element of $\mathbf{r}_i$.

involve any probabilistic operation such as taking expectation on the left-hand-side, instead, it is an equivalence relationship which hold for an arbitrary set of samples. Notice that, the right-hand-side also depends on the samples since $\hat{\mathcal{P}}(n)$ is defined through A and $\mathbf{b}$. Indeed, $\hat{\mathcal{P}}(n)$ represents the union of classes of distributions $\mathcal{P}(A,\Delta,\mathbf{b},c)$ such that the norm of each column of Δ is bounded, where $\mathcal{P}(A,\Delta,\mathbf{b},c)$ is the set of distributions corresponds to (see Proposition 2) disturbance in hyper-rectangle Borel sets $\mathcal{Z}_1,\cdots,\mathcal{Z}_n$ centered at $(b_i,\mathbf{r}_i^\top)$ with lengths $(2c,2\delta_{i1},\cdots,2\delta_{im})$.

Proof. The right-hand-side of Equation (8) equals to

$$\sup_{\Delta|\forall j,\sum_j \delta_{ij}^2=nc_j^2}\Big\{\sup_{\mu\in\mathcal{P}_n(A,\Delta,\mathbf{b},c)}\sqrt{n\int_{\mathbb{R}^{m+1}}(b'-\mathbf{r}'^\top\mathbf{x})^2d\mu(\mathbf{r}',b')}\Big\}.$$

Notice the left-hand-side equals to

$$\begin{aligned}&\max_{\|\delta\mathbf{b}\|\le\sqrt{n}c,\|\mathbf{a}_j\|_2\le\sqrt{n}c_j}\|\mathbf{b}+\delta\mathbf{b}-(A+\Delta A)\mathbf{x}\|_2\\=&\sup_{\Delta|\forall j,\sum_j \delta_{ij}^2=nc_j^2}\left\{\sup_{(\hat{b}_i,\hat{\mathbf{r}}_i)\in[b_i-c,b_i+c]\times\prod_{j=1}^m[a_{ij}-\delta_{ij},a_{ij}+\delta_{ij}]}\sqrt{\sum_{i=1}^n(\hat{b}_i-\hat{\mathbf{r}}_i^\top\mathbf{x})^2}\right\}\\=&\sup_{\Delta|\forall j,\sum_j \delta_{ij}^2=nc_j^2}\sqrt{\sum_{i=1}^n\sup_{(\hat{b}_i,\hat{\mathbf{r}}_i)\in[b_i-c,b_i+c]\times\prod_{j=1}^m[a_{ij}-\delta_{ij},a_{ij}+\delta_{ij}]}(\hat{b}_i-\hat{\mathbf{r}}_i^\top\mathbf{x})^2},\end{aligned}$$

furthermore, applying Proposition 2 yields

$$\begin{aligned}&\sqrt{\sum_{i=1}^n\sup_{(\hat{b}_i,\hat{\mathbf{r}}_i)\in[b_i-c,b_i+c]\times\prod_{j=1}^m[a_{ij}-\delta_{ij},a_{ij}+\delta_{ij}]}(\hat{b}_i-\hat{\mathbf{r}}_i^\top\mathbf{x})^2}\\=&\sqrt{\sup_{\mu\in\mathcal{P}_n(A,\Delta,\mathbf{b},c)}n\int_{\mathbb{R}^{m+1}}(b'-\mathbf{r}'^\top\mathbf{x})^2d\mu(\mathbf{r}',b')}\\=&\sup_{\mu\in\mathcal{P}_n(A,\Delta,\mathbf{b},c)}\sqrt{n\int_{\mathbb{R}^{m+1}}(b'-\mathbf{r}'^\top\mathbf{x})^2d\mu(\mathbf{r}',b')},\end{aligned}$$

which proves the corollary. □

Step 2: Next we show that robust regression has a form like that in the left hand side above. Also, the set of distributions we supremize over, in the right hand side above, includes a kernel density estimator for the true (unknown) distribution.

The *kernel density estimator* for a density $\hat{f}$ in $\mathbb{R}^d$, originally proposed in [24,25], is defined by

$$f_n(\mathbf{x})=(nc_n^d)^{-1}\sum_{i=1}^n K\left(\frac{\mathbf{x}-\hat{\mathbf{x}}_i}{c_n}\right),$$

where $\{c_n\}$ is a sequence of positive numbers, $\hat{\mathbf{x}}_i$ are i.i.d. samples generated according to $\hat{f}$, and K is a Borel measurable function (kernel) satisfying $K\ge 0$, $\int K=1$. See [26, 27] and

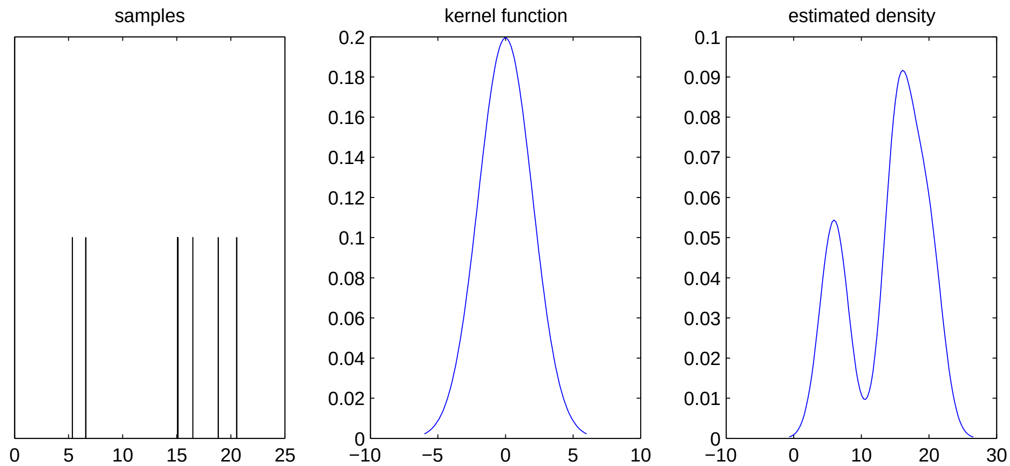

Figure 1: Illustration of Kernel Density Estimation

the reference therein for detailed discussions. Figure 1 illustrates a kernel density estimator using Gaussian kernel for a randomly generated sample-set.

