Robust Regression and Lasso

Huan Xu, Constantine Caramanis, *Member, IEEE*, and Shie Mannor, *Senior Member, IEEE*

***Abstract*—Lasso, or ℓ^1 regularized least squares, has been explored extensively for its remarkable sparsity properties. In this paper it is shown 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 a principled selection of the regularizer, and in particular, generalizations of Lasso that also yield convex optimization problems are obtained by considering different uncertainty sets. Second, robustness can itself be used as an avenue for exploring different properties of the solution. In particular, it is shown that robustness of the solution explains why the solution is sparse. The analysis as well as the specific results obtained differ from standard sparsity results, providing different geometric intuition. Furthermore, it is shown that the robust optimization formulation is related to kernel density estimation, and based on this approach, a proof that Lasso is consistent is given, using robustness directly. Finally, a theorem is proved which states that sparsity and algorithmic stability contradict each other, and hence Lasso is not stable.**

***Index Terms*—Lasso, regression, regularization, robustness, sparsity, stability, statistical learning.**

I. 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.

Manuscript received October 17, 2008; revised July 22, 2009. Current version published June 16, 2010. The work of C. Caramanis was supported in part by the NSF by Grants EFRI-0735905, CNS-0721532, CNS-0831580, and by DTRA by Grant HDTRA1-08-0029. The work of S. Mannor was partially supported by the Israel Science Foundation under Contract 890015 and by a Horev Fellowship. The material in this paper was presented in part at the 22nd Annual Conference on Neural Information Processing Systems, Vancouver, Canada, December 2008.

H. Xu was with the Department of Electrical and Computer Engineering, McGill University, Montréal, H3A2A7, Canada. He is now with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712 USA (e-mail: huan.xu@mail.utexas.edu).

C. Caramanis is with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712 USA (e-mail: caramanis@mail.utexas.edu).

S. Mannor was with the Department of Electrical and Computer Engineering, McGill University, Montréal, QC H3A2A7 Canada. He is now with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: shie@ee.technion.ac.il).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2048503

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 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 uncoupled [14], or the case where 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 preprocessing steps, and the magnitudes of such noises are known or bounded. Another situation of interest is where noise or disturbance across features is meaningfully coupled. We define *coupled* and *uncoupled* disturbances and uncertainty sets precisely in Section II-A below. Intuitively, a disturbance is feature-wise coupled if the variation or disturbance satisfies joint constraints across features, and it is uncoupled 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 for principled selection of the regularizer based on noise properties. Moreover, by considering different uncertainty sets, we construct generalizations of Lasso that also yield convex optimization problems.

Second, and perhaps most significantly, robustness is a strong property that can itself be used as an avenue for 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 as well as the organization of this paper.

- In Section II, 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 of Lasso from a robustness perspective.
- We generalize the robust regression formulation to loss functions of arbitrary norm in Section III. 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 to obtain solutions with additional properties.
- In Section IV, we present new sparsity results for the robust regression problem with feature-wise independent disturbances. This provides a new robustness-based explanation for the sparsity of Lasso. 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.
- We relate Lasso to kernel density estimation in Section V. This allows us to reprove 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 in Section VI a "no-free-lunch" theorem, stating that an algorithm that encourages sparsity cannot be stable.

A. Notation

We use capital letters to represent matrices, and boldface letters to represent column vectors. Row vectors are represented as the transpose of column vectors. For a vector $\mathbf{z}$, z_i denotes its ith element. Throughout the paper, $\mathbf{a}_i$ and $\mathbf{r}_j^\top$ are used to denote the ith column and the jth row of the observation matrix A, respectively. We use a_{ij} to denote the ij element of A, hence it is the jth element of $\mathbf{r}_i$, and ith element of $\mathbf{a}_j$. For a convex function $f(\cdot)$, $\partial f(\mathbf{z})$ represents any of its subgradients evaluated at $\mathbf{z}$. The all-ones vector of length n is denoted by $\mathbf{1}_n$.

II. Robust Regression With Feature-Wise Disturbance

In this section, 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 can be shown that these two coincide up to a change of the regularization coefficient, since the solution set to either formulation is the Pareto efficient set of the regression error and the regularization penalty. (We provide the detailed argument in Appendix A for completeness.)

A. 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

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

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

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

for given $c_i\ge 0$. We call these uncertainty sets *feature-wise uncoupled*, in contrast to *coupled* 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 uncoupled 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 (LHS) can be written as

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

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

$$\boldsymbol{\delta}_i^* \triangleq \begin{cases} -c_i \text{sgn}(x_i^*)\mathbf{u} & \text{if } x_i^* \neq 0; \\ -c_i \mathbf{u} & \text{otherwise.} \end{cases}$$

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

$$\begin{aligned} &\max_{\Delta A \in \mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}^*\|_2 \\ &\quad \geq \|\mathbf{b} - (A + \Delta A^*)\mathbf{x}^*\|_2 \\ &\quad = \|\mathbf{b} - (A + (\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_m^*))\mathbf{x}^*\|_2 \\ &\quad = \left\| (\mathbf{b} - A\mathbf{x}^*) - \sum_{i: x_i^* \neq 0} (-x_i^* c_i \text{sgn}(x_i^*)\mathbf{u}) \right\|_2 \\ &\quad = \left\| (\mathbf{b} - A\mathbf{x}^*) + \left(\sum_{i=1}^m c_i \, |x_i^*| \right) \mathbf{u} \right\|_2 \\ &\quad = \|\mathbf{b} - A\mathbf{x}^*\|_2 + \sum_{i=1}^m c_i \, |x_i^*| \, . \end{aligned} \tag{5}$$

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

Combining (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. Note that the theorem remains valid if we replace $\mathcal{U}$ by its boundary set $\{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \, \|\boldsymbol{\delta}_i\|_2 = c_i, \, i = 1, \ldots, m\}$. ■

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

In many setups of interest, the noise $\boldsymbol{\delta}_i$ is random with a known distribution and possibly an unbounded support. Thus, a robust regression formulation that allows all admissible ΔA can be too pessimistic or even meaningless. One remedy is to replace the deterministic guarantee by a probabilistic one, i.e., to find a parameter c_i such that $\|\boldsymbol{\delta}_i\|_2 \leq c_i$ holds with a given confidence $1 - \eta$. This can be achieved via line search and bisection, provided that we can evaluate $\Pr(\|\boldsymbol{\delta}_i\|_2 \geq c)$. Note that the distribution of $\boldsymbol{\delta}_i$ is known, thus we can evaluate this probability via Monte Carlo sampling.

