Taming heavy-tailed features by shrinkage

Ziwei Zhu
University of Michigan

Wenjing Zhou
University of Michigan

Abstract

In this work, we focus on a variant of the generalized linear model (GLM) called corrupted GLM (CGLM) with heavy-tailed features and responses. To robustify the statistical inference on this model, we propose to apply ℓ_4-norm shrinkage to the feature vectors in the low-dimensional regime and apply elementwise shrinkage to them in the high-dimensional regime. Under bounded fourth moment assumptions, we show that the maximum likelihood estimator (MLE) based on the shrunk data enjoys nearly the minimax optimal rate with an exponential deviation bound. Our simulations demonstrate that the proposed feature shrinkage significantly enhances the statistical performance in linear regression and logistic regression on heavy-tailed data. Finally, we apply our shrinkage principle to guard against mislabeling and image noise in the human-written digit recognition problem. We add an ℓ_4-norm shrinkage layer to the original neural net and reduce the testing misclassification rate by more than 30% relatively in the presence of mislabeling and image noise.

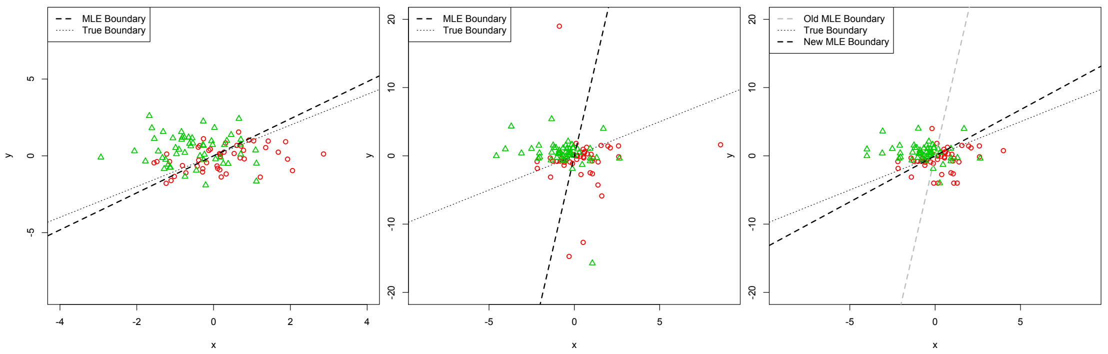
1 Introduction

Heavy-tailed data abound in modern data analytics. For instance, financial log-returns and macroeconomic variables usually exhibit heavy tails (Cont (2001)). In a genomic study, microarray data are always wildly fluctuated (Liu et al. (2003), Purdom et al. (2005)). In deep learning, features learned by deep neural nets are generated via highly nonlinear transformation of the original data and thus have no guarantee of exponential-tailed distribution. These real-world cases contradict the common sub-Gaussian or sub-exponential conditions in the statistics literature. A series of questions thus arise: with heavy-tailed data, can we still achieve good statistical properties of the previous standard estimators or testing statistics? If not, is there a solution to overcome heavy-tailed corruption and achieve equally well statistical performance as with exponential-tailed data?

To answer these questions, perhaps the easiest statistical problem to start with is the mean estimation problem. It turns out surprisingly, as first pointed out by Catoni (2012), that from a high-probability deviation perspective, the empirical mean is far from optimal when data only have finite low-order moments. Catoni (2012) proposed a novel M-estimator for the population mean and revealed its sub-Gaussian behavior around the true mean under merely bounded second moment assumptions. The score function therein is constructed to be logarithmic with respect to the deviation when it is large, thereby being insensitive to outliers and yielding a robust M-estimator. Since then, there has been a surge of interest in light-tailed mean estimators for heavy-tailed data, particularly through the median-of-means approach (Nemirovsky et al. (1982)). A partial list of the related literature includes Bubeck et al. (2013), Minsker (2015), Devroye et al. (2016), Hsu and Sabato (2016), Lugosi and Mendelson (2019e), Lugosi and Mendelson (2019d), Lugosi and Mendelson (2019b), among others. Hsu and Sabato (2016) and Lugosi and Mendelson (2019e) established the sub-Gaussianity of the median-of-means estimator under univariate and multivariate cases respectively. Minsker (2015),Hsu and Sabato (2016) and Lugosi and Mendelson (2019b) constructed and analyzed the median-of-means estimators in general metric spaces.

Beyond the mean estimation problem, robust risk minimization and the median-of-means approach are proved to be successful under a great variety of problem setups with heavy-tailed data, e.g., covariance matrix or general matrix estimation (Minsker (2018); Mendelson and Zhivotovskiy (2020); Fan et al. (2020+)), empirical risk minimization (Brownlees et al. (2015); Hsu and Sabato (2016); Lugosi and Mendelson (2019d,c); Lecué

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).



(a) Gaussian features (b) Student's t_2 features (c) Shrunk Student's t_2 features

Figure 1: Logistic Regression with 10% mislabeled data based on different features

and Lerasle (2020)), low-dimensional regression and high-dimensional sparse linear regression (Loh (2017); Bhatia et al. (2015); Fan et al. (2017); Bhatia et al. (2017); Pan et al. (2019); Sun et al. (2020); Wang et al. (2020+)), low-rank matrix recovery (Fan et al. (2020+)) and so forth. We refer our readers to Lugosi and Mendelson (2019a) for a comprehensive survey on recent advancement in mean estimation and regression under heavy-tailed distributions.

Despite heated research on statistics with heavy-tailed data, few have studied the effect of heavy tails of features or designs in regression. Previous works such as Loh (2017), Bhatia et al. (2015), Fan et al. (2017) and Avella-Medina and Ronchetti (2018) mainly focus on cases where only responses are heavy-tailed or contaminated. It remains unclear whether widely spread features or designs are blessings or curses to statistical efficiency. This motivates us to consider a variant of the generalized linear model (GLM) called corrupted GLM (CGLM) that accommodates both heavy-tailed designs and responses. The CGLM allows extra random corruption on the response of the traditional GLM, thereby enjoying much broader model capacity and embraces a myriad of important real-world problems.

One key message of our paper is that heavy-tailed features can aggravate the corruption on the response and jeopardize standard statistical approaches. To further illustrate this point, Panels (a) and (b) of Figure 1 contrast the performance of the standard MLE on light-tailed features and heavy-tailed features under a logistic regression model. When the data points are widely spread as in Panel (b), the boundary derived from the MLE deviates far from the true boundary. When the data points are Gaussian, however, Panel (a) shows nearly perfect alignment between the MLE boundary and the true boundary. The reason for this difference is that the outliers, especially those mislabeled, have severe influence on the log-likelihood and can undermine the validity of the MLE.

