On Robust Estimation of High Dimensional Generalized Linear Models

Eunho Yang
Department of Computer Science
University of Texas, Austin
eunho@cs.utexas.edu

Ambuj Tewari
Department of Statistics
University of Michigan, Ann Arbor
tewaria@umich.edu

Pradeep Ravikumar
Department of Computer Science
University of Texas, Austin
pradeepr@cs.utexas.edu

Abstract

We study *robust high-dimensional estimation* of generalized linear models (GLMs); where a small number k of the n observations can be arbitrarily corrupted, and where the true parameter is high dimensional in the "$p \gg n$" regime, but only has a small number s of non-zero entries. There has been some recent work connecting robustness and sparsity, in the context of linear regression with corrupted observations, by using an explicitly modeled outlier response vector that is assumed to be sparse. Interestingly, we show, in the GLM setting, such explicit outlier response modeling can be performed in *two distinct ways*. For each of these two approaches, we give ℓ_2 error bounds for parameter estimation for general values of the tuple (n, p, s, k).

1 Introduction

Statistical models in machine learning allows us to make strong predictions even from limited data, by leveraging specific assumptions imposed on the model space. However, on the flip side, when the specific model assumptions do not exactly hold, these standard methods could deteriorate severely. Constructing estimators that are *robust* to such departures from model assumptions is thus an important problem, and forms the main focus of Robust Statistics [Huber, 1981; Hampel *et al.*, 1986; Maronna *et al.*, 2006]. In this paper, we focus on the robust estimation of high-dimensional generalized linear models (GLMs). GLMs are a very general class of models for predicting a response given a covariate vector, and include many classical conditional distributions such as Gaussian, logistic, etc. In classical GLMs, the data points are typically low dimensional and are all assumed to be actually drawn from the assumed model. In our setting of high dimensional robust GLMs, there are two caveats: (a) the true parameter vector can be very high dimensional and furthermore, (b) certain observations are *outliers*, and could have arbitrary values with no quantitative relationship to the assumed generalized linear model.

Existing Research: Robust Statistics. There has been a long line of work [Huber, 1981; Hampel *et al.*, 1986; Maronna *et al.*, 2006] on robust statistical estimators. These are based on the insight that the typical log-likelihood losses, such as the squared loss for the ordinary least squares estimator for linear regression, are very sensitive to outliers, and that one could devise surrogate losses instead that are more resistant to such outliers. Rousseeuw [1984] for instance proposed the least median estimator as a robust variant of the ordinary least squares estimator. Another class of approaches fit trimmed estimators to the data after an initial removal of candidate outliers Rousseeuw and Leroy [1987]. There have also been estimators that model the outliers explicitly. [Gelman *et al.*, 2003] for instance model the responses using a mixture of two Gaussian distributions: one for the regular noise, and the other for the outlier noise, typically modeled as a Gaussian with high variance. Another instance is where the outliers are modeled as being drawn from heavy-tailed distributions such as the t distribution [Lange *et al.*, 1989].

Existing Research: Robustness and Sparsity. The past few years have actually led to an understanding that outlier robust estimation is intimately connected to *sparse signal recovery* [Candes and Tao, 2005; Antoniadis, 2007; Jin and Rao, 2010; Mitra *et al.*, 2010; She and Owen, 2011]. The main insight here is that if the number of outliers is small, it could be cast as a sparse error vector that is added to the standard noise. The problem of sparse signal recovery itself has seen a surge of recent research, and where a large body of work has shown that convex and tractable methods employing the likes of ℓ_1 regularization enjoy strong statistical guarantees [Donoho and Elad, 2003; Ng, 2004; Candes and Tao, 2006; Meinshausen and Bühlmann, 2006; Tropp, 2006; Zhao and Yu, 2007; Wainwright, 2009; Yang *et al.*, 2012]. Intriguingly, Antoniadis [2007]; She and Owen [2011] show that even classical robust statistics methods could be cast as sparsity encouraging M-estimators that specifically use non-convex regularization. Jin and Rao [2010]; Mitra *et al.* [2010] have also suggested the use of non-convex penalization based methods such as SCAD [Fan and Li, 2001] for robust statistical estimation. Convex regularization based estimators however have been enormously successful in high-dimensional statistical estimation, and in particular provide tractable methods that scale to very high-dimensional problems, and yet come with rigorous guarantees. To complete the story on the connection between robustness and sparsity, it is thus vital to obtain bounds on the performance of the convex regularization based estimators for general high-dimensional robust estimation. For the task of high dimensional robust *linear* regression, there has been some interesting recent work [Nguyen and Tran, 2011] that have provided precisely such bounds. In this paper, we provide such an analysis for GLMs beyond the standard Gaussian linear model.