Another setting of interest is when we have access only to the mean and variance of the distribution of the uncertainty (e.g., [15]–[17]). 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. Furthermore, this worst-case probability can be evaluated by solving a Semi-Definite Program. We elaborate the detailed uncertainty set construction for both settings in [18]; see also [19].

III. General Uncertainty Sets

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

In this section, 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 coupled uncertainty sets.

We first consider the case of an arbitrary norm $\|\cdot\|_a$ in $\mathbb{R}^n$ as a cost function rather than the squared loss. The proof of the next theorem is identical to that of Theorem 1, with only the ℓ^2 norm changed to $\|\cdot\|_a$.

Theorem 2: The robust regression problem

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

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 uncoupled. Allowing coupled 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:

$$\mathcal{U}' \triangleq \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \, f_j(\|\boldsymbol{\delta}_1\|_a, \ldots, \|\boldsymbol{\delta}_m\|_a) \leq 0; \quad j = 1, \ldots, k\}$$

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 Corollaries 1 and 2). The following theorem shows that this robust formulation is equivalent to a more general regularization-type problem, and thus converts this formulation to a tractable optimization problem. The proof is postponed to the Appendix.

Theorem 3: Assume that the set

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

has nonempty 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:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^k, \boldsymbol{\kappa} \in \mathbb{R}_+^m, \mathbf{x} \in \mathbb{R}^m} \{\|\mathbf{b} - A\mathbf{x}\|_a + v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \mathbf{x})\};$$
$$\text{where: } v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \mathbf{x}) \triangleq \max_{\mathbf{c} \in \mathbf{R}^m} \left[(\boldsymbol{\kappa} + |\mathbf{x}|)^\top \mathbf{c} - \sum_{j=1}^k \lambda_j f_j(\mathbf{c}) \right] \tag{6}$$

Remark: Problem (6) 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 subgradient 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

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

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 subgradient of $v(\cdot)$.

The next two corollaries are a direct application of Theorem 3.

Corollary 1: Suppose

$$\mathcal{U}' = \{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m) \,|\, \|\|\boldsymbol{\delta}_1\|_a,\ldots,\|\boldsymbol{\delta}_m\|_a\|_s \le l;\}$$

for a symmetric norm $\|\cdot\|_s$. Then the resulting regularized regression problem is

$$\min_{\mathbf{x}\in\mathbb{R}^m} \{\|\mathbf{b}-A\mathbf{x}\|_a + l\|\mathbf{x}\|_s^*\}$$

where $\|\cdot\|_s^*$ 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 bounded Frobenius norm, and Corollary 1 reduces to the robust formulation introduced in [13].

Corollary 2: Suppose

$$\mathcal{U}' = \{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m) \,|\, T\mathbf{c} \le \mathbf{s};\ \text{where: } c_j = \|\boldsymbol{\delta}_j\|_a\}$$

for a given matrix T and a vector $\mathbf{s}$. Then the resulting regularized regression problem is the following optimization problem on $\mathbf{x}$ and $\boldsymbol{\lambda}$:

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

Unlike previous results, this corollary considers general polytope uncertainty sets, i.e., the column-wise norm vector of the realizable uncertainty belongs to the polytope $\{T\mathbf{c} \le \mathbf{s}\}$. 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.

To further illustrate the power and flexibility of the robust optimization formulation, we provide in [18] other examples that are based on the robust regression formulation with different uncertainty sets that lead to tractable (i.e., convex) optimization problems. One notable example is the case where the perturbation of the matrix A can be decomposed as a column-wise uncoupled noise and a row-wise uncoupled noise. The resulting formulation resembles the elastic-net formulation [20], where there is a combination of ℓ^2 and ℓ^1 regularization. We refer the reader to [18] for the full details.

IV. 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]). There have been primarily two approaches to studying Lasso. 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 the conditions under which a pair $(A,\mathbf{b})$ defines a problem with sparse solutions (e.g., [21]).

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 4 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. For instance, using Theorem 4, we show, among other results, that "nearly" orthogonal features get zero weight (Theorem 5). Using similar tools, we provide additional results in [18]. There, we show, among other results, that the sparsity pattern of any optimal solution must satisfy certain angular separation conditions between the residual and the relevant features, and that "nearly" linearly dependent features get zero weight.

Substantial research regarding sparsity properties of Lasso can be found in the literature (cf [8]–[11], [22]–[26], and many others). In particular, results that rely on an *incoherence* property have been established in, e.g., [21], 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 with respect to 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 an index subset $I \subseteq \{1,\ldots,n\}$, and a matrix ΔA, let ΔA^I denote the restriction of ΔA on feature set I, i.e., ΔA^I equals ΔA on each feature indexed by $i \in I$, and is zero elsewhere. Similarly, given a feature-wise uncoupled uncertainty set $\mathcal{U}$, let $\mathcal{U}^I$ be the restriction of $\mathcal{U}$ to the feature set I, i.e., $\mathcal{U}^I \triangleq \{\Delta A^I | \Delta A \in \mathcal{U}\}$. Observe that any element $\Delta A \in \mathcal{U}$ can be written as $\Delta A^I + \Delta A^{I^c}$ (here $I^c \triangleq \{1,\ldots,n\} \setminus I$) such that $\Delta A^I \in \mathcal{U}^I$ and $\Delta A^{I^c} \in \mathcal{U}^{I^c}$.

Theorem 4: The robust regression problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \max_{\Delta A\in\mathcal{U}} \|\mathbf{b}-(A+\Delta A)\mathbf{x}\|_2 \right\} \tag{7}$$

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

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \max_{\Delta A\in\mathcal{U}^I} \|\mathbf{b} - (A + \Delta\tilde{A} + \Delta A)\mathbf{x}\|_2 \right\} \tag{8}$$

has a solution supported on the set I.

