Corruption-Tolerant Algorithms for Generalized Linear Models

Bhaskar Mukhoty[1*], **Debojyoti Dey**[2], **Purushottam Kar**[2]

[1]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2]Indian Institute of Technology Kanpur, Uttar Pradesh, India
bhaskar.mukhoty@mbzuai.ac.ae, {debojyot,purushot}@cse.iitk.ac.in

Abstract

This paper presents SVAM (Sequential Variance-Altered MLE), a unified framework for learning generalized linear models under adversarial label corruption in training data. SVAM extends to tasks such as least squares regression, logistic regression, and gamma regression, whereas many existing works on learning with label corruptions focus only on least squares regression. SVAM is based on a novel variance reduction technique that may be of independent interest and works by iteratively solving weighted MLEs over variance-altered versions of the GLM objective. SVAM offers provable model recovery guarantees superior to the state-of-the-art for robust regression even when a constant fraction of training labels are adversarially corrupted. SVAM also empirically outperforms several existing problem-specific techniques for robust regression and classification. Code for SVAM is available at https://github.com/purushottamkar/svam/

Introduction

Generalized linear models (GLMs) (Nelder and Wedderburn 1972) are effective models for a variety of discrete and continuous label spaces, allowing the prediction of binary or count-valued labels (logistic, Poisson regression) as well as real-valued labels (gamma, least-squares regression). Inference in a GLM involves two steps: given a feature vector $\mathbf{x} \in \mathbb{R}^d$ and model parameters $\mathbf{w}^*$, a *canonical parameter* is generated as $\theta := \langle \mathbf{w}^*, \mathbf{x} \rangle$ then the label y is sampled from the exponential family distribution

$$\mathbb{P}\left[y \mid \theta\right] = \exp(y \cdot \theta - \psi(\theta) - h(y)),$$

where the function $h(\cdot)$ is specific to the GLM and $\psi(\cdot)$ is a normalization term, also known as log partition function. It is common to use a *non-canonical link* such as $\theta := \exp(\langle \mathbf{w}^*, \mathbf{x} \rangle)$ for gamma distribution. GLMs also admit vector valued label $\mathbf{y} \in \mathbb{R}^n$ by substituting the scalar product by inner product $\langle \mathbf{y}, \boldsymbol{\eta} \rangle$ where $\boldsymbol{\eta} := \mathbf{X}\mathbf{w}^*$ is the canonical parameter and $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the covariate matrix.

Problem Description: Given data $\{(\mathbf{x}^i, y_i)\}_{i=1}^n$ generated using a known GLM but unknown model parameters $\mathbf{w}^*$, statistically efficient techniques exist to recover a consistent estimate of the model $\mathbf{w}^*$ (McCullagh and Nelder 1989). However, these techniques break down if several observed labels y_i are corrupted, not just by random statistical noise but by adversarially generated structured noise. Suppose $k < n$ labels are corrupted i.e. for some k data points $i_1, \ldots, i_k$, the actual label $y_{i_j}, j = 1, \ldots, k$ generated by the GLM are replaced by the adversary with corrupted ones say $\tilde{y}_{i_j}$. Can we still recover $\mathbf{w}^*$? Note that the learning algorithm is unaware of the points that are corrupted.

Breakdown Point: The largest fraction $\alpha = k/n$ of corruptions that a learning algorithm can tolerate while still offering an estimate of $\mathbf{w}^*$ with bounded error is known as its breakdown point. This paper proposes the SVAM algorithm that can tolerate $k = \Omega\left(n\right)$ corruptions i.e. $\alpha = \Omega\left(1\right)$.