To tame the heavy-tails of the features, we propose to shrink the features before calculating the M-estimator. Given feature vectors $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$, a threshold value τ and a norm $\|\cdot\|$ on the feature space, the shrunk features $\{\widetilde{\mathbf{x}}_i^s\}_{i=1}^n$ are defined as:

$$\widetilde{\mathbf{x}}_i^s = \min(\|\mathbf{x}_i\|, \tau)\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}.$$

In short, we restrict $\|\widetilde{\mathbf{x}}_i^s\|$ below the level τ. In the sequel, we illustrate both theoretically and numerically that the feature shrinkage trades little bias for great variance reduction such that the resulting MLE achieves (nearly) the minimax optimal statistical rate up to logarithmic factors of n and failure probability. Panels (b) and (c) of Figure 1 compare the performance of MLE based on original heavy-tailed features and shrunk features. One can see that after feature shrinkage, the new MLE boundary becomes much more aligned with the true boundary than the original one, because the shrinkage mitigates the perturbation of the outliers on the log-likelihood. Note that similar ideas have been explored to overcome adversarial corruption on features. For example, Chen et al. (2013) used the trimmed inner product to robustify standard high-dimensional regression methods and established strong statistical guarantees while allowing a certain fraction of observations to be arbitrarily corrupted. Feng et al. (2014) proposed to ignore observations with large feature values to prevent adversarial feature corruption in logistic regression and binary classification problems. The major difference between our work and theirs is that our focus is tail behavior, rather than corruption, of features in regression problems. We assume that the features have only few bounded moments, while

Chen et al. (2013) and Feng et al. (2014) assume them to be sub-Gaussian. Our theory does not assume any corruption on the features; all the corruption in this paper is imposed on responses.

The rest of the paper is organized as follows. In Section 2, we elucidate the CGLM and the log-likelihood based on the shrunk data. In Section 3, we introduce specific feature shrinkage methods for different regimes and present our main theoretical results. Under the low-dimensional regime, we prove that the MLE based on ℓ_4-norm shrunk features enjoys the same optimal statistical rate as the standard MLE with sub-Gaussian features up to a $(\log n)^{1/2}$ factor. For high-dimensional models, we show that the ℓ_1-regularized MLE based on elementwise shrunk features achieves (nearly) the minimax optimal rate. One technical contribution worth emphasis is that we provide a rigorous justification of the (restricted) strong convexity of the negative likelihood based on shrunk features. In Section 4, we demonstrate the numerical superiority of our proposed estimators over the standard MLEs under both low-dimensional and high-dimensional regimes. We investigate two important problem setups: linear regression with heavy-tailed noise and binary logistic regression with mislabeled data. Finally, motivated by the shrinakge principle, we add an ℓ_4-norm shrinkage layer to a convolutional neural network to classify human-written digits in the MNIST dataset. We show the significant improvement of the new architecture in the presence of mislabeling and image noise.

2 Problem setup

In this section, we formulate the corrupted GLM as aforementioned. Recall the definition of the standard GLM with the canonical link. Suppose we have n observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, where y_i is the response and $\mathbf{x}_i$ is the feature vector valued in $\mathbb{R}^d$. Under the GLM with the canonical link, the probability density function of the response y_i is defined as

$$
\begin{aligned}
f_n(\mathbf{y};\mathbf{X},\boldsymbol{\beta}^*) &= \prod_{i=1}^n f(y_i;\eta_i^*) \\
&= \prod_{i=1}^n \left\{ c(y_i)\exp\left(\frac{y_i\eta_i^* - b(\eta_i^*)}{\phi}\right)\right\},
\end{aligned} \tag{1}
$$

where $\mathbf{y} = (y_1, \cdots, y_n)^\top$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$, $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is the regression coefficient vector, $\eta_i^* := \mathbf{x}_i^\top\boldsymbol{\beta}^*$, $b(\cdot)$ is a known function that is twice differentiable with a positive second derivative and $\phi > 0$ is the dispersion parameter. The negative log-likelihood corresponding to (1) is given, up to an affine transformation, by

$$
\begin{aligned}
\ell_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n -y_i\mathbf{x}_i^\top\boldsymbol{\beta} + b(\mathbf{x}_i^\top\boldsymbol{\beta}) \\
&= \frac{1}{n}\sum_{i=1}^n -y_i\eta_i + b(\eta_i) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\beta}),
\end{aligned} \tag{2}
$$

and the gradient and Hessian of $\ell_n(\boldsymbol{\beta})$ are respectively

$$
\nabla\ell_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n (y_i - b'(\mathbf{x}_i^\top\boldsymbol{\beta}^*))\mathbf{x}_i \tag{3}
$$

$$
\nabla^2\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n b''(\mathbf{x}_i^\top\boldsymbol{\beta}^*)\mathbf{x}_i\mathbf{x}_i^\top. \tag{4}
$$

Note that $b'(\mathbf{x}_i^\top\boldsymbol{\beta}^*) = \mathbb{E}(y_i|\mathbf{x}_i)$. For ease of notation, we write the empirical hessian $\nabla^2\ell_n(\boldsymbol{\beta})$ as $\mathbf{H}_n(\boldsymbol{\beta})$ and $\mathbb{E}(b''(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i^\top)$ as $\mathbf{H}(\boldsymbol{\beta})$.

Under a CGLM, for the ith observation we can only observe its corrupted response

$$
z_i = y_i + \epsilon_i \tag{5}
$$

rather than the original response y_i, where ϵ_i is random noise. We emphasize that introducing ϵ_i significantly improves the flexibility of the original GLM, such that now the response is not limited within the exponential family. The CGLM embraces many more real-world problems with complex structures, e.g., the linear regression model with heavy-tailed noise, the logistic regression with mislabeled samples and so forth.

To handle the heavy-tailed features and noise on the response, we propose to shrink the data $\{(z_i, \mathbf{x}_i)\}_{i=1}^n$ first and use them to construct the log-likelihood (2). Formally, define

$$
\widetilde{\ell}_n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^n -\widetilde{z}_i\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta} + b(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}). \tag{6}
$$

We denote the hessian matrix of $\widetilde{\ell}_n(\boldsymbol{\beta})$ by $\widetilde{\mathbf{H}}_n(\boldsymbol{\beta})$ and its population version $\mathbb{E}\widetilde{\mathbf{H}}_n(\boldsymbol{\beta})$ by $\widetilde{\mathbf{H}}(\boldsymbol{\beta})$. In the next section, we elucidate the specific shrinkage methods to construct $\widetilde{\mathbf{x}}_i$ and $\widetilde{z}_i$ in both low-dimensional and high-dimensional regimes and explicitly derive the statistical error rates of the MLE based on $\widetilde{\ell}_n(\boldsymbol{\beta})$.

3 Main results

3.1 Notation

Here we collect all the notation that we use in the sequel. We use regular letters for scalars, bold regular letters for vectors and bold capital letters for matrices.