Thus, a robust regression has a solution supported on a set I, if we can find *one* (bounded) perturbation of the features corresponding to I^c that makes them *irrelevant* (i.e., not contributing to the regression error and hence with zero weight). An alternative interpretation is that if a certain matrix ($A+\Delta\tilde{A}^{I^c}$ in the theorem) is sparse under a more strict condition (note that $\mathcal{U}^I$ favors non-zero weight on I^c), then all its "neighboring" matrices are sparse under an "easier" condition. This leads to a novel technique of proving sparsity, by identifying a "neighboring" matrix that generates sparse solutions under a strict condition. Theorem 4 is indeed a special case of the following theorem with $c_j = 0$ for all $j \notin I$. The detailed proof is provided in Appendix C.

Theorem 4': Let $\mathbf{x}^*$ be an optimal solution of the robust regression problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \max_{\Delta A\in\mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_2 \right\} \tag{9}$$

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

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

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

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \max_{\Delta A\in\tilde{\mathcal{U}}} \|\mathbf{b} - (\tilde{A} + \Delta A)\mathbf{x}\|_2 \right\} \tag{10}$$

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

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

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

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

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

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

$$\{ A + \Delta A |\ \Delta A \in \mathcal{U}\} \subseteq \left\{ \tilde{A} + \Delta A \middle|\ \Delta A \in \tilde{\mathcal{U}} \right\}.$$

Finally, note that

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

Therefore, we have

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

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

A closer examination of the proof shows that we can replace the ℓ^2 norm loss by any loss function $f(\cdot)$ which satisfies the condition that if $x_j = 0$, A and A' only differ in the j^{th} column, then $f(\mathbf{b}, A, \mathbf{x}) = f(\mathbf{b}, A', \mathbf{x})$. This theorem thus suggests a methodology for constructing sparse algorithms by solving a robust optimization with respect to column-wise uncoupled uncertainty sets.

When we consider 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 4 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 5: Let $c_i = c$ for all i and consider ℓ^2 loss. Suppose that there exists $I \subset \{1,\ldots,m\}$ such that for all $\mathbf{v} \in \text{span}\left(\{\mathbf{a}_i, i\in I\}\bigcup\{\mathbf{b}\}\right), \|\mathbf{v}\| = 1$, we have $\mathbf{v}^\top\mathbf{a}_j \le c, \forall j \notin I$. Then there exists an optimal solution $\mathbf{x}^*$ that 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 $\|\mathbf{a}_j^=\| \le c$. 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

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

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 $\min_{\hat{\mathbf{x}}} \left\{ \left\|\mathbf{b} - \hat{A}\hat{\mathbf{x}}\right\|_2 + \sum_{i\in I} c|\hat{x}_i| \right\}$. 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 $\|\mathbf{a}_j - \hat{\mathbf{a}}_j\| = \|\mathbf{a}_j^=\| \le c\ \forall\ j \notin I$, (and recall that $\mathcal{U} = \{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \le c,\ \forall\ i\}$) applying Theorem 4 ' concludes the proof. ∎

To better understand the results of this section, we can consider a generative model[1] $b = \sum_{i\in I} w_i a_i + \tilde{\xi}$ where $I \subseteq \{1\ldots,m\}$ and $\tilde{\xi}$ is a random noise variable, i.e., b is generated by features belonging to I. If we further assume that the value of irrelevant features (i.e., features $\notin I$) are generated according to some zero-mean random rule (say Gaussian noise), it follows that as the number of samples increases the irrelevant features will be nearly orthogonal to the subspace spanned by the relevant features and $\mathbf{b}$ with high probability. Consequently, the irrelevant features will be assigned zero weight by Lasso.

V. Density Estimation and Consistency

In this section, we investigate the robust linear regression formulation from a statistical perspective and re-derive using only robustness properties that Lasso is asymptotically consistent. The basic idea of the consistency proof is as follows. We show that the robust optimization formulation can be seen to be the maximum error with respect to a class of probability measures. This class includes a kernel density estimator, and using this, we show that Lasso is consistent.

A. Robust Optimization, Worst-Case Expected Utility and Kernel Density Estimator

In this subsection, we present some notions and intermediate results. In particular, we link a robust optimization formulation with a worst case expected utility (with respect to a class of probability measures). Such results will be used in establishing the consistency of Lasso, as well as providing some additional insights on robust optimization. Proofs are postponed to the appendix. Throughout this section, we use $\mathcal{P}$ to represent the set of all probability measures (on Borel σ-algebra) of $\mathbb{R}^{m+1}$.

We first establish a general result on the equivalence between a robust optimization formulation and a worst-case expected utility:

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

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

The following holds:

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

This leads to the following corollary for Lasso, which states that for a given $\mathbf{x}$, the robust regression loss over the training data is equal to the worst-case expected *generalization error*.

Corollary 3: Given $\mathbf{b} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$, the following equation holds for any $\mathbf{x} \in \mathbb{R}^m$

$$\|\mathbf{b} - A\mathbf{x}\|_2 + \sqrt{n}c_n\|\mathbf{x}\|_1 + \sqrt{n}c_n = \sup_{\mu\in\hat{\mathcal{P}}(n)} \sqrt{n\int_{\mathbb{R}^{m+1}} (b' - \mathbf{r}'^{\top}\mathbf{x})^2 d\mu(\mathbf{r}',b')}. \quad (11)$$

[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.

Here[2]

$$\hat{\mathcal{P}}(n) \triangleq \bigcup_{\|\boldsymbol{\sigma}\|_2\leq\sqrt{n}c_n;\ \forall\, i:\|\boldsymbol{\delta}_i\|_2\leq\sqrt{n}c_n} \mathcal{P}_n(A,\Delta,\mathbf{b},\boldsymbol{\sigma});$$

$$\mathcal{P}_n(A,\Delta,\mathbf{b},\boldsymbol{\sigma}) \triangleq \{\mu \in \mathcal{P} | \mathcal{Z}_i = [b_i - \sigma_i, b_i + \sigma_i] \times \prod_{j=1}^{m} [a_{ij} - \delta_{ij}, a_{ij} + \delta_{ij}];$$

$$\forall\, S \subseteq \{1,\ldots,n\} : \mu\left(\bigcup_{i\in S} \mathcal{Z}_i\right) \geq |S|/n\}.$$

Remark 1: It is worth providing some further explanation of the meaning of Corollary 3. Equation (11) is nonprobabilistic. 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 involve any probabilistic operation such as taking expectation on the LHS. Instead, it is an equivalence relationship that holds for an arbitrary set of samples. Note that the right-hand side (RHS) 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 $\mathcal{P}_n(A,\Delta,\mathbf{b},\boldsymbol{\sigma})$ such that the norm of each column of Δ is bounded, where $\mathcal{P}_n(A,\Delta,\mathbf{b},\boldsymbol{\sigma})$ is the set of distributions corresponding to (see Proposition 1) disturbance in hyper-rectangle Borel sets $\mathcal{Z}_1,\ldots,\mathcal{Z}_n$ centered at $(b_i,\mathbf{r}_i^{\top})$ with lengths $(2\sigma_i, 2\delta_{i1},\ldots,2\delta_{im})$.