Adversary Models: Contamination of the training labels $y_1, \ldots, y_n$ by an adversary can misguide the learning algorithm into selecting model parameters of the adversary's choice. An adversary has to choose (1) which labels $i_1, \ldots, i_k$ to corrupt and (2) what corrupted labels $\tilde{y}_{i_1}, \ldots, \tilde{y}_{i_k}$ to put there. Adversary models emerge based on what information the adversary can consult while making these choices. The *oblivious* adversary must make both these choices with no access to the original data $\{(\mathbf{x}^i, y_i)\}_{i=1}^n$ or true model $\mathbf{w}^*$ and thus, can only corrupt a random/fixed subset of k labels by sampling $\tilde{y}_{i_j}$ from some predetermined noise distribution. This is also known as the *Huber* noise model. On the other hand, a *fully adaptive* adversary has full access to the original data and true model while making both choices. Finally, the *partially adaptive* adversary must choose the corruption locations without knowledge of original data or true model but has full access to these while deciding the corrupted labels. See Appendix B for details.

Contributions: This paper describes the SVAM (Sequential Variance-Altered MLE) framework that offers:
1. robust estimation with a breakdown point $\alpha = \Omega\left(1\right)$ against partially and fully adaptive adversaries for robust least-squares regression and mean estimation and $\alpha = \Omega\left(1/\sqrt{d}\right)$ for robust gamma regression. Prior works do not offer any breakdown point for gamma regression.
2. exact recovery of the true model $\mathbf{w}^*$ against a fully-adaptive adversary for the case of least squares regression,
3. the use of variance reduction technique (see §) in robust learning, which is novel to the best of our knowledge,
4. extensive empirical evaluation demonstrating that despite

[*] Work done while the author was a student at IIT Kanpur.

being a generic framework, SVAM is competitive to or outperforms algorithms specifically designed to solve problems such as least-squares and logistic regression.

Related Works

In the interest of space, we review aspects of literature most related to SVAM and refer to others (Diakonikolas et al. 2019a; Mukhoty et al. 2019) for a detailed review.

Robust GLM learning has been studied in a variety of settings. (Cantoni and Ronchetti 2001) considered an oblivious adversary (Huber's noise model) but offered a breakdown point of $\alpha = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ i.e. tolerate $k \leq \mathcal{O}\left(\sqrt{n}\right)$ corruptions. (Yang, Tewari, and Ravikumar 2013) solve robust GLM estimation by solving M-estimation problems. However, they require the magnitude of the corruptions to be upper-bounded by some constant i.e. $|y_i - \tilde{y}_i| \leq \mathcal{O}\left(1\right)$ and offer a breakdown point of $\alpha = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Moreover, their approach solves L_1-regularized problems using projected gradient descent that offers slow convergence. In contrast, SVAM offers a linear rate of convergence, offers a breakdown point of $\alpha = \Omega\left(1\right)$ i.e. tolerate $k = \Omega\left(n\right)$ corruptions and can tolerate corruptions with unbounded magnitude introduced by a partially or fully adaptive adversary.

Specific GLMs such as robust regression have received focused attention. Here the model is $\mathbf{y} = X\mathbf{w}^* + \mathbf{b}$ where $X \in \mathbb{R}^{n \times d}$ is the feature matrix and $\mathbf{b}$ is k-sparse corruption vector denoting the adversarial corruptions. A variant of this, studies a *hybrid* noise model that replaces the zero entries of $\mathbf{b}$ with Gaussian noise $\mathcal{N}(0, \sigma^2)$. (Nguyen and Tran 2013; Wright and Ma 2010) solve an L_1 minimization problem which is slow in practice. (see §). (Bhatia, Jain, and Kar 2015) use hard thresholding techniques to estimate the subset of uncorrupted points while (Mukhoty et al. 2019) modify the IRLS algorithm to do so. However, (Bhatia, Jain, and Kar 2015; Mukhoty et al. 2019) are unable to offer consistent model estimates in the hybrid noise model even if the corruption rate $\alpha = k/n \to 0$ which is surprising since $\alpha \to 0$ implies vanishing corruption. In contrast, SVAM offers consistent model recovery in the hybrid noise model against a fully adaptive adversary when $\alpha \to 0$. (Suggala et al. 2019) also offer consistent recovery with breakdown points $\alpha > 0.5$ but assume an oblivious adversary.