Now consider the following kernel estimator given samples $(b_i, \mathbf{r}_i)_{i=1}^n$,

$$\begin{aligned} & h_n(b, \mathbf{r}) \triangleq (nc^{m+1})^{-1} \sum_{i=1}^n K\left(\frac{b - b_i, \mathbf{r} - \mathbf{r}_i}{c}\right), \\ & \text{where: } K(\mathbf{x}) \triangleq I_{[-1,+1]^{m+1}}(\mathbf{x})/2^{m+1}. \end{aligned} \tag{10}$$

Observe that the estimated distribution given by Equation (10) belongs to the set of distributions

$$\mathcal{P}_n(A, \Delta, \mathbf{b}, c) \triangleq \{\mu \in \mathcal{P} | \mathcal{Z}_i = [b_i - c, b_i + c] \times \prod_{j=1}^m [a_{ij} - \delta_{ij}, a_{ij} + \delta_{ij}];$$
$$\forall S \subseteq \{1, \cdots, n\} : \mu(\bigcup_{i \in S} \mathcal{Z}_i) \geq |S|/n\},$$

and hence belongs to $\hat{\mathcal{P}}(n) = \hat{\mathcal{P}}(n) \triangleq \bigcup_{\Delta | \forall j, \sum_i \delta_{ij}^2 = nc_j^2} \mathcal{P}_n(A, \Delta, \mathbf{b}, c)$, which is precisely the set of distributions used in the representation from Proposition 2.

Step 3: Combining the last two steps, and using the fact that $\int_{b,\mathbf{r}} |h_n(b, \mathbf{r}) - h(b, \mathbf{r})| d(b, \mathbf{r})$ goes to zero almost surely when $c_n \downarrow 0$ and $nc_n^{m+1} \uparrow \infty$ since $h_n(\cdot)$ is a kernel density estimation of $f(\cdot)$ (see e.g. Theorem 3.1 of [26]), we prove consistency of robust regression.

We can remove the assumption that $\|\mathbf{x}(c_n, \mathcal{S}_n)\|_2 \leq H$, and as in Theorem 10, the proof technique rather than the result itself is of interest.

Theorem 11 *Let $\{c_n\}$ converge to zero sufficiently slowly. Then*

$$\lim_{n \to \infty} \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mathbb{P}(b, \mathbf{r})} = \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(\mathbb{P}))^2 d\mathbb{P}(b, \mathbf{r})},$$

almost surely.

Using the results above, we now prove Theorem 10, and then Theorem 11.

Proof. Let $\hat{\mu}_n$ be the estimated distribution using Equation (10) given $\mathcal{S}_n$ and c_n, and denote its density function $f_n(\cdot)$. Notice that, $\|\mathbf{x}(c_n, \mathcal{S}_n)\|_2 \leq H$ almost surely and $\mathbb{P}$ has a bounded support implies that there exists a universal constant C such that

$$\max_{b,\mathbf{r}}(b - \mathbf{r}^\top \mathbf{w}(c_n, \mathcal{S}_n))^2 \leq C,$$

almost surely.

By Corollary 8 and $\hat{\mu}_n \in \hat{\mathcal{P}}(n)$ we have

$$\begin{aligned}
&\sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\hat{\mu}_n(b, \mathbf{r})} \\
\leq & \sup_{\mu \in \hat{\mathcal{P}}(n)} \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mu(b, \mathbf{r})} \\
= & \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \mathbf{x}(c_n, \mathcal{S}_n))^2} + c_n \|\mathbf{x}(c_n, \mathcal{S}_n)\|_1 + c_n \\
\leq & \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \mathbf{x}(\mathbb{P}))^2} + c_n \|\mathbf{x}(\mathbb{P})\|_1 + c_n,
\end{aligned}$$

the last inequality holds by definition of $\mathbf{x}(c_n, \mathcal{S}_n)$.

Taking the square of both sides, we have

$$\begin{aligned}
&\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\hat{\mu}_n(b, \mathbf{r}) \\
\leq & \frac{1}{n} \sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \mathbf{x}(\mathbb{P}))^2 + c_n^2 (1 + \|\mathbf{x}(\mathbb{P})\|_1)^2 + 2c_n (1 + \|\mathbf{x}(\mathbb{P})\|_1) \sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \mathbf{x}(\mathbb{P}))^2}
\end{aligned}$$

Notice that, the right-hand side converges to $\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(\mathbb{P}))^2 d\mathbb{P}(b, \mathbf{r})$ as $n \uparrow \infty$ and $c_n \downarrow 0$ almost surely. Furthermore, we have

$$\begin{aligned}
&\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mathbb{P}(b, \mathbf{r}) \\
\leq & \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\hat{\mu}_n(b, \mathbf{r}) + \max_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 \times \int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \\
\leq & \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\hat{\mu}_n(b, \mathbf{r}) + C \int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}),
\end{aligned}$$

where the last inequality follows from the definition of C. Notice that $\int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r})$ goes to zero almost surely when $c_n \downarrow 0$ and $nc_n^{m+1} \uparrow \infty$ since $f_n(\cdot)$ is a kernel density estimation of $f(\cdot)$ (see e.g. Theorem 3.1 of [26]). Hence the theorem follows. □

Next we prove Theorem 11.

Proof. To prove the theorem, we need to consider a set of distributions belonging to $\hat{\mathcal{P}}(n)$. Hence we establish the following lemma first.

Lemma 1 *Partition the support of $\mathbb{P}$ as $V_1, \cdots, V_T$ such the ℓ^∞ radius of each set is less than c_n. If a distribution μ satisfies*

$$\mu(V_t) = \#\big((b_i, \mathbf{r}_i) \in V_t\big)/n; \quad t = 1, \cdots, T, \tag{11}$$

then $\mu \in \hat{\mathcal{P}}(n)$.