Denote the d-dimensional Euclidean unit sphere by $\mathcal{S}^{d-1}$. Denote the Euclidean and ℓ_1-norm balls with the center $\boldsymbol{\beta}^*$ and radius r by $\mathcal{B}_2(\boldsymbol{\beta}^*, r)$ and $\mathcal{B}_1(\boldsymbol{\beta}^*, r)$ respectively. We write the set $\{1, \cdots, d\}$ as $[d]$. For two scalar sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we say $a_n \asymp b_n$ if there exist two universal constants C_1 and C_2 such that $C_1 b_n \leq a_n \leq C_2 b_n$ for all $n \geq 1$. We use $\|\mathbf{v}\|_2$, $\|\mathbf{v}\|_1$ and $\|\mathbf{v}\|_4$ to denote the Euclidean norm, ℓ_1-norm and ℓ_4-norm of $\mathbf{v}$ respectively. Particularly, recall that $\|\mathbf{x}_i\|_4 := (\sum_{j=1}^d x_{ij}^4)^{1/4}$. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_{\mathrm{op}}$ and $\|\mathbf{A}\|_{\max}$ to denote the operator norm and elementwise max-norm of $\mathbf{A}$ respectively and use $\lambda_{\min}(\mathbf{A})$ to denote the minimum eigenvalue of $\mathbf{A}$. For any $\boldsymbol{\beta}^* \in \mathbb{R}^d$ and any differential map $f : \mathbb{R}^d \to \mathbb{R}$, define the first-order Taylor remainder of $f(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ to be

$$\delta f(\boldsymbol{\beta}; \boldsymbol{\beta}^*) := f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) - \nabla f(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

For a set of random variables $\{X_i\}_{i\in\mathcal{I}}$, we say that they are i.i.d. if they are independent and identically distributed. We refer to some quantities as *constants* if they are independent of the sample size n, the dimension d and the sparsity s of $\boldsymbol{\beta}^*$ in the high-dimensional regime.

3.2 Low-dimensional regime

The standard MLE estimator is defined as $\widehat{\boldsymbol{\beta}} := \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \ell_n(\boldsymbol{\beta})$, where $\ell_n(\cdot)$ is characterized as in (2). It is well established that under a standard GLM with bounded features, $\widehat{\boldsymbol{\beta}}$ enjoys $(d/n)^{1/2}$-consistency to the true parameter $\boldsymbol{\beta}^*$ in terms of the Euclidean norm. However, when the feature vectors have only bounded moments, there is no guarantee of $(d/n)^{1/2}$-consistency any more, let alone further perturbation on the response. To overcome the disruption due to heavy-tailed data, we apply ℓ_4-norm shrinkage to the feature vectors. Construct

$$\widetilde{\mathbf{x}}_i := \frac{\min(\|\mathbf{x}_i\|_4, \tau_1)}{\|\mathbf{x}_i\|_4}\mathbf{x}_i \tag{7}$$

and

$$\widetilde{z}_i := \min(|z_i|, \tau_2) z_i/|z_i|, \tag{8}$$

where τ_1 and τ_2 are predetermined thresholds. Clipping on the response is natural; when $|z_i|$ is abnormally large, clipping reduces its magnitude to prevent corruption by ϵ_i. Here we explain more on why we shrink features in terms of the ℓ_4-norm rather than other norms. The ℓ_4-norm shrinkage has been proven to be successful in low-dimensional covariance estimation in Fan et al. (2020+). Theorem 6 therein shows that when data have only bounded fourth moments, the ℓ_4-norm shrinkage sample covariance enjoys an operator-norm rate of order $O_{\mathbb{P}}\{(d \log d/n)^{1/2}\}$ in estimating the population covariance matrix. Intuitively, shrinking $\|\mathbf{x}_i\|_4$ implies thresholding the second moment of the random matrix $\mathbf{x}_i\mathbf{x}_i^\top$ in (4). This inspires us to apply ℓ_4-norm shrinkage to heavy-tailed features to ensure that the empirical hessian $\widetilde{\mathbf{H}}_n(\boldsymbol{\beta})$ is well concentrated around its population version $\mathbf{H}(\boldsymbol{\beta})$. Unlike the sample covariance matrix, the Hessian matrix varies with respect to $\boldsymbol{\beta}$. Therefore, we need to develop *uniform concentration* bounds to ensure that the Hessian matrix is well-behaved within a neighborhood of $\boldsymbol{\beta}^*$ (Lemmas 1 and 3). This is the main technical challenge that distinguishes our work from Fan et al. (2020+). After data shrinkage and clipping, we minimize the negative log-likelihood based on the new data $\{\widetilde{z}_i, \widetilde{\mathbf{x}}_i\}_{i=1}^n$ to derive the M-estimator, i.e., we choose $\widetilde{\boldsymbol{\beta}} := \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \widetilde{\ell}_n(\boldsymbol{\beta})$ to estimate $\boldsymbol{\beta}^*$, where $\widetilde{\ell}_n(\boldsymbol{\beta})$ is defined as in (6).'

We first establish the uniform strong convexity of $\widetilde{\ell}_n(\boldsymbol{\beta})$ over $\boldsymbol{\beta} \in \mathcal{B}_2(\boldsymbol{\beta}^*, r)$ (up to some small tolerance term) that is crucial to our subsequent statistical analysis.

Lemma 1. *Suppose the following conditions hold: (1) $\forall i \in [n]$, $b''(\mathbf{x}_i^\top\boldsymbol{\beta}^*) \leq M < \infty$, and $\forall\, \omega > 0$, $\exists\, m(\omega) > 0$ such that $b''(\eta) \geq m(\omega) > 0$ for $|\eta| \leq \omega$; (2) $\mathbb{E}\mathbf{x}_i = \mathbf{0}$, $\lambda_{\min}(\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top) \geq \kappa_0 > 0$ and $\mathbb{E}(\mathbf{v}^\top\mathbf{x}_i)^4 \leq R < \infty$ for all $\mathbf{v} \in \mathcal{S}^{d-1}$; (3) $\|\boldsymbol{\beta}^*\|_2 \leq L < \infty$. Choose the shrinkage threshold $\tau_1 \asymp (n/\log n)^{1/4}$. For any $0 < r < 1$ and $t > 0$, when $(d \log n/n)$ is sufficiently small, we have with probability at least $1-2\exp(-t)$ that for all $\boldsymbol{\Delta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\Delta}\|_2 \leq r$,*

$$\delta\widetilde{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \boldsymbol{\beta}^*) \geq \kappa\|\boldsymbol{\Delta}\|_2^2 - Cr^2\left\{\left(\frac{t}{n}\right)^{1/2} + \left(\frac{d}{n}\right)^{1/2}\right\},$$

where κ and C are constants.