The proof of consistency relies on showing that this set $\hat{P}_n$ of distributions contains a kernel density estimator. Recall the basic definition: The kernel density estimator for a density $\hat{h}$ in $\mathbb{R}^d$, originally proposed in [27] and [28], is defined by

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

where $\{c_n\}$ is a sequence of positive numbers, $\hat{\mathbf{x}}_i$ are i.i.d. samples generated according to $\hat{h}$, and K is a Borel measurable function (kernel) satisfying $K \geq 0$, $\int K = 1$. See [29], [30] and references therein for detailed discussions. Fig. 1 illustrates a kernel density estimator using a Gaussian kernel for a randomly generated sample-set. A celebrated property of a kernel density estimator is that it converges in $\mathcal{L}^1$ to $\hat{h}$ when $c_n \downarrow 0$ and $nc_n^d \uparrow \infty$ [29].

B. Consistency of Lasso

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 nonuniform case is straightforward. Throughout, we use c_n to represent c where there are n samples (we take c_n to zero as n grows).

[2]Recall that a_{ij} is the jth element of $\mathbf{r}_i$

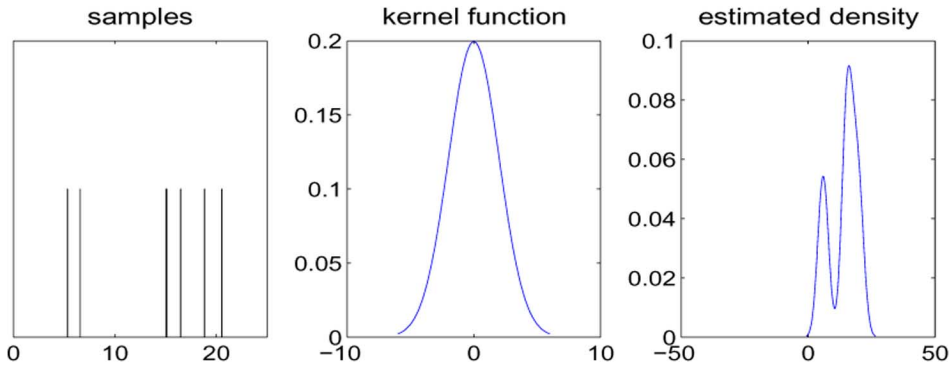

Fig. 1. Illustration of Kernel Density Estimation.

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

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

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 Vapnik-Chervonenkis (VC) dimension (e.g., [31]) and algorithmic stability (e.g., [32]) often work for a limited range of algorithms. For instance, the k-Nearest Neighbor is known to have infinite VC dimension, and we show in Section VI that Lasso is not stable. In contrast, a much wider range of algorithms have robustness interpretations, allowing a unified approach to prove their consistency. For example, it is shown in [33] that Support Vector Machines have a similar robust optimization interpretation, and one can prove consistency of SVMs from this robustness perspective.

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 $\|\mathbf{x}(c_n, \mathcal{S}_n)\|_2 \le H$ for all n. Then

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

almost surely.

Proof:

Step 1: We show that the RHS of (11) includes a kernel density estimator for the true (unknown) distribution. Consider the following kernel estimator given samples $\mathcal{S}_n = (b_i, \mathbf{r}_i)_{i=1}^n$ and tradeoff parameter c_n:

$$f_n(b,\mathbf{r}) \triangleq (nc_n^{m+1})^{-1} \sum_{i=1}^{n} K\left(\frac{b - b_i, \mathbf{r} - \mathbf{r}_i}{c_n}\right)$$
$$\text{where: } K(\mathbf{z}) \triangleq I_{[-1,+1]^{m+1}}(\mathbf{z})/2^{m+1}. \tag{12}$$

Let $\hat{\mu}_n$ denote the distribution given by the density function $f_n(b,\mathbf{r})$. It is easy to check that $\hat{\mu}_n$ belongs to $\mathcal{P}_n(A, (c_n\mathbf{1}_n, \ldots, c_n\mathbf{1}_n), \mathbf{b}, c_n\mathbf{1}_n)$ and hence it belongs to $\hat{\mathcal{P}}(n)$ by definition.

Step 2: Using the $\mathcal{L}^1$ convergence property of the kernel density estimator, we prove the consistency of robust regression and equivalently Lasso.

First note that the fact that $\|\mathbf{x}(c_n, \mathcal{S}_n)\|_2 \le H$ and $\mathbb{P}$ has a bounded support, implies that there exists a universal constant C such that

$$\max_{(b,\mathbf{r})\in \text{Support}(\mathbb{P})} (b - \mathbf{r}^\top \mathbf{x}(c_n, \mathcal{S}_n))^2 \le C.$$

By Corollary 3 and since $\hat{\mu}_n \in \hat{\mathcal{P}}(n)$, we have

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

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

Taking the square of both sides, we have

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

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

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

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 estimator of $f^*(\cdot)$ (see e.g.,[29, Theorem 3.1]). Hence the theorem follows. ∎

We can remove the assumption that $\|\mathbf{x}(c_n,\mathcal{S}_n)\|_2\le H$, and as in Theorem 6, the proof technique rather than the result itself is of interest.

Theorem 7: Let $\{c_n\}$ converge to zero sufficiently slowly. Then

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

almost surely.