Proof. Let $\mathcal{Z}_i = [b_i - c_n, b_i + c_n] \times \prod_{j=1}^m [a_{ij} - c_n, a_{ij} + c_n]$; recall that a_{ij} the j^{th} element of $\mathbf{r}_i$. Notice V_t has ℓ^∞ norm less than c_n we have

$$(b_i, \mathbf{r}_i \in V_t) \Rightarrow V_t \subseteq \mathcal{Z}_i.$$

Therefore, for any $S \subseteq \{1, \cdots, n\}$, the following holds

$$\begin{aligned}
&\mu(\bigcup_{i\in S} \mathcal{Z}_i) \geq \mu(\bigcup V_t | \exists i \in S : b_i, \mathbf{r}_i \in V_t) \\
= &\sum_{t|\exists i \in S: b_i, \mathbf{r}_i \in V_t} \mu(V_t) = \sum_{t|\exists i \in S: b_i, \mathbf{r}_i \in V_t} \#\big((b_i, \mathbf{r}_i) \in V_t\big)/n \geq |S|/n.
\end{aligned}$$

Hence $\mu \in \mathcal{P}_n(A, \Delta, b, c_n)$ where each element of Δ is c_n, which leads to $\mu \in \hat{\mathcal{P}}(n)$. □

Now we proceed to prove the theorem. Partition the support of $\mathbb{P}$ into T subsets such that ℓ^∞ radius of each one is smaller than c_n. Denote $\tilde{\mathcal{P}}(n)$ as the set of probability measures satisfying Equation (11). Hence $\tilde{\mathcal{P}}(n) \subseteq \hat{\mathcal{P}}(n)$ by Lemma 1. Further notice that there exists a universal constant K such that $\|\mathbf{x}(c_n, \mathcal{S}_n)\|_2 \leq K/c_n$ due to the fact that the square loss of the solution $\mathbf{x} = \mathbf{0}$ is bounded by a constant only depends on the support of $\mathbb{P}$. Thus, there exists a constant C such that $\max_{b,\mathbf{r}}(b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 \leq C/c_n^2$.

Follow a similar argument as the proof of Theorem 10, we have

$$\begin{aligned}
&\sup_{\mu_n \in \tilde{\mathcal{P}}(n)} \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mu_n(b, \mathbf{r}) \\
\leq &\frac{1}{n}\sum_{i=1}^n (b_i - \mathbf{r}_i^\top \mathbf{x}(\mathbb{P}))^2 + c_n^2(1 + \|\mathbf{x}(\mathbb{P})\|_1)^2 + 2c_n(1 + \|\mathbf{x}(\mathbb{P})\|_1)\sqrt{\frac{1}{n}\sum_{i=1}^n (b_i - \mathbf{r}_i^\top \mathbf{x}(\mathbb{P}))^2},
\end{aligned} \tag{12}$$

and

$$\begin{aligned}
&\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mathbb{P}(b, \mathbf{r}) \\
\leq &\inf_{\mu_n \in \tilde{\mathcal{P}}(n)} \Big\{ \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mu_n(b, \mathbf{r}) \\
&\qquad + \max_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 \int_{b,\mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \\
\leq &\sup_{\mu_n \in \tilde{\mathcal{P}}(n)} \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 d\mu_n(b, \mathbf{r}) + 2C/c_n^2 \inf_{\mu'_n \in \tilde{\mathcal{P}}(n)} \Big\{ \int_{b,\mathbf{r}} |f_{\mu'_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \Big\},
\end{aligned}$$

here f_μ stands for the density function of a measure μ. Notice that $\tilde{\mathcal{P}}_n$ is the set of distributions satisfying Equation (11), hence $\inf_{\mu'_n\in\tilde{\mathcal{P}}(n)}\int_{b,\mathbf{r}}|f_{\mu'_n}(b,\mathbf{r})-f(b,\mathbf{r})|d(b,\mathbf{r})$ is upper-bounded by $\sum_{t=1}^T|\mathbb{P}(V_t)-\#(b_i,\mathbf{r}_i\in V_t)|/n$, which goes to zero as n increases for any fixed c_n (see for example Proposition A6.6 of [28]). Therefore,

$$2C/c_n^2\inf_{\mu'_n\in\tilde{\mathcal{P}}(n)}\Big\{\int_{b,\mathbf{r}}|f_{\mu'_n}(b,\mathbf{r})-f(b,\mathbf{r})|d(b,\mathbf{r})\Big\}\to 0,$$

if $c_n\downarrow 0$ sufficiently slow. Combining this with Inequality (12) proves the theorem. □

6 Stability

Knowing that the robust regression problem (1) and in particular Lasso encourage sparsity, it is of interest to investigate another desirable characteristic of a learning algorithm, namely, stability. We show in this section that Lasso *is not stable*. This is a special case of a more general result we prove in [**?**], where we show that this is a common property for all algorithms that encourage sparsity. That is, if a learning algorithm achieves certain sparsity condition, then it cannot have a non-trivial stability bound.

We recall the definition of uniform stability bound [29] first. We let $\mathcal{Z}$ denote the space of points and labels (typically this will be a compact subset of $\mathbb{R}^{n+1}$) so that $S\in\mathcal{Z}^m$ denotes a collection of m labelled training points. We let $\mathbb{L}$ denote a learning algorithm, and for $S\in\mathcal{Z}^m$, we let $\mathbb{L}_S$ denote the output of the learning algorithm (i.e., the regression function it has learned from the training data). Then given a loss function l, and a labelled point $s=(\mathbf{z},b)\in\mathcal{Z}$, $l(\mathbb{L}_S,s)$ denotes the loss of the algorithm that has been trained on the set S, on the data point s. Thus for squared loss, we would have $l(\mathbb{L}_S,s)=\|\mathbb{L}_S(\mathbf{z})-b\|_2$.

Definition 1 *An algorithm $\mathbb{L}$ has uniform stability β_m with respect to the loss function l if the following holds*

$$\forall S\in\mathcal{Z}^m,\forall i\in\{1,\cdots,m\},\|l(\mathbb{L}_S,\cdot)-l(\mathbb{L}_{S^{\setminus i}},\cdot)\|_\infty\le\beta_m.$$

Here $\mathbb{L}_{S^{\setminus i}}$ stands for the learned solution with the i^{th} sample removed from S.

At first glance, this definition may seem too stringent for any reasonable algorithm to exhibit good stability properties. However, as shown in [29], *Tikhonov-regularized regression has stability that scales as* $1/m$. Stability that scales at least as fast as $o(\frac{1}{\sqrt{m}})$ can be used to establish strong PAC bounds.

In this section we show that not only is the stability (in the sense defined above) of Lasso much worse than the stability of ℓ^2-regularized regression, but in fact Lasso's stability is, in the following sense, as bad as it gets. To this end, we define the notion of the trivial bound, which is the worst possible error a training algorithm can have for arbitrary training set and testing sample labelled by zero.