Robust classification with $y_i \in \{-1, +1\}$ has been explored using robust surrogate loss functions (Natarajan et al. 2013) and ranking (Feng et al. 2014; Northcutt, Wu, and Chuang 2017) techniques. These works do not offer breakdown points but offer empirical comparisons.

Robust mean estimation entails recovering an estimate $\hat{\boldsymbol{\mu}} \in \mathbb{R}^d$ of the mean $\boldsymbol{\mu}^*$ of a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}^*, \Sigma)$ given n samples of which an α fraction are corrupted (Lai, Rao, and Vempala 2016). Estimation error is known to be lower bounded $\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\right\|_2 \geq \Omega\left(\alpha\sqrt{\log\frac{1}{\alpha}}\right)$ for this application even if $n \to \infty$ (Diakonikolas and Kane 2019). (Diakonikolas et al. 2019a) use convex programming techniques and offer $\mathcal{O}\left(\alpha \log^{\frac{3}{2}} \frac{1}{\alpha}\right)$ error given $n \geq \tilde{\Omega}\left(\frac{d^2}{\alpha^2}\right)$ samples and a poly $\left(n, d, \frac{1}{\alpha}\right)$ runtime. (Cheng, Diakonikolas, and Ge 2019) improve the running time to $\frac{\widetilde{\mathcal{O}}(nd)}{\text{poly}(\alpha)}$. The recent work of (Dalalyan and Minasyan 2022) uses an IRLS-style approach that internally relies on expensive SDP-calls but offers high breakdown points. SVAM uses $n = \mathcal{O}\left(\log^2 \frac{1}{\alpha}\right)$ samples and offers a recovery error of $\mathcal{O}\left(\mathit{trace}(\Sigma)(\log\frac{1}{\alpha})^{-1/2}\right)$. This is comparable to existing works if $\mathit{trace}(\Sigma) = \mathcal{O}\left(1\right)$. Moreover, SVAM is much faster and simpler to implement in practice.

Meta algorithms such as robust gradient techniques, median-of-means (Lecué and Lerasle 2020), tilted ERM (Li et al. 2021) and maximum correntropy criterion (Feng et al. 2015) have been studied. SEVER (Diakonikolas et al. 2019b) uses gradient covariance matrix to filter out the outliers along its largest eigenspace while RGD (Prasad et al. 2018) uses robust gradient estimates to perform robust first-order optimization directly. While convenient to execute, they may require larger training sets, e.g., SEVER requires $n > d^5$ samples for robust least-squares regression whereas SVAM requires $n > \Omega\left(d\log(d)\right)$. In terms of recovery guarantees, for least-squares regression without Gaussian noise, SVAM and other methods (Bhatia, Jain, and Kar 2015; Mukhoty et al. 2019)) offer exact recovery of $\mathbf{w}^*$ so long as the fraction of corrupted points is less than the breakdown point while SEVER's error continues to be bounded away from zero. RGD only considers an oblivious/Huber adversary while SVAM can tolerate partially/fully adaptive adversaries. SEVER does not report an explicit breakdown point, RGD offers a breakdown point of $\alpha = 1/\log d$ (see Thm 2 in their paper) while SVAM offers an explicit breakdown point independent of d. SVAM also offers faster convergence than existing methods such as SEVER and RGD.

The SVAM Algorithm

A popular approach in robust learning is to assign weights to data points, hoping that large weights would be given to uncorrupted points and low weights to corrupted ones, followed by weighted likelihood maximization. Often the weights are updated, and the process is repeated. (Cantoni and Ronchetti 2001) use Huber style weighing functions used in Mallow's type M-estimators, (Mukhoty et al. 2019) use truncated inverse residuals, and (Valdora and Yohai 2014) use Mahalanobis distance-based weights.