It turns out that the story for robust GLMs beyond the standard Gaussian linear model is more complicated. In particular, outlier modeling in GLMs could be done in two ways: (a) in the *parameter space* of the GLM, or (b) in the *output space*. For the linear model these two approaches are *equivalent*, but significant differences emerge in the general case. We show that the former approach always leads to convex optimization problems but only works under rather stringent conditions. On the other hand, the latter approach can lead to a non-convex M-estimator, but enjoys better guarantees. However, we show that all global minimizers of the M-estimation problem arising in the second approach are close to each other, so that the non-convexity in the problem is rather benign. Leveraging recent results [Agarwal *et al.*, 2010; Loh and Wainwright, 2011], we can then show that projected gradient descent will approach one of the global minimizers up to an additive error that scales with the statistical precision of the problem. Our main contributions are thus as follows:

1. For robust estimation of GLMs, we show that there are *two distinct ways* to use the connection between robustness and sparsity.
2. For each of these two distinct approaches, we provide M-estimators, that use ℓ_1 regularization, and *in addition*, appropriate constraints. For the first approach, the M-estimation problem is convex and tractable. For the second approach, the M-estimation problem is typically *non-convex*. But we provide a projected gradient descent algorithm that is *guaranteed* to converge to a global minimum of the corresponding M-estimation problem, up to an additive error that scales with the statistical precision of the problem.
3. One of the main contributions of the paper is to provide ℓ_2 error bounds for each of the two M-estimators, for general values of the tuple (n, p, s, k). The analysis of corrupted general non-linear models in high-dimensional regimes is highly non-trivial: it combines the twin difficulties of high-dimensional analysis of non-linear models, and analysis given corrupted observations. The presence of both these elements, specifically the interactions therein, required a subtler analysis, as well as slightly modified M-estimators.

2 Problem Statement and Setup

We consider generalized linear models (GLMs) where the response variable has an exponential family distribution, conditioned on the covariate vector,

$$\mathbb{P}(y|x, \theta^\star) = \exp\left\{ \frac{h(y) + y\langle\theta^\star, x\rangle - A\big(\langle\theta^\star, x\rangle\big)}{c(\sigma)} \right\}. \quad (1)$$

Examples. The standard linear model with Gaussian noise, the logistic regression and the Poisson model are typical examples of this model. In case of standard linear model, the domain of variable y, $\mathcal{Y}$, is the set of real numbers, $\mathbb{R}$, and with known scale parameter σ, the probability of y in (1) can be rewritten as

$$\mathbb{P}(y|x, \theta^\star) \propto \exp\left\{ \frac{-y^2/2 + y\langle\theta^\star, x\rangle - \langle\theta^\star, x\rangle^2/2}{\sigma^2} \right\}, \quad (2)$$

where the normalization function $A(a)$ in (1) in this case becomes $a^2/2$. Another very popular example in GLM models is logistic regression given a categorical output variable:

$$\mathbb{P}(y|x, \theta^\star) = \exp\left\{ y\langle\theta^\star, x\rangle - \log\left(1 + \exp(\langle\theta^\star, x\rangle)\right) \right\}, \quad (3)$$

where $\mathcal{Y}$ is $\{0, 1\}$, and the normalization function $A(a) = \log(1 + \exp(a))$. We can also derive the Poisson regression model from (1) as follows:

$$\mathbb{P}(y|x, \theta^\star) = \exp\left\{ -\log(y!) + y\langle\theta^\star, x\rangle - \exp(\langle\theta^\star, x\rangle) \right\}, \quad (4)$$

where $\mathcal{Y}$ is $\{0, 1, 2, ...\}$, and the normalization function $A(a) = \exp(a)$. Our final example is the case where the variable y follows an exponential distribution:

$$\mathbb{P}(y|x, \theta^\star) = \exp\left\{ y\langle\theta^\star, x\rangle + \log(-\langle\theta^\star, x\rangle) \right\}, \quad (5)$$

where $\mathcal{Y}$ is the set of non-negative real numbers, and the normalization function $A(a) = -\log(-a)$. Any distribution in the exponential family can be written as the GLM form (1) where the canonical parameter of exponential family is $\langle\theta^\star, x\rangle$. Note however that some distributions such as Poisson or exponential place restrictions on $\langle\theta^\star, x\rangle$ to be valid parameter, so that the density is normalizable, or equivalently the normalization function $A(\langle\theta^\star, x\rangle) < +\infty$.