Definition 2 *Given a subset from which we can draw m labelled points, $\mathcal{Z}\subseteq\mathbb{R}^{n\times(m+1)}$ and a subset for one unlabelled point, $\mathcal{X}\subseteq\mathbb{R}^m$, a trivial bound for a learning algorithm $\mathbb{L}$ w.r.t. $\mathcal{Z}$ and $\mathcal{X}$ is*

$$\mathfrak{b}(\mathbb{L},\mathcal{Z},\mathcal{X})\triangleq\max_{S\in\mathcal{Z},\mathbf{z}\in\mathcal{X}}l\big(\mathbb{L}_S,(\mathbf{z},0)\big).$$

As above, $l(\cdot,\cdot)$ is a given loss function.

Notice that the trivial bound does not diminish as the number of samples increases, since by repeatedly choosing the worst sample, the algorithm will yield the same solution.

Now we show that the uniform stability bound of Lasso can be no better than its trivial bound with the number of features halved.

Theorem 12 *Given $\hat{\mathcal{Z}} \subseteq \mathbb{R}^{n\times(2m+1)}$ be the domain of sample set and $\hat{\mathcal{X}} \subseteq \mathbb{R}^{2m}$ be the domain of new observation, such that*

$$(\mathbf{b}, A) \in \mathcal{Z} \Longrightarrow (\mathbf{b}, A, A) \in \hat{\mathcal{X}}$$
$$(\mathbf{z}^\top) \in \mathcal{X} \Longrightarrow (\mathbf{z}^\top, \mathbf{z}^\top) \in \hat{\mathcal{X}},$$

we have the uniform stability bound β of Lasso is lower bounded by $\mathfrak{b}(\text{Lasso}, \mathcal{Z}, \mathcal{X})$.

Proof. Let $(\mathbf{b}^*, A^*)$ and $(0, \mathbf{z}^{*\top})$ be the sample set and the new observation such that they jointly achieve $\mathfrak{b}(\text{Lasso}, \mathcal{Z}, \mathcal{X})$, and let $\mathbf{x}^*$ be the optimal solution to Lasso w.r.t $(\mathbf{b}^*, A^*)$. Consider the following sample set

$$\begin{pmatrix} \mathbf{b}^* & A^* & A^* \\ 0 & \mathbf{0}^\top & \mathbf{z}^{*\top} \end{pmatrix}.$$

Observe that $(\mathbf{x}^\top, \mathbf{0}^\top)^\top$ is an optimal solution of Lasso w.r.t to this sample set. Now remove the last sample from the sample set. Notice that $(\mathbf{0}^\top, \mathbf{x}^\top)^\top$ is an optimal solution for this new sample set. Using the last sample as a testing observation, the solution w.r.t the full sample set has zero cost, while the solution of the leave-one-out sample set has a cost $\mathfrak{b}(\text{Lasso}, \mathcal{Z}, \mathcal{X})$. And hence we prove the theorem. □

7 Conclusion

In this paper, we considered robust regression with a least-square-error loss. In contrast to previous work on robust regression, we considered the case where the perturbations of the observations are in the features. We show that this formulation is equivalent to a weighted ℓ^1 norm regularized regression problem if no correlation of disturbances among different features is allowed, and hence provide an interpretation of the widely used Lasso algorithm from a robustness perspective. We also formulated tractable robust regression problems for disturbance correlated among different features, and investigated the empirical performance of a class of such formulations which interpolate between Lasso and ℓ^∞ norm regularized regression.

The sparsity of the resulting formulation is also investigated, and in particular we present a "no-free-lunch" theorem saying that sparsity and algorithmic stability contradict each other. This result shows, although sparsity and algorithmic stability are both regarded as desirable properties of regression algorithms, it is not possible to achieve them simultaneously, and we have to tradeoff these two properties in designing a regression algorithm.

The main thrust of this work is to treat the widely used regularized regression scheme from a robust optimization perspective, and extend the result of [13] (i.e., Tikhonov regularization is equivalent to a robust formulation for Frobenius norm bounded disturbance set) to a broader range of disturbance set and hence regularization scheme. This provides us not only with new insight of why regularization schemes work, but also offer solid motivations for selecting regularization parameter for existing regularization scheme and facilitate designing new regularizing schemes.

A Proof of Theorem 2

Theorem 2 *Consider a random vector* $\mathbf{v} \in \mathbb{R}^n$*, such that* $\mathbb{E}(\mathbf{v}) = \mathbf{a}$*, and* $\mathbb{E}(\mathbf{v}\mathbf{v}^\top) = \Sigma$*,* $\Sigma \succeq 0$*. Then we have*

$$\Pr\{\|\mathbf{v}\|_2 \geq c_i\} \leq \begin{cases} \min_{P,\mathbf{q},r,\lambda} & \mathrm{Trace}(\Sigma P) + 2\mathbf{q}^\top \mathbf{a} + r \\ \textit{subject to:} & \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^\top & r \end{pmatrix} \succeq 0 \\ & \begin{pmatrix} I(m) & \mathbf{0} \\ \mathbf{0}^\top & -c_i^2 \end{pmatrix} \preceq \lambda \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^\top & r-1 \end{pmatrix} \\ & \lambda \geq 0. \end{cases} \tag{13}$$

Proof. Consider a function $f(\cdot)$ parameterized by $P, \mathbf{q}, r$ defined as $f(\mathbf{v}) = \mathbf{v}^\top P \mathbf{v} + 2\mathbf{q}^\top \mathbf{v} + r$. Notice $\mathbb{E}\big(f(\mathbf{v})\big) = \mathrm{Trace}(\Sigma P) + 2\mathbf{q}^\top \mathbf{a} + r$. Now we show that $f(\mathbf{v}) \geq \mathbf{1}_{\|\mathbf{v}\| \geq c_i}$ for all $P, \mathbf{q}, r$ satisfying the constraints in (13).