Remark 1. *Here we explain the conditions of Lemma 1. Condition (1) assumes that the response from the GLM has bounded variance and is non-degenerate when η is bounded. Note here that we do not assume a uniform lower bound of $b''(\eta)$. $m(\omega)$ is allowed to decay to zero as $\omega \to \infty$. Condition (2) says that the population covariance matrix of the design vector $\mathbf{x}_i$ is positive definite and $\mathbf{x}_i$ has bounded fourth moment. Condition (3) is natural: it holds if we have var$(\mathbf{x}_i^\top\boldsymbol{\beta}^*) < \infty$ and $\lambda_{\min}(\mathbb{E}\,\mathbf{x}_i\mathbf{x}_i^\top) \geq \kappa_0 > 0$. Note that the ordinary least square (OLS) estimator has been shown to enjoy consistency under similar bounded fourth moment conditions (Hsu et al. (2012), Audibert et al. (2011), Oliveira (2016)). Theorem 1 later establishes a similar result for the CGLM.*

Remark 2. *In the proof of Theorem 1, we let the radius of the local neighborhood r here decay to zero so that the tolerance term $r^2\{(t/n)^{1/2} + (d/n)^{1/2}\}$ is negligible.*

We are now in position to present the statistical rate of $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

Theorem 1. *Suppose the conditions of Lemma 1 hold. We further assume that (1) $\mathbb{E}z_i^4 \le M_1 < \infty$; (2) $\|\mathbb{E}(\epsilon_i \mathbf{x}_i)\|_2 \le M_2(d/n)^{1/2}$ for some constant M_2. Choose $\tau_1, \tau_2 \asymp (n/\log n)^{1/4}$. There exists a constant $C > 0$ such that when $(d \log n/n)$ is sufficiently small, for any $\xi > 1$,*

$$\mathbb{P}\biggl\{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \ge C\xi\biggl(\frac{d\log n}{n}\biggr)^{1/2}\biggr\} \le 3n^{1-\xi}.$$

Remark 3. *Condition 1 requires merely bounded fourth moments of the response from CGLM. Condition 2 requires the additional corruption to be nearly uncorrelated with the design, which is satisfied if* $\mathrm{E}(\epsilon_i|\mathbf{x}_i) = 0$.

Remark 4. *Some algebra yields the following equivalent form of the high-probability bound in Theorem 1:*

$$\mathbb{P}\biggl\{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \gtrsim \biggl(1 + \frac{\log(1/\delta)}{\log n}\biggr)\biggl(\frac{d\log n}{n}\biggr)^{1/2}\biggr\} \le \delta, \tag{9}$$

where δ is the failure probability. This suggests that $\widetilde{\boldsymbol{\beta}}$ is sub-exponential around $\boldsymbol{\beta}^*$. Lugosi and Mendelson (2019a), Lugosi and Mendelson (2019c) and Lugosi and Mendelson (2019b) study the optimal confidence band for a given δ. They achieved sub-Gaussian estimators for many mean estimation and regression problems. Compared with their optimal rates, our deviation bound has an extra term of $(\log\delta)^{1/2}$. Nevertheless, their results mainly focus on mean estimation and least squares problems with isotropic features; it remains to be an open problem if one can find sub-Gaussian estimators in a CGLM. Our error bounds in Corollary 1 and Theorem 2 and are both sub-exponential.

Remark 5. *Choosing τ_1, τ_2 to be of order $(n/\log n)^{1/2}$ is to achieve bias-and-variance tradeoff in controlling $\|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_2$, which determines the statistical rate of $\widetilde{\boldsymbol{\beta}}$. We refer interested readers to* (23) *to see how we balance the rates of two variance terms (T_1, T_3) and a bias term (T_2).*

In some cases, the covariance between ϵ_i and $\mathbf{x}_i$ does not vanish as n and d grow. For example, in binary logistic regression with mislabeling, we have that

$$\begin{aligned}\mathbb{P}(\epsilon_i = -1|y_i = 1) = p, \mathbb{P}(\epsilon_i = 0|y_i = 1) = 1 - p,\\ \mathbb{P}(\epsilon_i = 1|y_i = 0) = p, \mathbb{P}(\epsilon_i = 0|y_i = 0) = 1 - p,\end{aligned} \tag{10}$$

where $p < 0.5$. In other words, we flip the genuine label y_i with probability p. Then we have

$$\begin{aligned}\mathbb{E}(\epsilon_i\mathbf{x}_i) &= \mathbb{E}(\epsilon_i\mathbf{x}_i 1_{\{y_i=0\}}) + \mathbb{E}(\epsilon_i\mathbf{x}_i 1_{\{y_i=1\}})\\ &= p\mathbb{E}(\mathbf{x}_i(1_{\{y_i=0\}} - 1_{\{y_i=1\}})) = 2p\mathbb{E}(\mathbf{x}_i 1_{\{y_i=0\}}).\end{aligned}$$

The last equality holds because $\mathbb{E}\mathbf{x}_i = \mathbf{0}$. Therefore, $\mathbb{E}(\epsilon_i\mathbf{x}_i) \propto p$ and if p does not decay, neither does $\mathbb{E}(\epsilon_i\mathbf{x}_i)$. Natarajan et al. (2013) solve this noisy label problem through minimizing weighted negative log-likelihood

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^w &:= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell^w(\mathbf{x}_i, z_i; \boldsymbol{\beta})\\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \frac{(1-p)\ell(\mathbf{x}_i, z_i; \boldsymbol{\beta}) - p\ell(\mathbf{x}_i, 1 - z_i; \boldsymbol{\beta})}{1-2p}.\end{aligned} \tag{11}$$

Lemma 1 therein shows that $\mathbb{E}_{\epsilon_i}\ell^w(\mathbf{x}_i, z_i) = \ell(\mathbf{x}_i, y_i)$. This implies that when the sample size is sufficiently large, minimizing the weighted negative log-likelihood above is similar to minimizing the negative log-likelihood with true labels. In the presence of heavy-tailed features, we propose to replace $\mathbf{x}_i$ with the ℓ_4-norm shrunk feature $\widetilde{\mathbf{x}}_i$, i.e., we use

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}^w &:= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell^w(\widetilde{\mathbf{x}}_i, z_i; \boldsymbol{\beta})\\ &= \frac{1}{n}\sum_{i=1}^n \frac{(1-p)\ell(\widetilde{\mathbf{x}}_i, z_i; \boldsymbol{\beta}) - p\ell(\widetilde{\mathbf{x}}_i, 1 - z_i; \boldsymbol{\beta})}{1-2p}\end{aligned} \tag{12}$$

to estimate the regression vector $\boldsymbol{\beta}^*$. The following corollary establishes the statistical error rate of $\widetilde{\boldsymbol{\beta}}^w$ with an exponential deviation bound.