In the GLM setting, suppose that we are given n covariate vectors, $x_i \in \mathbb{R}^p$, drawn i.i.d. from some distribution, and corresponding response variables, $y_i \in \mathcal{Y}$, drawn from the distribution $\mathbb{P}(y|x_i, \theta^\star)$ in (1). A key goal in statistical estimation is to estimate the parameters θ^*, given just the samples $Z_1^n := \{(x_i, y_i)\}_{i=1}^n$. Such estimation becomes particularly challenging in a *high-dimensional* regime, where the dimension p is potentially even larger than the number of samples n. In this paper, we are interested in such high dimensional parameter estimation of a GLM under the additional caveat that some of the observations y_i are *arbitrarily* corrupted. We can model such corruptions by adding an "outlier error" parameter $e_i^\star$ in two ways: (i) we consider $e_i^\star$ in the "parameter space" to the uncorrupted parameter $\langle\theta^\star, x_i\rangle$, or (ii) introduce $e_i^\star$ in the output space, so that the output y_i is actually the sum of $e_i^\star$ and the uncorrupted output $\bar{y}_i$. For the specific case of the linear model, both these approaches are exactly the same. We assume that only some of the examples are corrupted, which translates to the error vector $e^\star \in \mathbb{R}^n$ being sparse. We further assume that the parameter $\theta^\star$ is also sparse. We thus assume:

$$\|\theta^\star\|_0 \leq s, \quad \text{and} \quad \|e^\star\|_0 \leq k,$$

with support sets S and T, respectively. We detail the two approaches (modeling outlier errors in the parameter space and output space respectively) in the next two sections.

3 Modeling gross errors in the parameter space

In this section, we discuss a robust estimation approach by modeling gross outlier errors in the parameter space. Specifically, we assume that the i-th response y_i is drawn from the conditional distribution in

(1) but with a "corrupted" parameter $\langle \theta^\star, x_i \rangle + \sqrt{n} e_i^\star$, so that the samples are distributed as $\mathbb{P}(y_i | x_i, \theta^\star, e_i^\star) = \exp\left\{ \frac{h(y_i) + y_i(\langle \theta^\star, x_i \rangle + \sqrt{n} e_i^\star) - A\left(\langle \theta^\star, x_i \rangle + \sqrt{n} e_i^\star\right)}{c(\sigma)} \right\}$. We can then write down the resulting negative log-likelihood as,

$$\mathcal{L}_p(\theta, e; Z_1^n) :=$$

$$-\langle \theta, \frac{1}{n} \sum_{i=1}^n y_i x_i \rangle - \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i e_i + \frac{1}{n} \sum_{i=1}^n A\big(\langle \theta, x_i \rangle + \sqrt{n} e_i\big).$$

We thus arrive at the following ℓ_1 regularized maximum likelihood estimator:

$$(\widehat{\theta}, \widehat{e}) \in \operatorname{argmin} \mathcal{L}_p(\theta, e; Z_1^n) + \lambda_{n,\theta} \|\theta\|_1 + \lambda_{n,e} \|e\|_1.$$

In the sequel, we will consider the following *constrained* version of the MLE (a_0, b_0 are constants independent of n, p):

$$(\widehat{\theta}, \widehat{e}) \in \operatorname*{argmin}_{\substack{\|\theta\|_2 \le a_0 \\ \|e\|_2 \le \frac{b_0}{\sqrt{n}}}} \mathcal{L}_p(\theta, e; Z_1^n) + \lambda_{n,\theta} \|\theta\|_1 + \lambda_{n,e} \|e\|_1. \quad (6)$$

The additional regularization provided by the constraints allow us to obtain tighter bounds for the resulting M-estimator.

We note that the M-estimation problem in (6) is *convex*: adding the outlier variables e does not destroy the convexity of the original problem with no gross errors. On the other hand, as we detail below, extremely stringent conditions are required for consistent statistical estimation. (The strictness of these conditions is what motivated us to also consider output space gross errors in the next section, where the conditions required are more benign).

3.1 ℓ_2 Error Bound

We require the following stringent condition:

Assumption 1. $\|\theta^\star\|_2 \le a_0$ *and* $\|e^\star\|_2 \le \frac{b_0}{\sqrt{n}}$.

We assume the covariates are multivariate Gaussian distributed as described in the following condition:

Assumption 2. *Let X be the $n \times p$ design matrix, with the n samples $\{x_i\}$ along the n rows. We assume that each sample x_i is independently drawn from $N(0, \Sigma)$. Let $\lambda_{\max}$ and $\lambda_{\min} > 0$ be the maximum and minimum eigenvalues of the covariance matrix Σ, respectively, and let ξ be the maximum diagonal entry of Σ. We assume that $\xi \lambda_{\max} = \Theta(1)$.*

Additionally, we place a mild restriction on the normalization function $A(\cdot)$ that all examples of GLMs in Section 2 satisfy:

Assumption 3. *The double-derivative $A''(\cdot)$ of the normalization function has at most exponential decay: $A''(\eta) \ge \exp(-c\eta)$ for some $c > 0$.*