To show $f(\mathbf{v}) \geq \mathbf{1}_{\|\mathbf{v}\|_2 \geq c_i}$, we need to establish (i) $f(\mathbf{v}) \geq 0$ for all $\mathbf{v}$, and (ii) $f(\mathbf{v}) \geq 1$ when $\|\mathbf{v}\|_2 \geq c_i$. Notice that

$$f(\mathbf{v}) = \begin{pmatrix} \mathbf{v} \\ 1 \end{pmatrix}^\top \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^\top & r \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ 1 \end{pmatrix},$$

hence (i) holds because

$$\begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^\top & r \end{pmatrix} \succeq 0.$$

To establish condition (ii), it suffices to show $\mathbf{v}^\top \mathbf{v} \geq c_i^2$ implies $\mathbf{v}^\top P \mathbf{v} + 2\mathbf{q}^\top \mathbf{v} + r \geq 1$, which is equivalent to show $\big\{\mathbf{v} \big| \mathbf{v}^\top P \mathbf{v} + 2\mathbf{q}^\top \mathbf{v} + r - 1 \leq 0\big\} \subseteq \big\{\mathbf{v} \big| \mathbf{v}^\top \mathbf{v} \leq c_i^2\big\}$. Noticing this is an ellipsoid-containment condition, by S-Procedure, we see that is equivalent to the condition that there exists a $\lambda \geq 0$ such that

$$\begin{pmatrix} I(m) & \mathbf{0} \\ \mathbf{0}^\top & -c_i^2 \end{pmatrix} \preceq \lambda \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^\top & r-1 \end{pmatrix}.$$

Hence we have $f(\mathbf{v}) \geq \mathbf{1}_{\|\mathbf{v}\|_2 \geq c_i}$, taking expectation over both side that notice that the expectation of a indicator function is the probability, we establish the theorem. □

B Proof of Theorem 4

Theorem 4 *Assume that the set*

$$\mathcal{Z} \triangleq \{\mathbf{z} \in \mathbb{R}^m | f_j(\mathbf{z}) \leq 0,\ j = 1, \cdots, k;\ \mathbf{z} \geq \mathbf{0}\}$$

has non-empty relative interior. Then the robust regression problem

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}'} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_a \right\}$$

is equivalent to the following regularized regression problem

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^k,\boldsymbol{\kappa}\in\mathbb{R}_+^m,\mathbf{x}\in\mathbb{R}^m}\Big\{\|\mathbf{b}-A\mathbf{x}\|_a+v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})\Big\};$$
$$\textit{where: } v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})\triangleq\max_{\mathbf{c}\in\mathbf{R}^m}\Big[(\boldsymbol{\kappa}+|\mathbf{x}|)^\top\mathbf{c}-\sum_{j=1}^k\lambda_j f_j(\mathbf{c})\Big] \tag{14}$$

Proof. Fix a solution $\mathbf{x}^*$. Notice that,

$$\mathcal{U}'=\{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)|\mathbf{c}\in\mathcal{Z};\ \|\boldsymbol{\delta}_i\|_a\le c_i,\ i=1,\cdots,m\}.$$

Hence we have:

$$\begin{aligned}&\max_{\Delta A\in\mathcal{U}'}\|\mathbf{b}-(A+\Delta A)\mathbf{x}^*\|_a\\ =&\max_{\mathbf{c}\in\mathcal{Z}}\Big\{\max_{\|\boldsymbol{\delta}_i\|_a\le c_i,\,i=1,\cdots,m}\|\mathbf{b}-\big(A+(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\big)\mathbf{x}^*\|_a\Big\}\\ =&\max_{\mathbf{c}\in\mathcal{Z}}\Big\{\|\mathbf{b}-A\mathbf{x}^*\|_a+\sum_{i=1}^m c_i|x_i^*|\Big\}\\ =&\|\mathbf{b}-A\mathbf{x}^*\|_a+\max_{\mathbf{c}\in\mathcal{Z}}\Big\{|\mathbf{x}^*|^\top\mathbf{c}\Big\}.\end{aligned} \tag{15}$$

The second equation follows from Theorem 3.

Now we need to evaluate $\max_{\mathbf{c}\in\mathcal{Z}}\{|\mathbf{x}^*|^\top\mathbf{c}\}$, which equals to $-\min_{\mathbf{c}\in\mathcal{Z}}\{-|\mathbf{x}^*|^\top\mathbf{c}\}$. Hence we are minimizing a linear function over a set of convex constraints. Furthermore, by assumption the Slater's condition holds. Hence the duality gap of $\min_{\mathbf{c}\in\mathcal{Z}}\{-|\mathbf{x}^*|^\top\mathbf{c}\}$ is zero. A standard duality analysis shows that

$$\max_{\mathbf{c}\in\mathcal{Z}}\Big\{|\mathbf{x}^*|^\top\mathbf{c}\Big\}=\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^k,\boldsymbol{\kappa}\in\mathbb{R}_+^m}v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x}^*). \tag{16}$$

We establish the theorem by substituting Equation (16) back into Equation (15) and taking minimum over $\mathbf{x}$ on both sides. □

C Proof of Theorem 5

Theorem 5 *Denote the j^{th} row of A as $\mathbf{r}_j^\top$. The robust regression problem (8) is equivalent to the following Second Order Cone Program on $(\mathbf{x},\mathbf{z},\mathbf{t},w)$:*

$$\begin{aligned}\textit{Minimize:}\quad & w+\sum_{i=1}^m c_i z_i\\ \textit{Subject to:}\quad & \mathbf{x}\le\mathbf{z};\\ & -\mathbf{x}\le\mathbf{z}\\ & \|\Sigma_j^{1/2}\mathbf{x}\|_2\le t_j-b_j+\mathbf{r}_j^\top\mathbf{x};\quad j=1,\cdots,n.\\ & \|\Sigma_j^{1/2}\mathbf{x}\|_2\le t_j+b_j-\mathbf{r}_j^\top\mathbf{x};\quad j=1,\cdots,n.\\ & \|\mathbf{t}\|_2\le w.\end{aligned}$$

Proof. To prove the theorem, it suffices to show that for given $\mathbf{x}$, the following holds

$$\max_{\Delta A_1\in\mathcal{U}_1,\Delta A_2\in\mathcal{U}_2}\|\mathbf{b}-(A+\Delta A_1+\Delta A_2)\mathbf{x}\|_2=\sqrt{\sum_{j=1}^{n}\left(|b_j-\mathbf{r}_j^\top\mathbf{x}|+\|\Sigma_j^{1/2}\mathbf{x}\|_2\right)^2}+\sum_{i=1}^{m}c_i|x_i|.$$

Notice that

$$\begin{aligned}&\max_{\Delta A_1\in\mathcal{U}_1,\Delta A_2\in\mathcal{U}_2}\|\mathbf{b}-(A+\Delta A_1+\Delta A_2)\mathbf{x}\|_2\\=&\max_{\Delta A_1\in\mathcal{U}_1}\{\max_{\Delta A_2\in\mathcal{U}_2}\|\mathbf{b}-(A+\Delta A_1+\Delta A_2)\mathbf{x}\|_2\}\\=&\max_{\Delta A_1\in\mathcal{U}_1}\{\|\mathbf{b}-(A+\Delta A_1)\mathbf{x}\|_2+\sum_{i=1}^{m}c_i|x_i|\}\\=&\max_{\Delta A_1\in\mathcal{U}_1}(\|\mathbf{b}-(A+\Delta A_1)\mathbf{x}\|_2)+\sum_{i=1}^{m}c_i|x_i|.\end{aligned}$$