SVAM notes that the label likelihood offers a natural measure of how likely the point is to be uncorrupted. Given a model estimate $\hat{\mathbf{w}}^t$ at iteration t, the weight $s_i = \mathbb{P}\left[y_i \,|\, \eta_i^t\right] = \exp(y_i \cdot \eta_i^t - \psi(\eta_i^t) - h(y_i))$ can be assigned to the i^{th} point where $\eta_i^t = \left\langle \hat{\mathbf{w}}^t, \mathbf{x}^i \right\rangle$. This gives us the weighted MLE[1] $\tilde{Q}(\mathbf{w} \,|\, \hat{\mathbf{w}}^t) = -\sum_{i=1}^n s_i \cdot \log \mathbb{P}\left[y_i \mid \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right]$ solving which gives us the next model iterate as

$$\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \tilde{Q}(\mathbf{w} \,|\, \hat{\mathbf{w}}^t) \qquad (1)$$

However, as § will show, this strategy does not perform well.

[1] Recall that for gamma/Poisson regression we need to set $\eta_i^t = \exp(\left\langle \hat{\mathbf{w}}^t, \mathbf{x}^i \right\rangle)$ given the non-canonical link for these problems.

If the initial model $\hat{\mathbf{w}}^1$ is far from $\mathbf{w}^*$, it may result in imprecise weights s_i that are large for the corrupted points. For example, if the adversary introduces corruptions using a different model $\tilde{\mathbf{w}}$ i.e. $\tilde{y}_{i_j} \sim \mathbb{P}\left[y_i \mid \left\langle \tilde{\mathbf{w}}, \mathbf{x}^{i_j} \right\rangle\right], j \in [k]$ and we happen to initialize close to $\tilde{\mathbf{w}}$ i.e. $\hat{\mathbf{w}}^1 \approx \tilde{\mathbf{w}}$, then it is the corrupted points that would get large weights initially that may cause the algorithm to converge to $\tilde{\mathbf{w}}$ itself.

Key Idea: It is thus better to avoid drastic decisions, say setting $s_i \gg 0$ in the initial stages no matter how much a data point appears to be clean. SVAM implements this intuition by setting weights using a label likelihood distribution with very large variance initially. This ensures that no data point (not even the uncorrupted ones) gets large weight (c.f. the uniform distribution that has large variance and assigns to point a high density). As SVAM progresses towards $\mathbf{w}^*$, it starts using likelihood distributions with progressively lower variance. Note that this allows data points (hopefully the uncorrupted ones) to get larger weights (c.f. the Dirac delta distribution that has vanishing variance and assigns high density to isolated points).

Mode-Preserving Variance-Altering Likelihood Transformations

To implement the above strategy, SVAM (Algorithm 1) needs techniques to alter the variance of a likelihood distribution at will. Note that the likelihood values of the altered distributions must be computable as they will be used as weights s_i i.e. merely being able to sample the distribution is not enough. Moreover, the transformation must be order-preserving – say the original and transformed distributions are $\mathbb{P}$ and $\tilde{\mathbb{P}}$ resp., then for every pair of labels y, y' and every parameter value η, we must have $\mathbb{P}\left[y \mid \eta\right] > \mathbb{P}\left[y' \mid \eta\right] \Leftrightarrow \tilde{\mathbb{P}}\left[y \mid \eta\right] > \tilde{\mathbb{P}}\left[y' \mid \eta\right]$. If this is not true, then SVAM could exhibit anomalous behavior.