Theorem 1. *Consider the optimal solution $(\widehat{\theta}, \widehat{e})$ of* (6) *with the regularization parameters:*

$$\lambda_{n,\theta} = 2c_1 \sqrt{\frac{\log p}{n}} \quad \textit{and} \quad \lambda_{n,e} = 2c_2 \sqrt{\frac{\log n}{n}},$$

where c_1 and c_2 are some known constants. Then, there exist positive constants K, c_3 and c_4 such that with probability at least $1 - \frac{K}{n}$, the error $(\widehat{\Delta}, \widehat{\Gamma}) := (\widehat{\theta} - \theta^\star, \widehat{e} - e^\star)$ is bounded by

$$\|\widehat{\Delta}\|_2 + \|\widehat{\Gamma}\|_2 \le c_3 \frac{\sqrt{s \log p} + \sqrt{k \log n}}{n^{1/2 - c_4/\sqrt{\log n}}}.$$

Note that the theorem requires the assumption that the outlier errors are bounded as $\|e^\star\|_2 \le \frac{b_0}{\sqrt{n}}$. Since the "corrupted" parameters are given by $\langle \theta^\star, x_i \rangle + \sqrt{n} e_i^\star$, the gross outlier error scales as $\sqrt{n} \|e^\star\|_2$, which the assumption thus entails be bounded by a constant (independent of n). Our search to find a method that can tolerate larger gross errors led us to introduce the error in the output space in the next section.

4 Modeling gross errors in the output space

In this section, we investigate the consequences of modeling the gross outlier errors directly in the response space. Specifically, we assume that a perturbation of the i-th response, $y_i - \sqrt{n} e_i^\star$ is drawn from the conditional distribution in (1) with parameter $\langle \theta^\star, x_i \rangle$, so that the samples are distributed as $\mathbb{P}(y_i | x_i, \theta^\star, e_i^\star) = \exp\left\{ \frac{h(y_i - \sqrt{n} e_i^\star) + (y_i - \sqrt{n} e_i^\star) \langle \theta^\star, x_i \rangle - A\left(\langle \theta^\star, x_i \rangle\right)}{c(\sigma)} \right\}$. We can then write down the resulting likelihood as, $\mathcal{L}_o(\theta, e; Z_1^n) := \frac{1}{n} \sum_{i=1}^n \Big[B\left(y_i - \sqrt{n} e_i\right) - (y_i - \sqrt{n} e_i) \langle \theta, x_i \rangle + A(\langle \theta, x_i \rangle) \Big]$, where $B(y) = -h(y)$, and the resulting ℓ_1 regularized maximum likelihood estimator as:

$$(\widehat{\theta}, \widehat{e}) \in \operatorname{argmin} \mathcal{L}_o(\theta, e; Z_1^n) + \lambda_{n,\theta} \|\theta\|_1 + \lambda_{n,e} \|e\|_1. \quad (7)$$

Note that when $B(y)$ is set to $-h(y)$ as above, the estimator has the natural interpretation of maximizing regularized log-likelihood, but in the sequel we allow it to be an arbitrary function taking the response variable as an input argument. As we will see, setting this to a function other than $-h(y)$ will allow us to obtain *stronger statistical guarantees*.

Just as in the previous section, we consider a constrained version of the MLE in the sequel:

$$(\widehat{\theta}, \widehat{e}) \in \operatorname*{argmin}_{\substack{\|\theta\|_1 \le a_0 \sqrt{s} \\ \|e\|_1 \le b_0 \sqrt{k}}} \mathcal{L}_o(\theta, e; Z_1^n) + \lambda_{n,\theta} \|\theta\|_1 + \lambda_{n,e} \|e\|_1. \quad (8)$$

A key reason we introduce these constraints will be seen in the next section: these constraints help in designing an efficient iterative optimization algorithm to solve the above optimization problem (by providing bounds on the iterates right from the first iteration). One unfortunate facet of the M-estimation problem in (8), and the allied problem in (7), is that it is not convex in general. We will nonetheless show that the computationally tractable algorithm we provide in the next section is guaranteed to converge to a global optimum (up to an additive error that scales at most with the statistical error of the global optimum).

We require the following bounds on the ℓ_2 norms of $\theta^\star, e^\star$.

Assumption 4. $\|\theta^\star\|_2 \le a_0$ *and* $\|e^\star\|_2 \le b_0$ *for some constants* a_0, b_0.

When compared with Assumption 1 in the previous section, the above assumption imposes a much weaker restriction on the magnitude of the gross errors. Specifically, with the $\sqrt{n}$ scaling included, the above Assumption 4 allows the ℓ_2 norm of the gross error to scale as $\sqrt{n}$, whereas Assumption 1 in the previous section required the norm $\|\theta^\star\|_2$ to be bounded above by a constant.