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 $V_1,\ldots,V_T$ such the ℓ^∞ radius of each set is less than c_n. If a distribution μ satisfies

$$\mu(V_t)=\left|\{i|\,(b_i,\mathbf{r}_i)\in V_t\}\right|/n;\quad t=1,\ldots,T\tag{13}$$

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 jth 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 $S\subseteq\{1,\ldots,n\}$, the following holds:

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

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 the ℓ^∞ radius of each one is smaller than c_n. Let $\tilde{\mathcal{P}}(n)$ denote the set of probability measures satisfying (13). Hence $\tilde{\mathcal{P}}(n)\subseteq\hat{\mathcal{P}}(n)$ by Lemma 1. Further notice that there exists a universal constant K such that $\|\mathbf{x}(c_n,\mathcal{S}_n)\|_2\le K/c_n$ due to the fact that the square loss of the solution $\mathbf{x}=\mathbf{0}$ is bounded by a constant which only depends on the support of $\mathbb{P}$. Thus, there exists a constant C such that $\max_{b,\mathbf{r}}(b-\mathbf{r}^\top\mathbf{x}(c_n,\mathcal{S}_n))^2\le C/c_n^2$.

Following a similar argument as the proof of Theorem 6, we have

$$\begin{aligned}&\sup_{\mu_n\in\tilde{\mathcal{P}}(n)}\int_{b,\mathbf{r}}(b-\mathbf{r}^\top\mathbf{x}(c_n,\mathcal{S}_n))^2d\mu_n(b,\mathbf{r})\\&\quad\le\frac{1}{n}\sum_{i=1}^n(b_i-\mathbf{r}_i^\top\mathbf{x}(\mathbb{P}))^2+c_n^2(1+\|\mathbf{x}(\mathbb{P})\|_1)^2\\&\qquad+2c_n(1+\|\mathbf{x}(\mathbb{P})\|_1)\sqrt{\frac{1}{n}\sum_{i=1}^n(b_i-\mathbf{r}_i^\top\mathbf{x}(\mathbb{P}))^2}\end{aligned}\tag{14}$$

and (see equation at the bottom of the next page) where f_μ stands for the density function of a measure μ. Notice that $\tilde{\mathcal{P}}_n$ is the set of distributions satisfying (13), 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, [34, Prop. A6.6]). Therefore

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

if $c_n\downarrow 0$ sufficiently slowly. Combining this with (14) proves the theorem. ∎

VI. 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. Indeed, this is a special case of the main result of [35], where we show that a learning algorithm that possesses certain sparsity

properties cannot be stable. While we elaborate the Lasso case in this section for completeness, we refer the readers to [35] for the more general case.

Before giving formal definitions, we briefly comment on the difference between the notions of robustness and stability. In this paper, "robustness" stands for the property of an algorithm such that its output regression function, if tested on a sample set "similar" to the training set, will have a *testing error* close to the training error. Therefore, a robust-optimization based learning algorithm (e.g., Lasso) is inherently robust since it essentially minimizes an upper bound of testing-errors on sample sets that are "similar" to the training set. In contrast, "stability," as we define formally below, refers to the property of an algorithm that if trained on "slightly different" sample sets, it will have "similar" *output functions*. Therefore, in principle, an algorithm can be robust but non-stable, because while it can output two significantly different regression functions given two similar sample sets, the *error* if tested on the other set need not be much higher. This section shows that Lasso falls into precisely this category.

Now we recall the definition of uniform stability [32] first. At a high level, an algorithm is stable if the *output function* does not have a heavy dependence on any one given training sample, i.e., the output function changes only slightly if any one training sample is removed. To be more specific, we let $\mathcal{Z}$ denote the space of (labeled) samples (typically this will be a compact subset of $\mathbb{R}^{m+1}$) so that $S \in \mathcal{Z}^n$ denotes a collection of n labeled training points. We let $\mathbb{L}$ denote a learning algorithm, and for $S \in \mathcal{Z}^n$, 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 labeled point $s = (\mathbf{z}, b) \in \mathcal{Z}$, we let $l(\mathbb{L}_S, s)$ denote 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 a uniform stability bound of β_n with respect to the loss function l if the following holds:

$$\forall\, S \in \mathcal{Z}^n, \forall\, i \in \{1, \ldots, n\}, \|l(\mathbb{L}_S, \cdot) - l(\mathbb{L}_{S^{\setminus i}}, \cdot)\|_\infty \leq \beta_n.$$

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

The notion of uniform stability is of interest because it implies tight (in terms of rates) generalization bounds if the stability scales as fast as $o(\frac{1}{\sqrt{n}})$ (see [32]). It is also shown in [32] that many learning algorithms have reasonable stability. For example, Tikhonov-regularized regression (i.e., ℓ^2 regularization) has stability that scales as $1/n$.

In contrast, 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 a precise sense, as bad as it gets. To this end, we define the notion of the trivial bound as the worst possible error a training algorithm can have, given an arbitrary training set and testing sample labeled by zero.

Definition 2: Given a loss function $l(\cdot, \cdot)$, a subset from which we can draw m labeled points, $\mathcal{Z} \subseteq \mathbb{R}^{n \times (m+1)}$ and a subset for one unlabeled point, $\mathcal{X} \subseteq \mathbb{R}^m$, a trivial bound for a learning algorithm $\mathbb{L}$ with respect to $\mathcal{Z}$ and $\mathcal{X}$ is

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

Note 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. The proof is constructive: we provide an example (recall the uniform requirement in Definition 1) where by adding or removing one training sample, Lasso will output significantly different regression functions. At a high level, the instability of Lasso is due to the fact that its minimizing objective is nonsmooth, and thus (in some sense) ill-defined.

Theorem 8: Let $\hat{\mathcal{Z}} \subseteq \mathbb{R}^{n \times (2m+1)}$ be the domain of the sample set and $\hat{\mathcal{X}} \subseteq \mathbb{R}^{2m}$ be the domain of the new observation, such that

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

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

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

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 with respect to $(\mathbf{b}^*, A^*)$. 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 with respect 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 with respect to 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. ∎

We note that the example in the proof would not work for ℓ^2 regularization, simply because ℓ^2 regularization spreads the weight between identical features to achieve a strictly smaller regularization penalty, i.e., $(\mathbf{x}^*/2, \mathbf{x}^*/2)$ is a better solution than both $(\mathbf{x}^*, \mathbf{0})$ and $(\mathbf{0}, \mathbf{x}^*)$. Indeed, in [35] we show that it is the ability to identify redundant features (i.e., select only one between identical features) that leads to instability of Lasso and many other *sparse* algorithms.