Corollary 1. *Under the logistic regression with random corruption ϵ_i satisfying* (10)*, choose $\tau_1 \asymp (n/\log n)^{1/4}$. Under the conditions of Lemma 1, it holds for some constant C and any $\xi > 1$ such that when $(d\log d/n)^{1/2}$ is sufficiently small,*

$$\mathbb{P}\biggl\{\|\widetilde{\boldsymbol{\beta}}^w - \boldsymbol{\beta}^*\|_2 \ge C\xi\biggl(\frac{d\log n}{n}\biggr)^{1/2}\biggr\} \le 2n^{1-\xi}.$$

Remark 6. *Here we do not need to truncate the response by τ_2 because in logistic regression the response is always bounded.*

3.3 High-dimensional regime

In this section, we consider the regime where the dimension d grows much faster than the sample size n. Recall that the standard ℓ_1-regularized MLE of the regression vector $\boldsymbol{\beta}^*$ under the GLM is

$$\widehat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n\bigl(-y_i\mathbf{x}_i^\top\boldsymbol{\beta} + b(\mathbf{x}_i^\top\boldsymbol{\beta})\bigr) + \lambda\|\boldsymbol{\beta}\|_1, \tag{13}$$

where $(y_i, \mathbf{x}_i)$ comes from the GLM (1) and $\lambda > 0$ is a tuning parameter. Negahban et al. (2012) show that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\{(s\log d/n)^{1/2}\}$ under the GLM when $\{\mathbf{x}_i\}_{i=1}^n$ are sub-Gaussian. However, in the presence of heavy-tailed features $\mathbf{x}_i$ and corruption ϵ_i, the statistical accuracy of $\widehat{\boldsymbol{\beta}}$ might deteriorate if we directly evaluate the log-likelihood (13) on $\{(z_i, \mathbf{x}_i)\}_{i=1}^n$. Our goal is to develop a robust ℓ_1-regularized MLE for $\boldsymbol{\beta}^*$. Let $\widetilde{\mathbf{x}}_i$ be the elementwise shrunk version of $\mathbf{x}_i$ such that for any $j \in [d]$,

$$\widetilde{x}_{ij} := \min(|x_{ij}|, \tau_1) x_{ij}/|x_{ij}|.$$

Construct $\widetilde{z}_i$ as in (8). We propose the following $\widetilde{\boldsymbol{\beta}}$ that minimizes the negative log-likelihood on the shrunk data with ℓ_1-norm regularization:

$$\widetilde{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} \widetilde{\ell}_n(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1,$$

where $\widetilde{\ell}_n(\boldsymbol{\beta})$ is defined as in (6), and where λ is a tuning parameter. For $\mathcal{S} \subset [d]$ and $|\mathcal{S}| = s$, define the restricted cone $\mathcal{C}(\mathcal{S}) := \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}_{\mathcal{S}^c}\|_1 \leq 3\|\mathbf{v}_{\mathcal{S}}\|_1\}$. By Lemma 1 in Negahban et al. (2012), when $\lambda > 2\|\nabla \widetilde{\ell}_n(\boldsymbol{\beta})\|_{\max}$, $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathcal{C}(\mathcal{S})$, which is a crucial property that gives rise to statistical consistency of $\widetilde{\boldsymbol{\beta}}$ under high-dimensional regimes. Therefore, in the following we first present a lemma that characterizes the order of $\|\nabla_{\boldsymbol{\beta}} \widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max}$.

Lemma 2. *Under the following conditions: (1) $\forall i \in [n]$, $b''(\mathbf{x}_i^\top \boldsymbol{\beta}^*) \leq M < \infty$ and $\forall\ \omega > 0$, $\exists\ m(\omega) > 0$ such that $b''(\eta) \geq m(\omega) > 0$ for $|\eta| \leq \omega$; (2) $\mathbb{E}x_{ij} = 0$, $\mathbb{E}x_{ij}^2 x_{ik}^2 \leq R < \infty$ for all $1 \leq j, k \leq d$; (3) $\mathbb{E}z_i^4 \leq M_1$ and $\mathbb{E}\epsilon_i^4 \leq M_1$; (4) $\|\boldsymbol{\beta}^*\|_1 \leq L < \infty$; (5) $|\mathbb{E}\epsilon_i x_{ij}| \leq M_2/n^{1/2}$ for some universal constant $M_2 < \infty$ and all $1 \leq j \leq d$. With $\tau_1, \tau_2 \asymp (n/\log d)^{1/4}$, for any $\xi > 1$ we have that*

$$\mathbb{P}\bigg\{\|\nabla \widetilde{\ell}(\boldsymbol{\beta}^*)\|_{\max} \geq C\xi \Big(\frac{\log d}{n}\Big)^{1/2}\bigg\} \leq 2d^{1-\xi}.$$

Remark 7. *In this lemma we show that $\|\nabla \widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max} = O_{\mathbb{P}}(\sqrt{\log d/n})$. In the sequel we will choose $\lambda \asymp \sqrt{\log d/n}$ to achieve the minimax optimal rate for $\widetilde{\boldsymbol{\beta}}$.*

Another requirement for the statistical guarantee of $\widetilde{\boldsymbol{\beta}}$ is the restricted strong convexity (RSC) of $\widetilde{\ell}_n$, which is first formulated in Negahban et al. (2012). RSC ensures that $\widetilde{\ell}_n(\boldsymbol{\beta})$ is “not too flat”, so that if $|\widetilde{\ell}_n(\widetilde{\boldsymbol{\beta}}) - \widetilde{\ell}_n(\boldsymbol{\beta}^*)|$ is small, then $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are close. In high-dimensional sparse linear regression, RSC is implied by the restricted eigenvalue (RE) condition (Bickel et al. (2009), van de Geer (2007), etc.), a widely studied and acknowledged condition for statistical error analysis of the Lasso estimator. Unlike the quadratic loss in linear regression, the negative log-likelihood $\widetilde{\ell}_n(\boldsymbol{\beta})$ has its hessian matrix $\widetilde{\mathbf{H}}_n(\boldsymbol{\beta})$ depend on $\boldsymbol{\beta}$, which creates technical difficulty of verifying its RSC. Here we establish localized RSC (LRSC) of $\widetilde{\ell}(\boldsymbol{\beta})$, i.e., RSC with $\boldsymbol{\beta}$ constrained within a small neighborhood of $\boldsymbol{\beta}^*$, which has been shown to suffice for statistical analysis of regularized M-estimators in the high-dimensional regime (Fan et al. (2018), Sun et al. (2020)). Formally, we say a loss function $\mathcal{L}(\boldsymbol{\beta})$ satisfies $\mathrm{LRSC}(\boldsymbol{\beta}^*, r, \mathcal{S}, \kappa, \tau_{\mathcal{L}})$ if for any $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{S}) \cap \mathcal{B}_2(\mathbf{0}, r)$,

$$\delta\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \boldsymbol{\beta}^*) \geq \kappa \|\boldsymbol{\Delta}\|_2^2 - \tau_{\mathcal{L}},$$

where $\tau_{\mathcal{L}}$ is a small tolerance term. The following lemma establishes the LRSC of $\widetilde{\ell}_n(\boldsymbol{\beta})$.