Furthermore, the following equation proves the theorem.

$$\begin{aligned}&\max_{\Delta A_1\in\mathcal{U}_1}(\|\mathbf{b}-(A+\Delta A_1)\mathbf{x}\|_2)\\=&\sqrt{\sum_{j=1}^{n}\max_{\mathbf{l}_j\Sigma_j^{-1}\mathbf{l}_j\le 1}(b_j-\mathbf{r}_j^\top\mathbf{x}-\mathbf{l}_j^\top\mathbf{x})^2}\\=&\sqrt{\sum_{j=1}^{n}\left(|b_j-\mathbf{r}_j^\top\mathbf{x}|+\|\Sigma_j^{1/2}\mathbf{x}\|_2\right)^2}.\end{aligned}$$

The last equality holds because

$$\min_{\mathbf{l}_j\Sigma_j^{-1}\mathbf{l}_j}\mathbf{l}_j^\top\mathbf{x}=-\|\Sigma_j^{1/2}\mathbf{x}\|_2;\quad\&\max_{\mathbf{l}_j\Sigma_j^{-1}\mathbf{l}_j}\mathbf{l}_j^\top\mathbf{x}=\|\Sigma_j^{1/2}\mathbf{x}\|_2.$$

□

D Proof of Theorem 8

Theorem 8 *Given a candidate solution* $\mathbf{x}^*$*, if there exists* $I\subset\{1,\cdots,m\}$ *such that for* $j\notin I$*,* x_j^* *are not all zero, and*

$$(\mathbf{b}-\sum_{i\in I}x_i^*\mathbf{a}_i)^\top(\sum_{j\notin I}x_j^*\mathbf{a}_j)<\frac{\|\mathbf{b}-\sum_{i\in I}x_i^*\mathbf{a}_i\|_2\|\sum_{j\notin I}x_j^*\mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)},$$

then $\mathbf{x}^*$ *is not an optimal solution to the robust regression problem*

$$\min_{\mathbf{x}\in\mathbb{R}^m}\left\{\max_{\Delta A\in\mathcal{U}}\|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_2\right\}.$$

Proof. Consider the following candidate solution $\hat{\mathbf{x}}^*$ defined as

$$\hat{x}_i^* \triangleq \begin{cases} x_i, & i \in I; \\ 0, & i \notin I. \end{cases}$$

Using the regularized form of the robust regression problem, we have

$$\begin{aligned} &\|\mathbf{b} - A\hat{\mathbf{x}}^*\|_2 + \sum_i |\hat{x}_i^*| \\ =&\|\mathbf{b} - \sum_{i\in I} x_i^* \mathbf{a}_i\|_2 - \frac{\|\sum_{j\notin I} x_j^* \mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)} + \sum_{i\in I} |x_i^*| + \frac{\|\sum_{j\notin I} x_j^* \mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)}. \end{aligned} \tag{17}$$

Notice that,

$$\begin{aligned} &\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t) \cdot \sum_{j\notin I} c_j |x_j^*| \\ \geq& \sum_{j\notin I} (\|\mathbf{a}_j\|_2/c_j \cdot c_j |x_j^*|) \\ =& \sum_{j\notin I} (\|\mathbf{a}_j\|_2 |x_j^*|) \\ \geq& \|\sum_{j\notin I} x_j^* \mathbf{a}_j\|_2, \end{aligned}$$

which implies that

$$\frac{\|\sum_{j\notin I} x_j^* \mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)} \leq \sum_{j\notin I} c_j |x_j^*|. \tag{18}$$

Furthermore, consider the vector $\sum_{j\notin I} x_j^* \mathbf{a}_j$, denote $(\sum_{j\notin I} x_j^* \mathbf{a}_j)^=$ as the projection of this vector on the direction of the vector $\mathbf{b} - \sum_{i\in I} x_i^* \mathbf{a}_i$, and $(\sum_{j\notin I} x_j^* \mathbf{a}_j)^+$ the residual. Therefore, the assumption that

$$(\mathbf{b} - \sum_{i\in I} x_i^* \mathbf{a}_i)^\top (\sum_{j\notin I} x_j^* \mathbf{a}_j) < \frac{\|\mathbf{b} - \sum_{i\in I} x_i^* \mathbf{a}_i\|_2 \|\sum_{j\notin I} x_j^* \mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)},$$

implies

$$\|(\sum_{j\notin I} x_j^* \mathbf{a}_j)^=\|_2 < \frac{\|\sum_{j\notin I} x_j^* \mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)}.$$

Hence we have

$$\begin{aligned}
&\|\mathbf{b} - A\mathbf{x}^*\|_2 \\
=&\|\mathbf{b} - \sum_{i\in I} x_i^*\mathbf{a}_i - \sum_{j\notin I} x_j^*\mathbf{a}_j\|_2 \\
=&\|\mathbf{b} - \sum_{i\in I} x_i^*\mathbf{a}_i - (\sum_{j\notin I} x_j^*\mathbf{a}_j)^= - (\sum_{j\notin I} x_j^*\mathbf{a}_j)^\perp\|_2 \\
\geq&\|\mathbf{b} - \sum_{i\in I} x_i^*\mathbf{a}_i - (\sum_{j\notin I} x_j^*\mathbf{a}_j)^=\|_2 \\
\geq&\|\mathbf{b} - \sum_{i\in I} x_i^*\mathbf{a}_i\|_2 - \|(\sum_{j\notin I} x_j^*\mathbf{a}_j)^=\|_2 \\
>&\|\mathbf{b} - \sum_{i\in I} x_i^*\mathbf{a}_i\|_2 - \frac{\|\sum_{j\notin I} x_j^*\mathbf{a}_j\|_2}{\max_{t\notin I}(\|\mathbf{a}_t\|_2/c_t)},
\end{aligned} \tag{19}$$

where the first inequality holds because $(\sum_{j\notin I} x_j^*\mathbf{a}_j)^\perp$ is orthogonal to both $\mathbf{b} - \sum_{i\in I} x_i^*\mathbf{a}_i$ and $(\sum_{j\notin I} x_j^*\mathbf{a}_j)^=$. Substitute Inequality (18) and (19) into Equation (equ.sparsityanglemain) implies that

$$\|\mathbf{b} - A\hat{\mathbf{x}}^*\|_2 + \sum_i |\hat{x}_i^*| < \|\mathbf{b} - A\mathbf{x}^*\|_2 + \sum_i |x_i^*|.$$

Hence $\mathbf{x}^*$ is not an optimal solution. □

E Proof of Theorem 9

Theorem 9 *Given $I \subseteq \{1, \cdots, m\}$, if there exists a non-zero vector $(w_i)_{i\in I}$ satisfying*

$$\|\sum_{i\in I} w_i\mathbf{a}_i\|_2 \leq \min_{\sigma_i\in\{-1,+1\}} |\sum_{i\in I} \sigma_i c_i w_i|,$$

then there exists an optimal solution $\mathbf{x}^*$ such that $\exists i \in I: x_i^* = 0$.

