Robust Regression and Lasso

Huan Xu
Department of Electrical and Computer Engineering
McGill University
Montreal, QC Canada
xuhuan@cim.mcgill.ca

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

Shie Mannor
Department of Electrical and Computer Engineering
McGill University
Montreal, QC Canada
shie.mannor@mcgill.ca

Abstract

We consider robust least-squares regression with feature-wise disturbance. We show that this formulation leads to tractable convex optimization problems, and we exhibit a particular uncertainty set for which the robust problem is equivalent to ℓ_1 regularized regression (Lasso). This provides an interpretation of Lasso from a robust optimization perspective. We generalize this robust formulation to consider more general uncertainty sets, which all lead to tractable convex optimization problems. Therefore, we provide a new methodology for designing regression algorithms, which generalize known formulations. The advantage is that robustness to disturbance is a physical property that can be exploited: in addition to obtaining new formulations, we use it directly to show sparsity properties of Lasso, as well as to prove a general consistency result for robust regression problems, including Lasso, from a unified robustness perspective.

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, 2]. Many regularization methods have been proposed to decrease this sensitivity. Among them, Tikhonov regularization [3] and Lasso [4, 5] 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 [6, 7, 8, 9, 10] and references therein). In many of these approaches, the choice of regularization parameters often has no fundamental connection to an underlying noise model [2].

In [11], the authors propose an alternative approach to reducing sensitivity of linear regression, by considering a *robust version* of the regression problem: they minimize the worst-case residual for the observations under some unknown but bounded disturbances. They show that their robust least squares formulation is equivalent to ℓ_2-regularized least squares, and they explore computational aspects of the problem. In that paper, and in most of the subsequent research in this area and the more general area of Robust Optimization (see [12, 13] and references therein) the disturbance is taken to be either row-wise and uncorrelated [14], or given by bounding the Frobenius norm of the disturbance matrix [11].

In this paper we investigate the robust regression problem under more general uncertainty sets, focusing in particular on the case where the uncertainty set is defined by feature-wise constraints, and also the case where features are meaningfully correlated. This 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. We prove that all our formulations are computationally tractable. Unlike much of the previous literature, we provide a focus on *structural properties* of the robust solution. In addition to giving new formulations, and new properties of the solutions to these robust problems, we focus on the inherent importance of robustness, and its ability to prove from scratch important properties such as sparseness, and asymptotic consistency of Lasso in the statistical learning context. In particular, our main contributions in this paper are as follows.

- 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. This can be helpful in choosing the regularization parameter. We generalize the robust regression formulation to loss functions given by an arbitrary norm, and uncertainty sets that allow correlation between disturbances of different features.
- We investigate the sparsity properties for the robust regression problem with feature-wise independent disturbances, showing that such formulations encourage sparsity. We thus easily recover standard sparsity results for Lasso using a robustness argument. This also implies a fundamental connection between the *feature-wise independence* of the disturbance and the sparsity.
- 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.

Notation. We use capital letters to represent matrices, and boldface letters to represent column vectors. For a vector $\mathbf{z}$, we let z_i denote the i^{th} element. Throughout the paper, $\mathbf{a}_i$ and $\mathbf{r}_j^\top$ denote the i^{th} column and the j^{th} row of the observation matrix A, respectively; a_{ij} is 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

We show that our robust regression formulation recovers Lasso as a special case. 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 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. In Section 3, we show the robust formulation gives rise to new sparsity results. Some of our results there (e.g. Theorem 4) fundamentally depend on (and follow from) the robustness argument, which is not found elsewhere in the literature. Then in Section 4, we establish consistency of Lasso directly from the robustness properties of our formulation, thus explaining consistency from a more physically motivated and perhaps more general perspective.

2.1 Formulation

Robust linear regression considers the case that the observed matrix A is corrupted by some disturbance. We seek the optimal weight for the uncorrupted (yet unknown) sample matrix. We consider the following min-max formulation:

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

Here, $\mathcal{U}$ is the set of admissible disturbances of the matrix A. In this section, we consider the specific setup where the disturbance is feature-wise uncorrelated, and norm-bounded for each feature:

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

for given $c_i \ge 0$. This formulation recovers the well-known Lasso:

Theorem 1. *The robust regression problem (1) with the uncertainty set (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. We defer the full details to [15], and give only an outline of the proof here. Showing that the robust regression is a lower bound for the regularized regression follows from the standard triangle inequality. Conversely, one can take the worst-case noise to be $\boldsymbol{\delta}_i^* \triangleq -c_i \mathrm{sgn}(x_i^*)\mathbf{u}$, where $\mathbf{u}$ is given by

$$\mathbf{u} \triangleq \begin{cases} \frac{\mathbf{b}-A\mathbf{x}^*}{\|\mathbf{b}-A\mathbf{x}^*\|_2} & \text{if } A\mathbf{x}^* \neq \mathbf{b}, \\ \text{any vector with unit } \ell_2 \text{ norm} & \text{otherwise;} \end{cases},$$

from which the result follows after some algebra. □

If we take $c_i = c$ and normalized $\mathbf{a}_i$ for all i, Problem (3) is the well-known Lasso [4, 5].