4.1 ℓ_2 Error bound

It turned out, given our analysis, that a natural selection of the function $B(\cdot)$ is to use the quadratic function (we defer discussion due to lack of space). Thus, in the spirit of classical robust statistics, we considered the modified log-likelihood objective in (7) with the above setting of $B(\cdot)$: $\mathcal{L}_o(\theta, e; Z_1^n) := \frac{1}{n}\sum_{i=1}^n \left[\frac{1}{2}\left(y_i - \sqrt{n}e_i\right)^2 - (y_i - \sqrt{n}e_i)\langle\theta, x_i\rangle + A(\langle\theta, x_i\rangle)\right]$.

Similarly here, we assume the random design matrix has rows sampled from a sub-Gaussian distribution:

Assumption 5. *Let X be the $n \times p$ design matrix which has each sample x_i in its ith row. Let $\lambda_{\max}$ and $\lambda_{\min} > 0$ be the maximum and minimum eigenvalues of the covariance matrix of x, respectively. For any $v \in \mathbb{R}^p$, the variable $\langle\theta, x_i\rangle$ is sub-Gaussian with parameter at most $\kappa_u^2\|v\|_2^2$.*

Theorem 2. *Consider the optimal solution $(\widehat{\theta}, \widehat{e})$ of* (8) *with the regularization parameters:*

$$\lambda_{n,\theta} = \max\left\{2c_1\sqrt{\frac{\log p}{n}}, c_2\sqrt{\frac{\max(s,k)\log p}{sn}}\right\} \quad \textit{and}$$

$$\lambda_{n,e} = \max\left\{\frac{2}{c'' n^{1/2 - \frac{c'}{\sqrt{\log n}}}}, c_3\sqrt{\frac{\max(s,k)\log p}{kn}}\right\},$$

where c', c'', c_1, c_2, c_3 are some known constants. Then, there exist positive constants K, L and c_4 such that if $n \geq L\max(s,k)\log p$, then with probability at least $1 - n^K$, the error $(\widehat{\Delta}, \widehat{\Gamma}) := (\widehat{\theta} - \theta^\star, \widehat{e} - e^\star)$ is bounded by

$$\|\widehat{\Delta}\|_2 + \|\widehat{\Gamma}\|_2 \leq\ c_4 \max\left\{\frac{\sqrt{k}}{n^{\frac{1}{2} - \frac{c'}{\sqrt{\log n}}}}, \sqrt{\frac{\max(s,k)\log p}{n}}\right\}.$$

Remarks. Nguyen and Tran [2011] analyze the specific case of the standard linear regression model (which nonetheless is a member of the GLM family), and provide the bound:

$$\|\widehat{\Delta}\|_2 + \|\widehat{\Gamma}\|_2 \leq c\max\left\{\sqrt{\frac{s\log p}{n}}, \sqrt{\frac{k\log n}{n}}\right\},$$

which is asymptotically equivalent to the bound in Theorem 2. As we noted earlier, for the linear regression model, both approaches of modeling outlier errors in the parameter space or the output space are equivalent, so that we could also compare the linear regression bound to our bound in Theorem 1.There too, the bounds can be seen to be asymptotically equivalent. We thus see that the generality of the GLM family does not adversely affect ℓ_2 norm convergence rates even when restricted to the simple linear regression model.

5 A Tractable Optimization Method for the Output Space Modeling Approach

In this section we focus on the M-estimation problem (8) that arises in the second approach where we model errors in the output space. Unfortunately, this is not a tractable optimization problem: in particular, the presence of the bilinear term $e_i\langle\theta, x_i\rangle$ makes the objective function $\mathcal{L}_o$ non-convex. A tractable seemingly-approximate method would be to solve for a local minimum of the objective, by using a gradient descent based method. In particular, projected gradient descent (PGD) applied to the M-estimation problem (8) produces the iterates:

$$(\theta^{t+1}, e^{t+1}) \leftarrow \underset{\substack{\|\theta\|_1 \leq a_0\sqrt{s} \\ \|e\|_1 \leq b_0\sqrt{k}}}{\operatorname{argmin}} \Big\{\langle\theta, \nabla_\theta\mathcal{L}_o(\theta^t, e^t; Z_1^n)\rangle + \frac{\eta}{2}\|\theta - \theta^t\|_2^2$$
$$+ \langle e, \nabla_e\mathcal{L}_o(\theta^t, e^t, Z_1^n)\rangle + \frac{\eta}{2}\|e - e^t\|_2^2 + \lambda_{n,\theta}\|\theta\|_1 + \lambda_{n,e}\|e\|_1\Big\},$$

where $\eta > 0$ is a step-size parameter. Note that even though $\mathcal{L}_o$ is non-convex, the problem above is convex, and decouples in θ, e. Moreover, minimizing a composite objective over the ℓ_1 ball can be solved very efficiently by performing two projections onto the ℓ_1-ball (see Agarwal *et al.* [2010] for instance for details).