Proof. Suppose there is an optimal solution $\hat{\mathbf{x}}$ which puts non-zero weights on all $i \in I$, and we show that there exists a solution $\mathbf{x}^*$ which is at least as good as $\hat{\mathbf{x}}$ and satisfies $\prod_{i\in I} x_i^* = 0$.

Since for all $i \in I$, $\hat{x}_i \neq 0$, there exists $\lambda^+ > 0$ such that for all $i \in I$ $(\hat{x}_i + \lambda^+ w_i)\hat{x}_i \geq 0$ with at least one equality holds. That is, we push the solution until one of the elements reach zero. Similarly, there exists $\lambda^- > 0$ such that for all $i \in I$ $(\hat{x}_i - \lambda^- w_i)\hat{x}_i \geq 0$ with at least one equality holds. Notice, either $\sum_{i\in I} c_i|\hat{x}_i + \lambda^+ w_i| \leq \sum_{i\in I} |\hat{x}_i|$ holds or $\sum_{i\in I} c_i|\hat{x}_i - \lambda^- w_i| \leq \sum_{i\in I} |\hat{x}_i|$ holds, since $x_i \neq 0$ for all $i \in I$. Without loss of generality we assume that $\sum_{i\in I} c_i|\hat{x}_i + \lambda^+ w_i| \leq \sum_{i\in I} |\hat{x}_i|$. Notice that

$$\begin{aligned}
&\sum_{i\in I} c_i|\hat{x}_i + \lambda^+ w_i| - \sum_{i\in I} c_i|\hat{x}_i| \\
=&\sum_{i\in I} \mathrm{sign}(w_i\hat{x}_i) \cdot c_i\lambda^+ w_i \\
\leq& -\lambda^+ \min_{\sigma_i\in\{-1,+1\}} |\sum_{i\in I} \sigma_i c_i w_i| \\
\leq& -\lambda^+ \|\sum_{i\in I} w_i\mathbf{a}_i\|_2.
\end{aligned}$$

Construct $\mathbf{x}^*$ as

$$x_i^* \triangleq \begin{cases} \hat{x}_i + \lambda^+ w_i, & i \in I; \\ \hat{x}_i, & i \notin I. \end{cases}$$

Using the regularized form, we have

$$\begin{aligned}
&\|\mathbf{b} - A\mathbf{x}^*\|_2 + \sum_i c_i |x_i^*| \\
=&\|\mathbf{b} - A\hat{\mathbf{x}} + \lambda^+ \sum_{i\in I} w_i \mathbf{a}_i\|_2 + \sum_i c_i|\hat{x}_i| + \sum_i c_i|x_i^*| - \sum_i c_i|\hat{x}_i| \\
\leq&\|\mathbf{b} - A\hat{\mathbf{x}}\|_2 + \sum_i c_i|\hat{x}_i| + \big(\lambda^+ \| \sum_{i\in I} w_i\mathbf{a}_i\|_2 + \sum_{i\in I} c_i|\hat{x}_i + \lambda^+ w_i| - \sum_{i\in I} c_i|\hat{x}_i|\big) \\
\leq&\|\mathbf{b} - A\hat{\mathbf{x}}\|_2 + \sum_i c_i|\hat{x}_i|.
\end{aligned}$$

Therefore, $\mathbf{x}^*$ is at least as good as $\hat{\mathbf{x}}$ and has zero elements for at least one $i \in I$. $\square$

Notice in the proof, $\mathbf{x}^*$ has strictly fewer non-zero elements than $\hat{\mathbf{x}}$.

F Proof of Proposition 2

Proposition 2 *Given a function $g : \mathbb{R}^{m+1} \to \mathbb{R}$ and Borel sets $\mathcal{Z}_1, \cdots, \mathcal{Z}_n \subseteq \mathbb{R}^{m+1}$, let*

$$\mathcal{P}_n \triangleq \{\mu \in \mathcal{P} | \forall S \subseteq \{1, \cdots, n\} : \mu(\bigcup_{i\in S} \mathcal{Z}_i) \geq |S|/n\}.$$

The following holds

$$\frac{1}{n}\sum_{i=1}^n \sup_{(\mathbf{r}_i, b_i)\in\mathcal{Z}_i} h(\mathbf{r}_i, b_i) = \sup_{\mu\in\mathcal{P}_n} \int_{\mathbb{R}^{m+1}} h(\mathbf{r}, b) d\mu(\mathbf{r}, b).$$

To prove Proposition 2, we first establish the following lemma.

Lemma 2 *Given a function $f : \mathbb{R}^{m+1} \to \mathbb{R}$, and a Borel set $\mathcal{Z} \subseteq \mathbb{R}^{m+1}$, the following holds:*

$$\sup_{\mathbf{x}'\in\mathcal{Z}} f(\mathbf{x}') = \sup_{\mu\in\mathcal{P}|\mu(\mathcal{Z})=1} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}).$$

Proof. Let $\hat{\mathbf{x}}$ be a $\epsilon-$optimal solution to the left hand side, consider the probability measure μ' that put mass 1 on $\hat{\mathbf{x}}$, which satisfy $\mu'(\mathcal{Z}) = 1$. Hence, we have

$$\sup_{\mathbf{x}'\in\mathcal{Z}} f(\mathbf{x}') - \epsilon \leq \sup_{\mu\in\mathcal{P}|\mu(\mathcal{Z})=1} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}),$$

since ϵ can be arbitrarily small, this leads to

$$\sup_{\mathbf{x}'\in\mathcal{Z}} f(\mathbf{x}') \leq \sup_{\mu\in\mathcal{P}|\mu(\mathcal{Z})=1} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}). \tag{20}$$