2.2 Arbitrary norm and correlated disturbance

It is possible to generalize this result to the case where the ℓ_2-norm is replaced by an arbitrary norm, and where the uncertainty is correlated from feature to feature. For space considerations, we refer to the full version ([15]), and simply state the main results here.

Theorem 2. *Let $\|\cdot\|_a$ denote an arbitrary norm. Then the robust regression problem*

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

is equivalent to the regularized regression problem $\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \|\mathbf{b} - A\mathbf{x}\|_a + \sum_{i=1}^m c_i |x_i| \right\}$.

Using feature-wise uncorrelated disturbance may lead to overly conservative results. We relax this, allowing the disturbances of different features to be correlated. Consider the following uncertainty set:

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

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. The following theorem converts this formulation to a convex and tractable optimization problem.

Theorem 3. *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. 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

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^k_+,\boldsymbol{\kappa}\in\mathbb{R}^m_+,\mathbf{x}\in\mathbb{R}^m} \Big\{\|\mathbf{b}-A\mathbf{x}\|_a + v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x})\Big\};$$
$$\text{where: } v(\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{x}) \triangleq \max_{\mathbf{c}\in\mathbf{R}^m}\Big[(\boldsymbol{\kappa}+|\mathbf{x}|)^\top\mathbf{c} - \sum_{j=1}^k \lambda_j f_j(\mathbf{c})\Big]. \tag{4}$$

Example 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*

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

The robust regression formulation (1) considers disturbances that are bounded in a set, while in practice, often the disturbance is a random variable with unbounded support. In such cases, it is not possible to simply use an uncertainty set that includes all admissible disturbances, and we need to construct a meaningful $\mathcal{U}$ based on probabilistic information. In the full version [15] we consider computationally efficient ways to use chance constraints to construct uncertainty sets.

3 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 discussed (cf [6, 7, 8, 9]), and 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., [16]).

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. In particular, we show that (i) as a direct result of *feature-wise independence* of the uncertainty set, a slight change of a feature that was originally assigned zero weight still gets zero weight (Theorem 4); (ii) using Theorem 4, we show that "nearly" orthogonal features get zero weight (Corollary 1); and (iii) "nearly" linearly dependent features get zero weight (Theorem 5). Substantial research regarding sparsity properties of Lasso can be found in the literature (cf [6, 7, 8, 9, 17, 18, 19, 20] and many others). In particular, similar results as in point (ii), that rely on an *incoherence* property, have been established in, e.g., [16], and are used as standard tools in investigating sparsity of Lasso from a 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.

Theorem 4. *Given* $(\tilde{A},\mathbf{b})$*, let* $\mathbf{x}^*$ *be an optimal solution of the robust regression problem:*

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

Let $I\subseteq\{1,\cdots,m\}$ *be such that for all* $i\in I$*,* $x_i^*=0$*. Now let*

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

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

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

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

Proof. Notice that for $i \in I$, $x_i^* = 0$, hence the i^{th} column of both A and ΔA has no effect on the residual. We have

$$\max_{\Delta A \in \tilde{\mathcal{U}}} \left\| \mathbf{b} - (A + \Delta A)\mathbf{x}^* \right\|_2 = \max_{\Delta A \in \mathcal{U}} \left\| \mathbf{b} - (A + \Delta A)\mathbf{x}^* \right\|_2 = \max_{\Delta A \in \mathcal{U}} \left\| \mathbf{b} - (\tilde{A} + \Delta A)\mathbf{x}^* \right\|_2.$$

For $i \in I$, $\|\mathbf{a}_i - \tilde{\mathbf{a}}_i\| \leq l_i$, and $\mathbf{a}_j = \tilde{\mathbf{a}}_j$ for $j \notin I$. Thus $\left\{\tilde{A} + \Delta A \middle| \Delta A \in \mathcal{U}\right\} \subseteq \left\{A + \Delta A \middle| \Delta A \in \tilde{\mathcal{U}}\right\}$. Therefore, for any fixed $\mathbf{x}'$, the following holds:

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

By definition of $\mathbf{x}^*$,

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

Therefore we have

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

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

Theorem 4 is established *using the robustness argument*, and is a direct result of the *feature-wise independence* of the uncertainty set. It explains why Lasso tends to assign zero weight to non-relative features. Consider 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 $i' \notin I$, 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. This is also shown in the next corollary, in which we apply Theorem 4 to show that the problem has a sparse solution as long as an incoherence-type property is satisfied (this result is more in line with the traditional sparsity results).