The Transformation: If $\mathbb{P}\left[y \mid \eta\right] = \exp(y \cdot \eta - \psi(\eta) - h(y))$ is an exponential family distribution with parameter η and log-partition function $\psi(\eta) = \log \int \exp(y \cdot \eta - h(y))\, dy$, then for any $\beta > 0$, we get the variance-altered density,

$$\tilde{\mathbb{P}}_\beta\left[y \mid \eta\right] = \frac{1}{Z(\eta, \beta)} \exp(\beta \cdot (y \cdot \eta - \psi(\eta) - h(y))),$$

where $Z(\eta, \beta) = \int \exp(\beta \cdot (y \cdot \eta - \psi(\eta) - h(y)))\, dy$. This transformation is order and mode preserving since x^β is an increasing function for any $\beta > 0$. This generalized likelihood distribution has variance (Nelder and Wedderburn 1972) $\frac{1}{\beta}\nabla^2\psi(\eta)$, which tends to 0 as $\beta \to \infty$. Table 1 lists a few popular distributions, their variance altered versions, and asymptotic versions as $\beta \to \infty$.

We note that (Jiang, Kulis, and Jordan 2012) also study variance altering transformations for learning hidden Markov models, topic models, etc.. However, their transformations are unsuitable for use in SVAM for a few reasons:

1. SVAM's transformed distributions are always available in closed form whereas those of (Jiang, Kulis, and Jordan 2012) are not necessarily available in closed form.

2. SVAM's transformations are *order*-preserving while (Jiang, Kulis, and Jordan 2012) offer *mean*-preserving that are not assured to be order-preserving.

Algorithm 1: SVAM: Sequential Variance-Altered MLE

```
Input: Data $\left\{(\mathbf{x}^i, y_i)\right\}_{i=1}^n$, initial model $\hat{\mathbf{w}}^1$, initial scale $\beta_1$,
    scale increment $\xi > 1$, likelihood dist. $\mathbb{P}\left[\cdot \mid \cdot\right]$
for $t = 1, 2, \ldots, T-1$ do
    $s_i^t \leftarrow \tilde{\mathbb{P}}_{\beta_t}[y_i \mid \left\langle \hat{\mathbf{w}}^t, \mathbf{x}^i \right\rangle]$   // $\beta_t$-var altered $\mathbb{P}[\cdot \mid \cdot]$
    $\tilde{Q}_{\beta_t}(\mathbf{w} \mid \hat{\mathbf{w}}^t) \overset{\text{def}}{=} -\sum_{i=1}^n s_i^t \cdot \log \mathbb{P}\left[y_i \mid \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right]$
    $\hat{\mathbf{w}}^{t+1} = \arg\min_{\mathbf{w}} \tilde{Q}_{\beta_t}(\mathbf{w} \mid \hat{\mathbf{w}}^t)$
    $\beta_{t+1} \leftarrow \xi \cdot \beta_t$     // Variance of $\tilde{\mathbb{P}}_\beta[\cdot \mid \cdot] \downarrow$ as $\beta \uparrow$
end for
return $\hat{\mathbf{w}}^T$
```

The Algorithm: As presented in Algorithm 1, SVAM repeatedly constructs weighted MLEs $\tilde{Q}_\beta(\mathbf{w} \mid \hat{\mathbf{w}}^t)$ that take *β-variance altered* weights $s_i = \tilde{\mathbb{P}}_\beta[y_i \mid \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle]$ for all $i \in [n]$ and solves them to get new model estimates.

We take a pause to assert that whereas the approach in (Mukhoty et al. 2019), although similar at first to Eq (1), applies only to least-squares regression as it relies on notions of residuals missing from other GLMs. In contrast, SVAM works for all GLMs e.g. least-squares/logistic/gamma regression and offers stronger theoretical guarantees.

Theorem 1 shows that SVAM enjoys a linear rate of convergence. However, we first define notions of *Local Weighted Strong Convexity and Lipschitz Continuity*. Let $\mathcal{B}_2(\mathbf{v}, r) := \{\mathbf{w} : \left\|\mathbf{w} - \mathbf{v}\right\|_2 \leq r\}$ denote the L_2 ball of radius r centered at the vector $\mathbf{v} \in \mathbb{R}^d$.