Lemma 3. *Suppose the conditions of Lemma 2 hold. Let $\mathcal{S}$ be the true support of $\boldsymbol{\beta}^*$ with $|\mathcal{S}| = s$. Assume that for any $\mathbf{v} \in \mathbb{R}^d$ such that $\mathbf{v} \in \mathcal{C}(\mathcal{S})$ and $\|\mathbf{v}\|_2 = 1$, $0 < \kappa_0 \leq \mathbf{v}^\top \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top)\mathbf{v} \leq \kappa_1 < \infty$. Set $\tau_1 \asymp (n/\log d)^{1/4}$. For any $0 < r < 1$ and $t > 0$, as long as $s^2 \log d/n$ is sufficiently small, we have with probability at least $1 - 2\exp(-t)$ that for any $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{S}) \cap \mathcal{B}_2(\mathbf{0}, r)$,*

$$\delta\widetilde{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \boldsymbol{\beta}^*) \geq \kappa \|\boldsymbol{\Delta}\|_2^2 - C_0 r^2 \bigg\{\Big(\frac{t}{n}\Big)^{1/2} + \Big(\frac{s \log d}{n}\Big)^{1/2}\bigg\},$$

where κ and C_0 are constants.

Remark 8. *This lemma is a high-dimensional analogue of Lemma 1. Similarly, we let r converge to zero when analyzing the statistical rate of $\widetilde{\boldsymbol{\beta}}$, so that the tolerance term $r^2\{(t/n)^{1/2} + (s\log d/n)^{1/2}\}$ becomes negligible.*

Combining Lemmas 2 and 3 yields the statistical guarantee of $\widetilde{\boldsymbol{\beta}}$ as follows.

Theorem 2. *Under the assumptions of Lemma 2 and 3, choose $\lambda = 2C\xi(\log d/n)^{1/2}$ and $\tau_1, \tau_2 \asymp (n/\log d)^{1/4}$, where ξ and C are the same constants as in Lemma 2. Then there exists a constant $C_1 > 0$ such that*

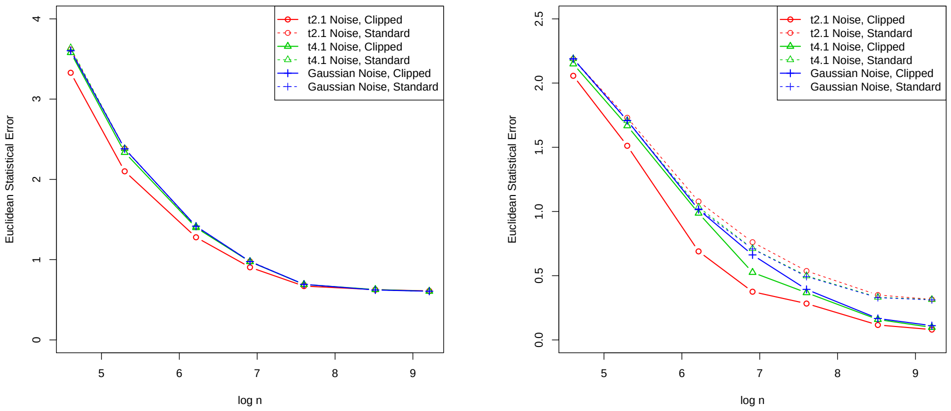
$$\mathbb{P}\bigg\{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \geq C_1 \xi \Big(\frac{s \log d}{n}\Big)^{1/2}\bigg\} \leq 4d^{1-\xi}.$$

Remark 9. *Similarly to Theorem 1, our choice of $\tau_1, \tau_2 \asymp (n/\log d)^{1/4}$ is to achieve bias-and-variance tradeoff in bounding $\|\nabla \ell_n(\boldsymbol{\beta}^*)\|_{\max}$. We refer interested readers to* (28) *for the technical details.*

4 Numerical study

4.1 High-dimensional sparse linear regression

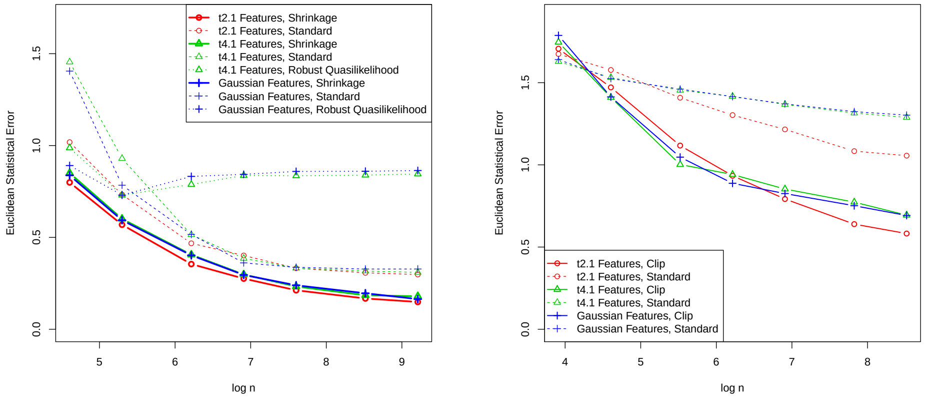
We first consider the high-dimensional sparse linear model $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i$. We set $d = 1000$, $n = 100, 200, 500, 1000, 2000, 5000, 10000$ and $\boldsymbol{\beta}^* = (1, 1, 1, 1, 1, 0, \ldots, 0)^\top$. Recall that in the high-dimensional regime, we propose elementwise shrinkage on the heavy-tailed features and clip the responses. In Figure 2, we compare estimation error of the ℓ_1-regularized least squares estimators based on the shrunk



Standard Gaussian features

$t_{4.1}$ features

Figure 2: High dimensional sparse linear regression with light-tailed features (left) and heavy-tailed features (right)



Low dimension

High dimension

Figure 3: Statistical error of the MLEs based on minimizing $\widetilde{\ell}_n^w(\boldsymbol{\beta})$ with 10% mislabeled data

data and original data under standard Gaussian features and $t_{4.1}$ features respectively. All feature vectors $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d., and within each $\mathbf{x}_i$, $\{x_{ij}\}_{j=1}^d$ are i.i.d. $\{\epsilon_i\}_{i=1}^n$ are i.i.d. noises that are independent of the features and we adjust the magnitude of the noise such that $\text{SD}(\epsilon_i) = 5$ regardless of its distribution. τ_1, τ_2 and λ are selected by cross-validation. The plot is based on $1,000$ independent Monte Carlo simulations. From Figure 2, we first observe that under both light-tailed and heavy-tailed features, the heavier tail ϵ_i has, the more the data shrinkage approach improves the statistical accuracy. More importantly, the benefit from data shrinkage is much more significant in the presence of heavy-tailed features, which justifies our theory. Besides, Table 1 compares the average false discovery proportion (FDP) and true positive rate (TPR) of the shrinakge and standard approaches with $t_{4.1}$ features and $t_{2.1}$ noise. While both methods select much denser models than the true one (because of the ℓ_1-norm penalty), the shrinkage method exhibits higher TPR than the standard method, especially when n is small.