Corollary 1. *Suppose that for all i, $c_i = c$. If there exists $I \subset \{1, \cdots, m\}$ such that for all $\mathbf{v} \in \operatorname{span}\big(\{\mathbf{a}_\mathrm{i}, \mathrm{i} \in \mathrm{I}\} \bigcup \{\mathbf{b}\}\big)$, $\|\mathbf{v}\| = 1$, 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 $\min_{\hat{\mathbf{x}}} \left\{ \max_{\Delta A \in \hat{\mathcal{U}}} \left\| \mathbf{b} - (\hat{A} + \Delta A)\hat{\mathbf{x}} \right\|_2 \right\}$, which is equivalent to $\min_{\hat{\mathbf{x}}} \left\{ \left\| \mathbf{b} - \hat{A}\hat{\mathbf{x}} \right\|_2 + \sum_{i \in I} c|\hat{x}_i| \right\}$. Now we show that there exists an optimal solution $\hat{\mathbf{x}}^*$ such that $\hat{x}_j^* = 0$ for all $j \notin I$. This is because $\hat{\mathbf{a}}_j$ are orthogonal to the span of 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 4 we establish the corollary. □

The next corollary follows easily from Corollary 1.

Corollary 2. *Suppose there exists $I \subseteq \{1, \cdots, m\}$, such that for all $i \in I$, $\|\mathbf{a}_i\| < c_i$. Then any optimal solution $\mathbf{x}^*$ satisfies $x_i^* = 0$, for $i \in I$.*

[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 shows that sparsity is achieved when a set of features are "almost" linearly dependent. Again we refer to [15] for the proof.

Theorem 5. *Given $I \subseteq \{1, \cdots, m\}$ such that 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 $\exists i \in I : \; x_i^* = 0$.

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 3. *Given $I \subseteq \{1, \cdots, m\}$, let $A_I \triangleq \Big(\mathbf{a}_i \Big)_{i \in I}$, and $t \triangleq \operatorname{rank}(A_I)$. 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 4. *If $n < m$, then there exists an optimal solution with no more than n non-zero coefficients.*

4 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 the full version ([15]) 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 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

$$\mathbf{x}(c_n, \mathcal{S}_n) \triangleq \arg\min_{\mathbf{x}} \Big\{ \sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \mathbf{x})^2} + c_n \|x\|_1 \Big\} = \arg\min_{\mathbf{x}} \Big\{ \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \mathbf{x})^2} + c_n \|x\|_1 \Big\};$$

$$\mathbf{x}(\mathbb{P}) \triangleq \arg\min_{\mathbf{x}} \Big\{ \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \mathbf{x})^2 d\mathbb{P}(b, \mathbf{r})} \Big\}.$$

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 the full version ([15]) 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 6. *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.

The full proof and results we develop along the way are deferred to [15], but we provide the main ideas and outline here. 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 1. *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

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

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

$$h_n(b, \mathbf{r}) \triangleq (nc^{m+1})^{-1} \sum_{i=1}^{n} K\left(\frac{b - b_i, \mathbf{r} - \mathbf{r}_i}{c}\right), \quad \text{where: } K(\mathbf{x}) \triangleq I_{[-1,+1]^{m+1}}(\mathbf{x})/2^{m+1}. \tag{5}$$

Observe that the estimated distribution given by Equation (5) 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 1.

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 [21]), 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 6, the proof technique rather than the result itself is of interest. We postpone the proof to [15].

Theorem 7. *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.

5 Conclusion

In this paper, we consider robust regression with a least-square-error loss, and extend the results of [11] (i.e., Tikhonov regularization is equivalent to a robust formulation for Frobenius norm-bounded disturbance set) to a broader range of disturbance sets and hence regularization schemes. A special case of our formulation recovers the well-known Lasso algorithm, and we obtain an interpretation of Lasso from a robustness perspective. We consider more general robust regression formulations, allowing correlation between the feature-wise noise, and we show that this too leads to tractable convex optimization problems.

We exploit the new robustness formulation to give direct proofs of sparseness and consistency for Lasso. As our results follow from robustness properties, it suggests that they may be far more general than Lasso, and that in particular, consistency and sparseness may be properties one can obtain more generally from robustified algorithms.

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

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

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

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

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

[8] 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.

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

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

[11] 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.

[12] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, August 1999.

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

[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] H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. Submitted, available from `http://arxiv.org/abs/0811.1790v1`, 2008.

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

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

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

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

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

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