While the projected gradient descent algorithm with the iterates above might be tractable, one concern might be that these iterates would atmost converge to a local minimum, which might not satisfy the consistency and ℓ_2 convergence rates as outlined in Theorem 2. However, the following theorem shows that the concern is unwarranted: the iterates converge to a global minimum of the optimization problem in (8), up to an additive error that scales at most as the *statistical error*, $\left(\|\widehat{\theta} - \theta^\star\|_2^2 + \|\widehat{e} - e^\star\|_2^2\right)$.

Theorem 3. *Suppose all conditions of Theorem 2 hold and that $n > c_0^2(k+s)^2\log(p)$. Let $F(\theta, e)$ denote the objective function in* (8) *and let $(\widehat{\theta}, \widehat{e})$ be a global optimum of the problem. Then, when we apply the PGD steps above with appropriate step-size η, there exist universal constants $C_1, C_2 > 0$ and a contraction coefficient $\gamma < 1$, independent of (n, p, s, k), such that $\|\theta^t - \widehat{\theta}\|_2^2 + \|e^t - \widehat{e}\|_2^2 \leq \underbrace{C_1\left(\|\widehat{\theta} - \theta^\star\|_2^2 + \|\widehat{e} - e^\star\|_2^2\right)}_{\delta^2}$ for all iterates $t \geq T$ where $T = C_2\log\frac{F(\theta^0, e^0) - F(\widehat{\theta}, \widehat{e})}{\delta^2} / \log(1/\gamma)$.*

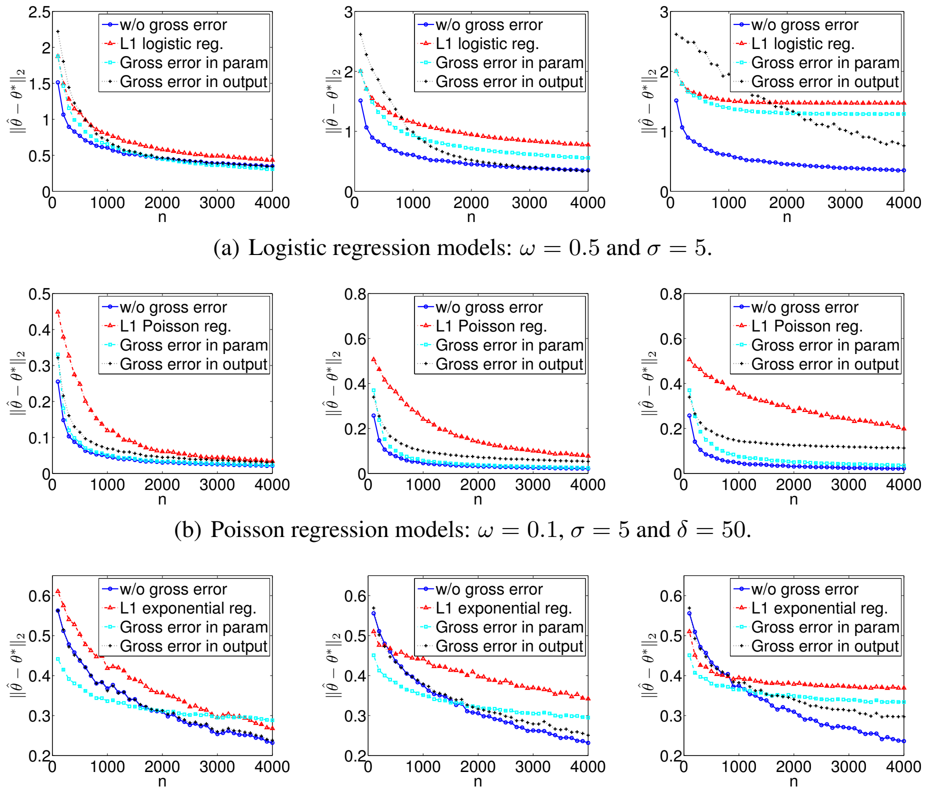
6 Experimental Results

In this section, we provide experimental validation, over both simulated as well as real data, of the performance of our M-estimators.

6.1 Simulation Studies

In this section, we provide simulations corroborating Theorems 1 and 2. The theorems are applicable to any distribution in the GLM family (1), and as canonical instances, we consider the cases of logistic regression (3), Poisson regression (4), and exponential regression (5). (The case of the standard linear regression model under gross errors has been previously considered in Nguyen and Tran [2011].)