Table 1: Average FDP and TPR of the standard and shrinkage methods with $t_{4.1}$ features and $t_{2.1}$ noise

n	100	200	500	1000
FDP (Shrinkage)	81.3%	82%	80.9%	78.6%
FDP (Standard)	88.6%	82.8%	81.4%	77.6%
TPR (Shrinkage)	60.9%	92.7%	100%	100%
TPR (Standard)	35.1%	82.2%	99.6%	100%

Following one referee's suggestion, we investigated a more high-dimensional (relative to n) setup where

$d = 1000$, $n = 125, 200, 300, 500, 1125$, and where $\boldsymbol{\beta} = (1, \dots, 1, 0, \dots, 0)^\top$ has 15 ones. We set a higher signal-to-noise ratio here to offset the smaller n and achieve reasonable statistical accuracy. We choose $t_{2.1}$ noise and keep the other configurations the same as that in the right panel of Figure 2. In the table below, we compare the average ℓ_2 error of estimating $\boldsymbol{\beta}^*$ by the shrinkage and standard methods with 100 independent Monte Carlo experiments. We can see that the shrinkage approach still outperforms the standard one.

n	125	200	300	500	1125
Shrinkage	3.37	2.56	1.89	1.32	0.80
Standard	3.60	2.97	2.29	1.65	1.01

We also ran Robust Lasso proposed by Chen et al. (2013) for comparisons, where there are two tuning parameters: R, an upper bound of $\|\boldsymbol{\beta}^*\|_1$, and n_1, an upper bound of the number of outliers. We set $R = \|\boldsymbol{\beta}^*\|_1 = \sqrt{5}$ and set n_1 such that the resulting statistical error is minimized, which leads to the oracle performance of the method. When $n = 500$, the mean l_2 estimation errors of Robust Lasso are 1.05 and 1.32 under $t_{4.1}$ and standard Gaussian features respectively, significantly higher than those of our shrinkage approach (0.68, 1.23). Since the performance curves of Robust Lasso are quite close to those of the standard methods, we do not present Robust Lasso in Figure 2 for clarity. From this comparative study, one can see that our shrinkage approach is better than a typical adversarial learning approach in terms of guarding against heavy tails.

Finally, we assessed the sensitivity of our shrinkage methods with respect to the thresholds τ_1 and τ_2 with $t_{4.1}$ features, $t_{2.1}$ noise and $n = 200$. Let $\mathcal{Q}_x$ be the set of the upper $2\%, 1\%, .5\%, .2\%, .1\%$ quantiles of the feature values, and let $\mathcal{Q}_y$ be the set of the upper $2\%, 1\%, .5\%, .2\%, .1\%$ quantiles of the responses. Based on 100 independent Monte Carlo experiments, we compute the average ℓ_2 error $e(\tau_1, \tau_2)$ of estimating $\boldsymbol{\beta}^*$ for all $(\tau_1, \tau_2) \in \mathcal{Q}_x \times \mathcal{Q}_y$ respectively. We found that $\max_{(\tau_1,\tau_2)\in\mathcal{Q}_x\times\mathcal{Q}_y} e(\tau_1, \tau_2) = 1.67$, which is still less than the error of the standard method: 1.79. The error given by CV is 1.41. Therefore, our shrinkage method is not very sensitive to τ_1 or τ_2.

4.2 Logistic regression with mislabeled data

In this subsection, we consider the logistic regression with mislabeled data as characterized by (10). We minimize the weighted negative log-likelihood to derive $\widehat{\boldsymbol{\beta}}^w$ and $\widetilde{\boldsymbol{\beta}}^w$ as described in (11) and (12) to estimate the regression vector $\boldsymbol{\beta}^*$ and compare their performance. The tuning parameters λ and τ_1 are chosen based on cross-validation. We investigate both the low-dimensional and high-dimensional regimes and three distributions of features: $t_{2.1}$, $t_{4.1}$ and Gaussian features. We scale the features so that the marginal variance of each dimension is always 21 regardless of its distribution.

In the low-dimensional regime, let $d = 10$, n range from 10^2 to 10^4, $\boldsymbol{\beta}^* = (0.5\mathbf{1}_5^\top, -0.5\mathbf{1}_5^\top)^\top$ and $p = 0.1$. The left panel of Figure 3 compares $\|\widehat{\boldsymbol{\beta}}^w - \boldsymbol{\beta}^*\|_2$ and $\|\widetilde{\boldsymbol{\beta}}^w - \boldsymbol{\beta}^*\|_2$ under $t_{2.1}, t_{4.1}$ and Gaussian features. We can observe that $\widetilde{\boldsymbol{\beta}}^w$ significantly outperforms $\widehat{\boldsymbol{\beta}}^w$ under $t_{2.1}$ and $t_{4.1}$ features, and they perform equally well when features are Gaussian. This perfectly validates our theory. We also implemented a robust quasi-likelihood approach with ℓ_1-norm regularization (Avella-Medina and Ronchetti, 2018), whose performance is presented in the left panel of Figure 3. We can see that the corresponding error does not go down as n increases; the reason is that the quasi-likelihood method does not take into account the random flippling of the labels and is thus biased in terms of estimating $\boldsymbol{\beta}^*$.

In the high-dimensional regime, we apply elementwise shrinakge to $\mathbf{x}_i$ to derive $\widetilde{\boldsymbol{\beta}}^w$. Let $d = 100$, n range from 50 to 5,000, $\boldsymbol{\beta}^* = (1, 1, -1, 0, \dots, 0)$ and $p = 0.1$. As shown in the right panel of Figure 3, $\widetilde{\boldsymbol{\beta}}^w$ enjoys sharper statistical accuracy than $\widehat{\boldsymbol{\beta}}^w$ under all the three types of features. The outstanding performance of $\widetilde{\boldsymbol{\beta}}^w$ under the Gaussian feature scenario is particularly surprising. We conjecture that feature shrinkage here downsizes $\|\nabla\widetilde{\ell}_n^w(\boldsymbol{\beta}^*)\|_{\max}$ and thus leads to more effective regularization. We did not manage to report the performance of the robust quasi-likelihood method since the provided code did not run through.

4.3 Experiments on the MNIST dataset

Motivated by the effectiveness of feature shrinkage, we incorporate a shrinkage layer to a convolutional neural network (CNN) to robustify its classification performance on corrupted images. Figure 4 illustrates this new architecture, which we call a shrinkage CNN. The new shrinkage layer applies the ℓ_4-norm shrinkage as in (7) to the feature vector $\mathbf{x}$ learned by the original CNN to guard against its heavy tail if any. Then the shrunk features are used to derive the posterior probability of each class.