VII. Conclusion

In this paper, we consider 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 coupled among different features and hence generalize Lasso to a wider class of regularization schemes.

The sparsity and consistency of Lasso are also investigated based on its robustness interpretation. 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 trade off 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 sets and hence regularization schemes. This not only provides us with new insights on why regularization schemes work, but also offers a solid motivation for selecting a regularization parameter for existing regularization schemes, and facilitate the design of new regularization schemes.

Appendix A
Equivalence of Two Formulations

In this appendix we show that the following two formulations:

$$\min_{\mathbf{x}} ||\mathbf{b} - A\mathbf{x}||_2 + c_1||\mathbf{x}||_1; \text{ and } \min_{\mathbf{x}} ||\mathbf{b} - A\mathbf{x}||_2^2 + c_2||\mathbf{x}||_1.$$

are equivalent up to a change of tradeoff parameters c_1 and c_2. We first define the concept of weak efficiency.

Definition 3: Given two functions $f(\cdot) : \mathbb{R}^m \to \mathbb{R}$ and $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$, $\mathbf{x}^* \in \mathbb{R}^m$ is *weakly efficient* if at least one of the following three conditions holds:

1) $\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^m} f(\mathbf{x})$;
2) $\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^m} g(\mathbf{x})$;
3) $\mathbf{x}^*$ is Pareto efficient. That is, there exists no $\mathbf{x}'$ such that $f(\mathbf{x}') \leq f(\mathbf{x})$ and $g(\mathbf{x}') \leq g(\mathbf{x})$ with at least one strict inequality holds.

Indeed, a standard result in convex analysis states that the set of optimal solutions for the weighted sum of two convex functions coincides with the set of weakly efficient solutions.

Lemma 2: If $f(\mathbf{x})$ and $g(\mathbf{x})$ are both convex, then the solution set of

$$\min \lambda_1 f(\mathbf{x}) + \lambda_2 g(\mathbf{x})$$

where λ_1 and λ_2 range over $[0, +\infty)$ and cannot be zero simultaneously, is the set of weakly efficient solutions.

Proof: Observe that if $\mathbf{x}^*$ is an optimal solution for some nonzero λ_1 and λ_2, then it must be Pareto efficient. If $\mathbf{x}^*$ is an optimal solution for $\lambda_1 = 0$, then $\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^m} g(\mathbf{x})$. If $\mathbf{x}^*$ is an optimal solution for $\lambda_2 = 0$, then $\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^m} f(\mathbf{x})$. Thus, an optimal solution must be weakly efficient.

If $\mathbf{x}^*$ belongs to $\arg\max f(\mathbf{x})$, then it is optimal with respect to $\lambda_2 = 0$. Similarly for $\arg\max g(\mathbf{x})$. Now suppose that $\mathbf{x}^*$ is Pareto efficient. Since $f(\cdot)$ and $g(\cdot)$ are convex functions, we have the following set $\mathcal{T} \triangleq \{(a, b) | \exists \mathbf{x} \in \mathbb{R}^m : f(\mathbf{x}) \leq a; \& g(\mathbf{x}) \leq b\}$ is a convex set and observe that all Pareto efficient solutions are on the boundary. Given an arbitrary point $\mathbf{x}^*$ on the boundary, it follows from the supporting hyperplane theorem there exists a line (note that the set $\mathcal{T}$ is in 2-d space) that passes the $\mathbf{x}^*$ and is lower-left to $\mathcal{T}$ (note that $\mathcal{T}$ is unbounded up and right). In another word, there exists λ_1, λ_2 both nonnegative such that $\mathbf{x}^*$ minimizes $\lambda_1 f(\mathbf{x}) + \lambda_2 g(\mathbf{x})$. Combining the three conditions, we conclude that a weakly efficient solution is optimal to some tradeoff parameters. ∎

Thus, we have that the set of optimal solutions to $\min_{\mathbf{x}} ||\mathbf{b} - A\mathbf{x}||_2 + c_1||\mathbf{x}||_1$ where c_1 ranges over $[0, +\infty]$ is the set of weakly efficient solutions of $||\mathbf{b} - A\mathbf{x}||_2$ and $||\mathbf{x}||_1$. Similarly, the set of optimal solutions to $\min_{\mathbf{x}} ||\mathbf{b} - A\mathbf{x}||_2^2 + c_2||\mathbf{x}||_1$ where c_2 ranges over $[0, +\infty]$ is the set of weakly efficient solutions of $||\mathbf{b} - A\mathbf{x}||_2^2$ and $||\mathbf{x}||_1$. Further note that these two sets of weakly efficient solutions are identical, due to the fact that taking square of non-negative value is monotonic. We therefore conclude that these two formulations are identical up to change of the tradeoff parameters.

APPENDIX B
PROOF OF THEOREM 3

Theorem 3: Assume that the set

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

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:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^k_+, \boldsymbol{\kappa} \in \mathbb{R}^m_+, \mathbf{x} \in \mathbb{R}^m} \{\|\mathbf{b} - A\mathbf{x}\|_a + v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \mathbf{x})\};$$

$$\text{where: } v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \mathbf{x}) \triangleq \max_{\mathbf{c} \in \mathbf{R}^m} \left[(\boldsymbol{\kappa} + |\mathbf{x}|)^\top \mathbf{c} - \sum_{j=1}^k \lambda_j f_j(\mathbf{c}) \right].$$

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

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

Hence we have:

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

The second equation follows from Theorem 2.

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

$$\max_{\mathbf{c} \in \mathcal{Z}} \left\{ |\mathbf{x}^*|^\top \mathbf{c} \right\} = \min_{\boldsymbol{\lambda} \in \mathbb{R}^k_+, \boldsymbol{\kappa} \in \mathbb{R}^m_+} v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \mathbf{x}^*). \qquad (16)$$

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

APPENDIX C
PROOF OF THEOREM 4

Theorem 4: The robust regression problem

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}} \|\mathbf{b} - (A + \Delta A)\mathbf{x}\|_2 \right\} \qquad (17)$$

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

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}^I} \|\mathbf{b} - (A + \Delta \tilde{A} + \Delta A)\mathbf{x}\|_2 \right\} \qquad (18)$$

has a solution supported on the set I.