We instantiated our models as follows. We first randomly selected a subset S of $\{1, \ldots, p\}$ of size $\sqrt{p}$ as the support set (indexing non-zero values) of the true parameter θ^*. We then set the nonzero elements, θ_S^*, to be equal to ω, which we vary as noted in the plots. We then randomly generated n i.i.d samples, $\{x_1, ..., x_n\}$, from the normal distribution $N(0, \sigma^2 I_{p\times p})$. Given each feature vector x_i, we drew



(a) Logistic regression models: $\omega = 0.5$ and $\sigma = 5$.

(b) Poisson regression models: $\omega = 0.1$, $\sigma = 5$ and $\delta = 50$.

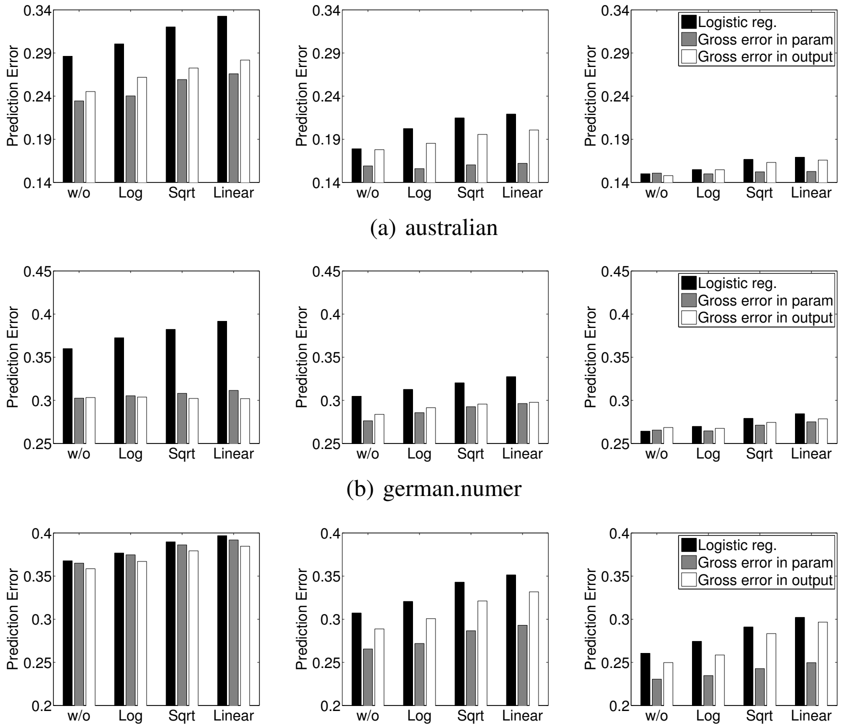
(c) Exponential regression models: $\omega = 0.1$, $\sigma = 1$ and $\delta = 20$.

Figure 1: Comparisons of ℓ_2 error norm for θ vs. n, and where $p = 196$, for different regression cases: logistic regression (top row), Poisson regression (middle row), and exponential regression (bottom row). Three different types of corruptions are presented: $k = \log(n)$ (Left column), $k = \sqrt{(n)}$ (Center), and $k = 0.1(n)$ (Right column).

the corresponding true class label $\bar{y}_i$ from the corresponding GLM distribution. To simulate the worst instance of gross errors, we selected the k samples with the highest value of $\langle \theta^*, x_i \rangle$ and corrupted them as follows. For logistic regression, we just flipped their class labels, to $y_i = (1 - \bar{y}_i)$. For the Poisson and exponential regression models, the corrupted response y_i is obtained by adding a gross error term δ_i to $\bar{y}_i$. The learning algorithms were then given the *corrupted dataset* $\{x_i, y_i\}_{i=1}^n$. We scaled the number of corrupted samples k with the total number of samples n in three different ways: logarithmic scaling with $k = \Theta(\log n)$, square root scaling with $k = \Theta(\sqrt{n})$, and linear scaling with $k = \Theta(n)$. For each tuple of (n, p, s, k), we drew 50 batches of n samples, and plot the average.

Figure 1 plots the ℓ_2 norm error $\|\hat{\theta} - \theta^*\|_2$ of the parameter estimates, against the number of samples n. We compare three methods: (a) the standard ℓ_1 penalized GLM MLE (e.g. ℓ_1 logistic reg.), which directly learns a GLM regression model over the corrupted data; (b) our first M-estimator (6), which models error in the parameter space (Gross error in param); and (c) our second M-estimator, which models error in the output space (8) (Gross error in output). As a gold standard, we also include the performance of the standard ℓ_1 penalized GLM regression *over the uncorrupted version of the dataset*, $\{x_i, \bar{y}_i\}_{i=1}^n$ (w/o gross error). Note that the ℓ_2 norm error is just on the parameter estimates, and we exclude the error in estimating the outliers e themselves, so that we could compare against the gold-standard GLM regression on the uncorrupted data.