Next construct function $\hat{f}: \mathbb{R}^{m+1} \to \mathbb{R}$ as

$$\hat{f}(\mathbf{x}) \triangleq \begin{cases} f(\hat{\mathbf{x}}) & \mathbf{x} \in \mathcal{Z}; \\ f(\mathbf{x}) & \text{otherwise.} \end{cases}$$

By definition of $\hat{\mathbf{x}}$ we have $f(\mathbf{x}) \leq \hat{f}(\mathbf{x}) + \epsilon$ for all $\mathbf{x} \in \mathbb{R}^{m+1}$. Hence, for any probability measure μ such that $\mu(\mathcal{Z}) = 1$, the following holds

$$\int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(x) \leq \int_{\mathbb{R}^{m+1}} \hat{f}(\mathbf{x}) d\mu(x) + \epsilon = f(\hat{\mathbf{x}}) + \epsilon \leq \sup_{\mathbf{x}' \in \mathcal{Z}} f(\mathbf{x}') + \epsilon.$$

This leads to

$$\sup_{\mu \in \mathcal{P} | \mu(\mathcal{Z})=1} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(x) \leq \sup_{\mathbf{x}' \in \mathcal{Z}} f(\mathbf{x}') + \epsilon.$$

Notice ϵ can be arbitrarily small, we have

$$\sup_{\mu \in \mathcal{P} | \mu(\mathcal{Z})=1} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(x) \leq \sup_{\mathbf{x}' \in \mathcal{Z}} f(\mathbf{x}') \tag{21}$$

Combining (20) and (21), we prove the lemma. □

Now we proceed to prove the proposition. Let $\hat{\mathbf{x}}_i$ be an ϵ-optimal solution to $\sup_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i)$. Observe that the empirical distribution for $(\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_n)$ belongs to $\mathcal{P}_n$, since ϵ can be arbitrarily close to zero, we have

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i) \leq \sup_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}). \tag{22}$$

Without loss of generality, assume

$$f(\hat{\mathbf{x}}_1) \leq f(\hat{\mathbf{x}}_2) \leq \cdots \leq f(\hat{\mathbf{x}}_n). \tag{23}$$

Now construct the following function

$$\hat{f}(\mathbf{x}) \triangleq \begin{cases} \min_{i | \mathbf{x} \in \mathcal{Z}_i} f(\hat{\mathbf{x}}_i) & \mathbf{x} \in \bigcup_{j=1}^{n} \mathcal{Z}_j; \\ f(\mathbf{x}) & \text{otherwise.} \end{cases} \tag{24}$$

Observe that $f(\mathbf{x}) \leq \hat{f}(\mathbf{x}) + \epsilon$ for all $\mathbf{x}$.

Furthermore, given $\mu \in \mathcal{P}_n$, we have

$$\begin{aligned} & \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}) - \epsilon \\ = & \int_{\mathbb{R}^{m+1}} \hat{f}(\mathbf{x}) d\mu(\mathbf{x}) \\ = & \sum_{k=1}^{n} f(\hat{\mathbf{x}}_k) \Big[\mu(\bigcup_{i=1}^{k} \mathcal{Z}_i) - \mu(\bigcup_{i=1}^{k-1} \mathcal{Z}_i) \Big] \end{aligned}$$

Denote $\alpha_k \triangleq \left[\mu(\bigcup_{i=1}^{k}\mathcal{Z}_i) - \mu(\bigcup_{i=1}^{k-1}\mathcal{Z}_i)\right]$, we have

$$\sum_{k=1}^{n}\alpha_k = 1, \quad \sum_{k=1}^{t}\alpha_k \geq t/n.$$

Hence by Equation (23) we have

$$\sum_{k=1}^{n}\alpha_k f(\hat{\mathbf{x}}_k) \leq \frac{1}{n}\sum_{k=1}^{n} f(\hat{\mathbf{x}}_k).$$

Thus we have for any $\mu \in \mathcal{P}_n$,

$$\int_{\mathbb{R}^{m+1}} f(\mathbf{x})d\mu(\mathbf{x}) - \epsilon \leq \frac{1}{n}\sum_{k=1}^{n} f(\hat{\mathbf{x}}_k).$$

Therefore,

$$\sup_{\mu\in\mathcal{P}_n}\int_{\mathbb{R}^{m+1}} f(\mathbf{x})d\mu(\mathbf{x}) - \epsilon \leq \sup_{\mathbf{x}_i\in\mathcal{Z}_i}\frac{1}{n}\sum_{k=1}^{n} f(\mathbf{x}_k).$$

Notice ϵ can be arbitrarily close to 0, we proved the proposition by combining with (22).

References

[1] L. Elden. Perturbation theory for the least-square problem with linear equality constraints. *BIT*, 24:472–476, 1985.

[2] G. Golub and C. Van Loan. *Matrix Computation.* John Hopkins University Press, Baltimore, 1989.

[3] D. Higham and N. Higham. Backward error and condition of structured linear systems. *SIAM Journal on Matrix Analysis and Applications*, 13:162–175, 1992.

[4] R. Fierro and J. Bunch. Collinearity and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 15:1167–1181, 1994.

[5] A. Tikhonov and V. Arsenin. *Solution for Ill-Posed Problems*. Wiley, New York, 1977.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[8] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[9] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.

[10] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[11] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[12] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. Technical Report Available from: `http://http://www.stat.berkeley.edu/tech-reports/709.pdf`, Department of Statistics, UC Berkeley, 2006.

[13] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.

[14] P. Shivaswamy, C. Bhattacharyya, and A. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, July 2006.

[15] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal of Optimization*, 15(3):780–800, 2004.

[16] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, January 2004.

[17] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal Of The Royal Statistical Society Series B*, 67(2):301–320, 2005.

[18] J. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory*, 51(3):1030–1051, 2006.

[19] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1445–1480, 1998.

[20] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[21] S. Mallat and Z. Zhang. Matching Pursuits with time-frequence dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[22] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[23] H. Xu, C. Caramanis, and S. Mannor. Robust optimization and max min expected utility. Technical report, The University of Texas at Austin, 2008.

[24] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[25] E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[26] L. Devroye and L. Györfi. *Nonparametric Density Estimation: the l_1 View*. John Wiley & Sons, 1985.

[27] D. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons, 1992.

[28] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 2000.

[29] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.