Definition 1 (LWSC/LWLC). *Given data $\left\{(\mathbf{x}^i, y_i)\right\}_{i=1}^n$ and $\beta > 0$ an exponential family distribution $\mathbb{P}\left[\cdot \mid \cdot\right]$ is said to satisfy λ_β-Local Weighted Strongly Convexity and Λ_β-Local Weighted Lipschitz Continuity if for any* true *model $\mathbf{w}^*$ and any $\mathbf{u}, \mathbf{v} \in \mathcal{B}_2\left(\mathbf{w}^*, \sqrt{\frac{1}{\beta}}\right)$ the following hold,*

1. $\nabla^2 \tilde{Q}_\beta(\mathbf{v} \mid \mathbf{u}) \overset{\text{def}}{=} \nabla^2 \tilde{Q}_\beta(\cdot \mid \mathbf{u})\Big|_{\mathbf{v}} \succeq \lambda_\beta \cdot I$

2. $\left\|\nabla \tilde{Q}_\beta(\mathbf{w}^* \mid \mathbf{u})\right\|_2 \overset{\text{def}}{=} \left\|\nabla \tilde{Q}_\beta(\cdot \mid \mathbf{u})\Big|_{\mathbf{w}^*}\right\|_2 \leq \Lambda_\beta$

The above requires the $\tilde{Q}_\beta$-function to be strongly convex and Lipschitz continuous in a ball of radius $\frac{1}{\sqrt{\beta}}$ around the true model $\mathbf{w}^*$ i.e. as $\beta \uparrow$, the neighborhood in which these properties are desired also shrinks. We will show that likelihood functions corresponding to GLMs e.g., least squares and gamma regression satisfy these properties for appropriate ranges of β, even in the presence of corrupted samples.

Theorem 1 (SVAM Convergence). *If the data and likelihood distribution satisfy the LWSC/LWLC properties for all $\beta \in (0, \beta_{\max}]$ and if* SVAM *is initialized at $\hat{\mathbf{w}}^1$ and scale $\beta_1 > 0$ s.t. $\beta_1 \cdot \left\|\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\|_2^2 \leq 1$, then for any $\epsilon > 1/\beta_{\max}$, for small-enough scale increment $\xi > 1$,* SVAM *ensures $\left\|\hat{\mathbf{w}}^T - \mathbf{w}^*\right\|_2^2 \leq \epsilon$ after $T = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ iterations.*

It is useful to take a moment to analyze this result. Note that if the LWSC/LWLC properties hold for larger values of β, SVAM is able to offer smaller model recovery errors. Lets take least-squares regression with hybrid noise (see §) as an example. The proofs will show