While the M-estimation problem with gross errors in the output space is not convex, it can be seen that the proximal gradient descent (PGD) iterates converge to the true θ^*, corroborating Theorem 3. In the figure, the three rows correspond to the three different GLMs, and the three columns correspond to different outlier scalings, with logarithmic (first column), square-root (second column), and linear (third column) scalings of the number of outliers k as a function of the number of samples n. As the figure shows, the approaches modeling the outliers in the output and parameter spaces perform overwhelmingly better than the baseline ℓ_1 penalized GLM regression estimator, and their error even approaches the estimator that is trained from *uncorrupted data*, even under settings where the number of outliers is a *linear* fraction of the number of samples. The approach modeling outliers in the output space seems preferable in some cases (logistic, exponential), while the approach modeling outliers in the parameter space seems preferable in some cases (Poisson).



(a) australian

(b) german.numer

(c) splice

Figure 2: Comparisons of the empirical prediction errors for different types of outliers on 3 real data examples. Percentage of the used samples in the training dataset: 10% (Left column), 50% (Center) and 100% (Right column)

6.2 Real Data Examples

In this section, we evaluate the performance of our estimators on some real binary classification datasets, obtained from LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/). We focused on the logistic regression case, and compared our two proposed approaches against the standard logistic regression. Note that the datasets we consider have $p < n$ so that we can set $\lambda_{n,\theta} = 0$, thus not adding further sparsity encouraging ℓ_1 regularization to the parameters. We created variants of the datasets by adding varying number of outliers (by randomly flipped the values of the responses y as in the experiments on the simulated datasets). Given each dataset, we split it into three groups; 0.2 of training dataset, 0.4 of validation dataset and 0.4 of test dataset. The validation dataset is used to choose the best performance of $\lambda_{n,e}$ and the constraint constant ρ where we solved the optimization problem under the constraint $\|e\| \leq \rho$.

Figure 2 plots performance comparisons on 3 datasets, one row for each dataset. We varied the fraction of training dataset provided to each algorithm, and columns correspond to these varying fractions: 10% (Left column), 50% (Center) and 100% (Right column). Each graph has a group of four bar-plots corresponding to the four different types of outliers: original dataset without adding artificial outliers (w/o), and where the number of outliers scales as $\log(n)$ (Log), $\sqrt{n}$ (Sqrt) or $0.1(n)$ (Linear), given n training examples. Our proposed robust methods perform as well or better, with particularly strong performance, with more outliers, and/or where less samples are used for the training. We found the latter phenomenon interesting, and worthy of further research: that robustness might help the performance of regression models even in the absence of outliers by preventing overfitting.

7 Conclusion

We have provided a comprehensive analysis of statistical estimation of high dimensional GLMs with grossly corrupted observations. We detail *two distinct approaches* for modeling sparse outlier errors in GLMs: incidentally these are equivalent in the linear case, though distinct for general GLMs. For both approaches, we provide tractable M-estimators, and analyze their consistency by providing ℓ_2 error bounds. The parameter space approach is nominally more intuitive and computationally tractable, but requires stronger conditions for the error bounds to hold (and in turn for ℓ_2 consistency). In contrast, the second output space based approach leads to a nonconvex problem, which makes statistical and computational analyses harder. Nonetheless, we show that this second approach is better than the first on the statistical front, since we obtain better bounds that require weaker conditions to hold, and on the computational front it is comparable, as we show a simple projected gradient descent algorithm converges to one of the global optima up to statistical precision.

Acknowledgments. We acknowledge the support of NSF via IIS-1149803, and DoD via W911NF-12-1-0390.

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *NIPS 23*, pages 37–45, 2010.

A. Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

D. Donoho and M. Elad. Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100:2197–2202, March 2003.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2003.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

Y. Jin and B. Rao. Algorithms for robust linear regression by exploiting the connection to sparse signal recovery. In *ICASSP*, 2010.

K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modeling using the t distribution. *J. Amer. Stat. Assoc.*, 84:881–896, 1989.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *NIPS 24*, pages 2726–2734, 2011.

R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, 2006.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3), 2006.

K. Mitra, A. Veeraraghavan, and R. Chellappa. Robust rvm regression using sparse outlier model. In *IEEE CVPR*, 2010.

A. Y. Ng. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *International Conference on Machine Learning*, 2004.

N. H. Nguyen and T. D. Tran. Robust Lasso with missing and grossly corrupted observations. *IEEE Trans. Info. Theory*, 2011. submitted.

P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, 1987.

P. J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880, 1984.

Y. She and A. B. Owen. Outlier detection using nonconvex penalized regression. *JASA*, 106(494):626–639, 2011.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Info. Theory*, 51(3):1030–1051, March 2006.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Info. Theory*, 55:2183–2202, 2009.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via generalized linear models. In *NIPS 25*, pages 1367–1375, 2012.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7:2541–2567, 2007.