Proof: We show that Theorem 4 is a special case of Theorem 4'. To see this, we map the notations of Theorem 4 to the notations of Theorem 4'. To avoid conflict of notations, we use ' for the notations in Theorem 4'. For example A' is the matrix A in Theorem 4'. Given $\mathbf{b}$, A, $\mathcal{U}$ and I, we do the following conversion:

$$\begin{aligned}
&c'_i := c_i\ \forall\, i \in I;\ l'_j := c_j, c'_j := 0,\ \forall\, j \in I^c; \\
&\tilde{A}' := A;\ A' := A + \Delta \tilde{A};\ \mathbf{b}' := \mathbf{b}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\mathcal{U} &= \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c_i,\ \forall\, I\} \\
&= \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c'_i,\ i \in I; \\
&\qquad \|\boldsymbol{\delta}_j\|_2 \leq c'_j + l'_j,\ j \notin I\} \\
&= \tilde{\mathcal{U}}'; \\
\mathcal{U}^I &= \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c_i,\ i \in I;\ \|\boldsymbol{\delta}_j\|_2 = 0,\ j \notin I\} \\
&= \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c'_i,\ i \in I;\ \|\boldsymbol{\delta}_j\|_2 \leq c'_j,\ j \notin I\} \\
&= \mathcal{U}';
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{U}^{I^c} &= \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 = 0,\ i \in I;\ \|\boldsymbol{\delta}_j\|_2 \leq c_j,\ j \notin I\} \\
&= \{(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 = 0,\ i \in I;\ \|\boldsymbol{\delta}_j\|_2 \leq l'_j,\ j \notin I\}.
\end{aligned}$$

Thus, (10) is equivalent to (17), and (9) is equivalent to (18). Furthermore, $A' - \tilde{A}' = \Delta \tilde{A}$ implies that $\|\tilde{\mathbf{a}}'_j - \mathbf{a}'_j\| \leq l'_j$ for $j \notin I$, and $\tilde{\mathbf{a}}'_i = \mathbf{a}'_i$ for $i \in I$ due to the fact that $\Delta \tilde{A} \in \mathcal{U}^{I^c}$. Applying Theorem 4' completes the proof. ∎

APPENDIX D
PROOF OF PROPOSITION 1

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

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

The following holds

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

Proof: To prove Proposition 1, we first establish the following lemma.

Lemma 3: 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 ϵ-optimal solution to the LHS, 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

$$\sup_{\mathbf{x}' \in \mathcal{Z}} f(\mathbf{x}') \leq \sup_{\mu \in \mathcal{P} | \mu(\mathcal{Z})=1} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}). \quad (19)$$

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:

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

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

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

Combining (19) and (20), we prove the lemma. ∎

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

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i) \leq \sup_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^{m+1}} f(\mathbf{x}) d\mu(\mathbf{x}). \quad (21)$$

Without loss of generality, assume

$$f(\hat{\mathbf{x}}_1) \leq f(\hat{\mathbf{x}}_2) \leq \cdots \leq f(\hat{\mathbf{x}}_n). \quad (22)$$

Now construct the following function:

$$\hat{f}(\mathbf{x}) \triangleq \begin{cases} \min_{i | \mathbf{x} \in \mathcal{Z}_i} f(\hat{\mathbf{x}}_i) & \mathbf{x} \in \bigcup_{j=1}^{n} \mathcal{Z}_j; \\ f(\mathbf{x}) & \text{otherwise.} \end{cases} \quad (23)$$

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

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

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

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

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

Hence by (22) 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 (21). ∎

Appendix E
Proof of Corollary 3

Corollary 3: Given $\mathbf{b} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$, the following equation holds for any $\mathbf{x} \in \mathbb{R}^m$:

$$\begin{aligned} &\|\mathbf{b} - A\mathbf{x}\|_2 + \sqrt{n} c_n \|\mathbf{x}\|_1 + \sqrt{n} c_n \\ &\qquad = \sup_{\mu \in \hat{\mathcal{P}}(n)} \sqrt{n \int_{\mathbb{R}^{m+1}} (b' - \mathbf{r}'^{\top} \mathbf{x})^2 d\mu(\mathbf{r}', b')}. \end{aligned} \quad (24)$$

Here,

$$\begin{aligned} \hat{\mathcal{P}}(n) &\triangleq \bigcup_{\|\boldsymbol{\sigma}\|_2 \leq \sqrt{n} c_n;\ \forall\, i : \|\boldsymbol{\delta}_i\|_2 \leq \sqrt{n} c_n} \mathcal{P}_n(A, \Delta, \mathbf{b}, \boldsymbol{\sigma}); \\ &\mathcal{P}_n(A, \Delta, \mathbf{b}, \boldsymbol{\sigma}) \\ &\triangleq \{\mu \in \mathcal{P} | \mathcal{Z}_i = [b_i - \sigma_i, b_i + \sigma_i] \\ &\qquad \times \prod_{j=1}^{m} [a_{ij} - \delta_{ij}, a_{ij} + \delta_{ij}]; \\ &\forall\, S \subseteq \{1, \ldots, n\} : \mu\left(\bigcup_{i \in S} \mathcal{Z}_i \right) \geq |S|/n\}. \end{aligned}$$

$$\max_{\|\boldsymbol{\sigma}\|_2\le\sqrt{n}c_n;\forall\, i:\|\boldsymbol{\delta}_i\|_2\le\sqrt{n}c_n} \|\mathbf{b}+\boldsymbol{\sigma}-(A+[\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m])\,\mathbf{x}\|_2$$
$$= \sup_{\|\boldsymbol{\sigma}\|_2\le\sqrt{n}c_n;\ \forall\, i:\|\boldsymbol{\delta}_i\|_2\le\sqrt{n}c_n} \left\{ \sup_{(\hat{b}_i,\hat{\mathbf{r}}_i)\in[b_i-\sigma_i,b_i+\sigma_i]\times\prod_{j=1}^m[a_{ij}-\delta_{ij},a_{ij}+\delta_{ij}]} \sqrt{\sum_{i=1}^n(\hat{b}_i-\hat{\mathbf{r}}_i^\top\mathbf{x})^2} \right\}$$
$$= \sup_{\|\boldsymbol{\sigma}\|_2\le\sqrt{n}c_n;\ \forall\, i:\|\boldsymbol{\delta}_i\|_2\le\sqrt{n}c_n} \sqrt{\sum_{i=1}^n \sup_{(\hat{b}_i,\hat{\mathbf{r}}_i)\in[b_i-\sigma_i,b_i+\sigma_i]\times\prod_{j=1}^m[a_{ij}-\delta_{ij},a_{ij}+\delta_{ij}]} (\hat{b}_i-\hat{\mathbf{r}}_i^\top\mathbf{x})^2}$$