We classify the digits 4's and 9's in the MNIST (LeCun (1998)) dataset when the images are randomly mislabeled with probability 0.4 and corrupted by "salt" noise. We train both the original CNN and its shrinkage variant by minimizing the weighted negative log-likelihood $\widetilde{\ell}_n^w(\boldsymbol{\beta})$ in (12). We choose $\tau_1 = 2$ in (7) in the ℓ_4-norm shrinakge layer. We repeat flipping labels, adding noise

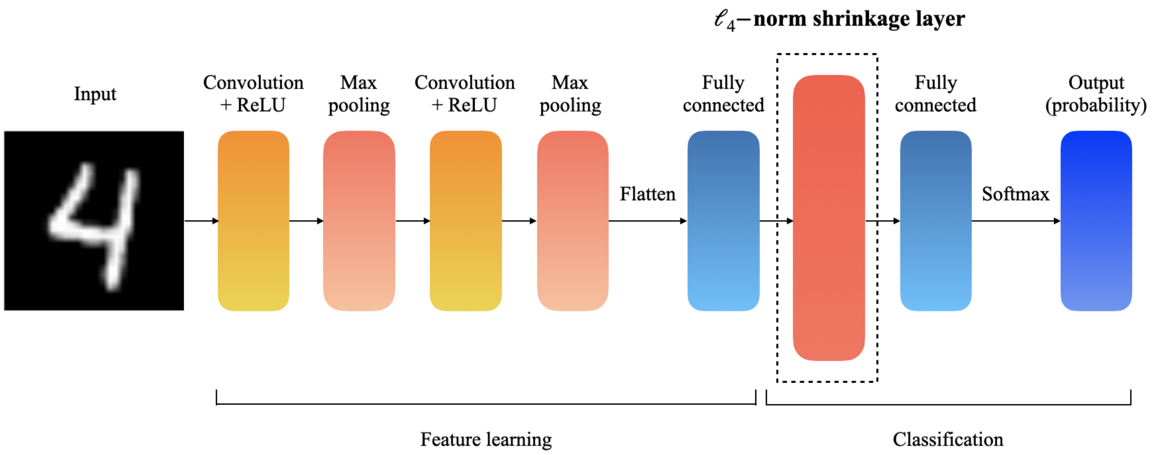


Figure 4: Architecture of the shrinkage CNN

and training for 100 times independently to evaluate the average misclassification rate. The result is presented in Table 2. We can see that the feature shrinkage layer reduces the testing misclassification rate by more than 30% relatively in the presence of noisy pixels.

Table 2: Average testing misclassification rate (with standard error in the parentheses) on noisy MNIST images under mislabeling probability 40%

Noisy Pixel Ratio	Original CNN	Shrinkage CNN
0	$3.64\%_{(0.20\%)}$	$2.93\%_{(0.09\%)}$
0.1	$6.88\%_{(0.22\%)}$	$4.18\%_{(0.17\%)}$
0.2	$6.90\%_{(0.21\%)}$	$4.37\%_{(0.16\%)}$
0.4	$10.69\%_{(0.29\%)}$	$6.65\%_{(0.24\%)}$
0.6	$18.82\%_{(0.88\%)}$	$12.80\%_{(0.65\%)}$

5 Discussion

This paper proposes and studies several shrinkage principles for CGLMs under both low-dimensional and high-dimensional regimes. Assessing the tail behavior of features and shrinking the features appropriately is crucial to achieve reliable statistical inference. There are two future research directions to pursue: (1) designing computationally efficient algorithms or guidlines to find the appropriate shrinkage thresholds; (2) handling adversarial corruption on features.

6 Acknowledgement

The authors thank the four anonymous referees for their insightful comments and suggestions that substantially improve the quality of the paper. Ziwei Zhu gratefully acknowledges funding from the U.S. National Science Foundation (NSF) grant DMS-2015366.

References

Audibert, J.-Y., Catoni, O. et al. (2011). Robust linear least squares regression. *The Annals of Statistics* **39** 2766–2794.

Avella-Medina, M. and Ronchetti, E. (2018). Robust and consistent variable seletion in high-dimensional generalized linear models. *Biometrika* **105** 31–44.

Bhatia, K., Jain, P., Kamalaruban, P. and Kar, P. (2017). Consistent robust regression. In *Advances in Neural Information Processing Systems*.

Bhatia, K., Jain, P. and Kar, P. (2015). Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Brownlees, C., Joly, E. and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics* **43** 2507–2536.

Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory* **59** 7711–7717.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré* **48** 1148–1185.

Chen, Y., Caramanis, C. and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* **1**.

Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics* **44** 2695–2725.

Fan, J., Li, Q. and Wang, Y. (2017). Robust estimation of high-dimensional mean regression. *Journal of Royal Statistical Society, Series B* **79** 247–265.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics* **46** 814–841.

Fan, J., Wang, W. and Zhu, Z. (2020+). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of Statistics* To appear.

Feng, J., Xu, H., Mannor, S. and Yan, S. (2014). Robust logistic regression and classification. In *Advances in Neural Information Processing Systems*.

Hsu, D., Kakade, S. M. and Zhang, T. (2012). Random design analysis of ridge regression. In *Conference on Learning Theory*.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research* **17** 1–40.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics* **48** 906–931.

LeCun, Y. (1998). The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* .

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: Isoperimetry and Processesrocesses*. Springer Science & Business Media.

Liu, L., Hawkins, D. M., Ghosh, S. and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences* **100** 13167–13172.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics* **45** 886–896.

Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* **19** 1145–1190.

Lugosi, G. and Mendelson, S. (2019b). Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields* **175** 957–973.

Lugosi, G. and Mendelson, S. (2019c). Risk minimization by median-of-means tournaments. *Journal of European Mathematical Society* **22** 925–965.

Lugosi, G. and Mendelson, S. (2019d). Robust multivariate mean estimation: the optimality of trimmed mean. *arXiv preprint arXiv:1907.11391* .

Lugosi, G. and Mendelson, S. (2019e). Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics* **47** 783–794.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability* **28** 863–884.

Mendelson, S. and Zhivotovskiy, N. (2020). Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics* **48** 1648–1664.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335.

Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics* **46** 2871–2903.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K. and Tewari, A. (2013). Learning with noisy labels. In *Advances in Neural Information Processing Systems*.

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **24** 538–577.

Nemirovsky, A.-S., Yudin, D.-B. and Dawson, E.-R. (1982). Problem complexity and method efficiency in optimization. *SIAM Review* **27** 264–265.

Oliveira, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields* **166** 1175–1194.

Pan, X., Sun, Q. and Zhou, W.-X. (2019). Nonconvex regularized robust regression with oracle properties in polynomial time. *arXiv preprint arXiv:1907.04027* .

Purdom, E., Holmes, S. P. et al. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology* **4** 1070.

Sun, Q., Zhou, W. and Fan, J. (2020). Adaptive Huber regression: Optimality and phase transition. *Journal of the American Statistical Association* **115** 254–265.

van de Geer, S. (2007). The deterministic LASSO. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Wang, L., Zheng, C., Zhou, W. and Zhou, W.-X. (2020+). A new principle for tuning-free Huber regression. *Statistica Sinica* .