Proof: The RHS of (24) equals

$$\sup_{\|\boldsymbol{\sigma}\|_2\le\sqrt{n}c_n;\ \forall\, i:\|\boldsymbol{\delta}_i\|_2\le\sqrt{n}c_n} \left\{ \sup_{\mu\in\mathcal{P}_n(A,\Delta,\mathbf{b},\boldsymbol{\sigma})} \sqrt{n\int_{\mathbb{R}^{m+1}}(b'-\mathbf{r}'^\top\mathbf{x})^2 d\mu(\mathbf{r}',b')} \right\}$$

Notice by the equivalence to robust formulation, the LHS equals the equation shown at the top of the page, furthermore, applying Proposition 1 yields

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

which proves the corollary. ■

REFERENCES

[1] L. Elden, "Perturbation theory for the least square problem with linear equality constraints," *SIAM J. Numer. Anal.*, vol. 17, no. 3, pp. 338–350, 1980.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1989.

[3] D. J. Higham and N. J. Higham, "Backward error and condition of structured linear systems," *SIAM J. Matrix Anal. Appl.*, vol. 13, pp. 162–175, 1992.

[4] R. D. Fierro and J. R. Bunch, "Collinearity and total least squares," *SIAM J. Matrix Anal. Appl.*, vol. 15, pp. 1167–1181, 1994.

[5] A. N. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*. New York: Wiley, 1977.

[6] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Royal Statist. Soc., Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientif. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[9] A. Feuer and A. Nemirovski, "On sparse representation in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1579–1581, 2003.

[10] E. J. Candés, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, pp. 489–509, 2006.

[11] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231–2242, 2004.

[12] M. J. Wainwright, Sharp Thresholds for Noisy and High-Dimensional Recovery of Sparsity Using ℓ_1-Constrained Quadratic Programming Dep. Statist., UC Berkeley, 2006 [Online]. Available: http://www.stat.berkeley.edu/tech-reports/709.pdf

[13] L. E. Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM J. Matrix Anal. Appl.*, vol. 18, pp. 1035–1064, 1997.

[14] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *J. Mach. Learn. Res.*, vol. 7, pp. 1283–1314, Jul. 2006.

[15] H. Scarf, "A min-max solution of an inventory problem," in *Studies in Mathematical Theory of Inventory and Production*. Stanford, CA: Stanford Univ. Press, 1958, pp. 201–209.

[16] P. Kall, "Stochastic programming with recourse: Upper bounds and moment problems, a review," in *Advances in Mathematical Optimization*. Berlin, Germany: Academie-Verlag, 1988.

[17] I. Popescu, "Robust mean-covariance solutions for stochastic optimization," *Operat. Res.*, vol. 55, no. 1, pp. 98–112, 2007.

[18] H. Xu, "Robust decision making and its applications to machine learning," Ph.D., McGill Univ., Montreal, QC, Canada, 2009.

[19] A. Ben-Tal, L. G. El, and A. Nemirovski, *Robust Optimization*. Princeton, NJ: Princeton Univ. Press, 2009.

[20] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Statist. Soc. Ser. B*, vol. 67, no. 2, pp. 301–320, 2005.

[21] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1030–1051, 2006.

[22] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Computat.*, vol. 10, no. 6, pp. 1445–1480, 1998.

[23] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, pp. 713–718, 1992.

[24] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, pp. 3397–3415, 1993.

[25] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, 2006.

[26] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, pp. 1341–1344, 2004.

[27] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, pp. 832–837, 1956.

[28] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[29] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The l_1 View*. New York: Wiley , 1985.

[30] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley, 1992.

[31] V. N. Vapnik and A. Chervonenkis, "The necessary and sufficient conditions for consistency in the empirical risk minimization method," *Pattern Recogn. Image Anal.*, vol. 1, no. 3, pp. 260–284, 1991.

[32] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, 2002.

[33] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Mach. Learn. Res.*, vol. 10, no. Jul., pp. 1485–1510, 2009.

[34] A. W. van der vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 2000.
[35] H. Xu, S. Mannor, and C. Caramanis, "Sparse algorithms are not stable: A no-free-lunch theorem," in *Proc. 46th Allerton Conf. Commun., Control, Comput.*, 2008, pp. 1299–1303.

Huan Xu received the B.Eng. degree in automation from Shanghai Jiaotong University, Shanghai, China, in 1997, the M.Eng. degree in electrical engineering from the National University of Singapore in 2003, and the Ph.D. degree in electrical engineering from McGill University, Canada, in 2009.

Since 2009, he has been a Postdoctoral Associate with the Wireless Networking and Communications Group (WNCG), The University of Texas at Austin. His research interests include statistics, machine learning, robust optimization, and decision making and control under uncertainty.

Constantine Caramanis (M'06) received the A.B. degree in mathematics from Harvard University, Cambridge, MA, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT).

Since 2006, he has been on the faculty in Electrical and Computer Engineering, The University of Texas at Austin. His research interests include robust and adaptable optimization, combinatorial optimization, statistics, machine learning and control, with applications to large-scale networks.

Shie Mannor (S'00–M'03–SM'09) received the B.Sc. degree in electrical engineering, the B.A. degree in mathematics, and the Ph.D. degree in electrical engineering, all from the Technion-Israel Institute of Technology, Haifa, in 1996, 1996, and 2002, respectively.

From 2002 to 2004, he was a Fulbright scholar and a Postdoctoral associate with the Massachusetts Institute of Technology (MIT). From 2004–2009, he was with the Department of Electrical Engineering, McGill University, Canada, where he was a Canada Research chair in Machine Learning. He has been a Horev Fellow and an Associate Professor with the Faculty of Electrical Engineering, Technion, since 2008. His research interests include machine learning and pattern recognition, planning and control, multi-agent systems, and communications.