Name	Standard Form (Mass/Density function)	Variance Altered Form (β)	Variance	Asymptotic Form ($\beta \to \infty$)
Gaussian (univariate) $\mathcal{N}(y \,\vert\, \eta)$	$\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}\left(y-\eta\right)^2)$	$\sqrt{\frac{\beta}{2\pi}}\exp(-\frac{\beta}{2}\left(y-\eta\right)^2)$	$\frac{1}{\beta}$	$\delta_\eta(y)$
Gaussian (multivariate) $\mathcal{N}(\mathbf{y} \,\vert\, \boldsymbol{\eta})$	$\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\exp(-\frac{1}{2}\left\Vert\mathbf{y}-\boldsymbol{\eta}\right\Vert_2^2)$	$\left(\frac{\beta}{2\pi}\right)^{\frac{d}{2}}\exp(-\frac{\beta}{2}\left\Vert\mathbf{y}-\boldsymbol{\eta}\right\Vert_2^2)$	$\frac{1}{\beta}$	$\delta_{\boldsymbol{\eta}}(\mathbf{y})$
Bernoulli $y \in \{-1,+1\}$	$\mathbb{P}\left[y=1 \,\vert\, \eta\right]=\pi$ $\pi=(1+\exp(-y\eta))^{-1}$	$\mathbb{P}\left[y=1 \,\vert\, \eta\right]=\tilde{\pi}$ $\tilde{\pi}=(1+\exp(-\beta y\eta))^{-1}$	$< \frac{1}{\beta\eta}$	$\delta_{\text{sign}(\eta)}(y)$
Gamma $\mathcal{G}(y \,\vert\, \eta,\phi)$	$\frac{1}{y\Gamma(\frac{1}{\phi})}\left(\frac{y\eta}{\phi}\right)^{\frac{1}{\phi}}\exp(-\frac{y\eta}{\phi})$ $\phi < 1$ **Note**: $\eta = \exp(\langle\mathbf{w},\mathbf{x}\rangle)$	$\frac{1}{y\Gamma(\frac{1}{\tilde{\phi}_\beta})}\left(\frac{y\tilde{\eta}_\beta}{\tilde{\phi}_\beta}\right)^{\frac{1}{\tilde{\phi}_\beta}}\exp(-\frac{y\tilde{\eta}_\beta}{\tilde{\phi}_\beta})$ $\tilde{\phi}_\beta = \phi/\left(\phi+\beta(1-\phi)\right)$ $\tilde{\eta}_\beta = \eta\beta/\left(\phi+\beta(1-\phi)\right)$	$\frac{\phi}{\eta^2}\frac{\phi+\beta(1-\phi)}{\beta^2}$	$\delta_{\frac{1-\phi}{\eta}}(y)$

Table 1: Some common distributions and their variance altered forms. Note that in all cases, the form of the distribution is preserved after transformation, as well as that the variance asymptotically goes down at the rate $\Theta(1/\beta)$ as $\beta \to \infty$.

that LWSC/LWLC properties are assured for β as large as $\beta_{\max} = \widetilde{\mathcal{O}}\left(\min\left\{\frac{1}{\alpha^{2/3}}, \sqrt{\frac{n}{d}}\right\}\right)$ (see §). Thus, with proper initialization of $\hat{\mathbf{w}}^1, \xi$ and β_1 (discussed below), SVAM ensures $\left\Vert\hat{\mathbf{w}}^T - \mathbf{w}^*\right\Vert_2^2 \leq \widetilde{\mathcal{O}}\left(\max\left\{\alpha^{2/3}, \sqrt{\frac{d}{n}}\right\}\right)$ within $T = \mathcal{O}\left(\ln(n)\right)$ steps. This proof will hold so long as SVAM is offered at least $n = \Omega(d \log d)$ training samples.

Initialization: SVAM needs to be invoked with $\hat{\mathbf{w}}^1, \beta_1$ that satisfy the requirements of Thm 1 and small enough ξ. If we initialize at the origin i.e. $\hat{\mathbf{w}}^1 = \mathbf{0}$, then Theorem 1's requirement translates to $\beta_1 \leq \frac{1}{\left\Vert\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\Vert_2^2}$ i.e. we need only find a small enough β_1. Thus, SVAM needs to tune two scalars ξ, β_1 to take small enough values which it does as described below. In practice, SVAM offered stable performance for a wide range of β_1, ξ (see Fig 1).

Hyperparameter Tuning: SVAM's two hyperparameters β_1, ξ were tuned using a held-out validation set. As the validation data could also contain corruptions, validation error was calculated by rejecting the top α fraction of validation points with the highest prediction error. The true value of α was provided to competitor algorithms as a handicap but not to SVAM. Thus, α itself was treated as a (third) hyperparameter for SVAM.

Robust GLM Applications with SVAM

This section adapts SVAM to robust least-squares/gamma/logistic regression and robust mean estimation and establishes breakdown points and LWSC/LWLC guarantees for their respective $\tilde{Q}_\beta$ functions (see Defn 1). We refer the reader to § for definitions of **partially/fully adaptive** adversaries.

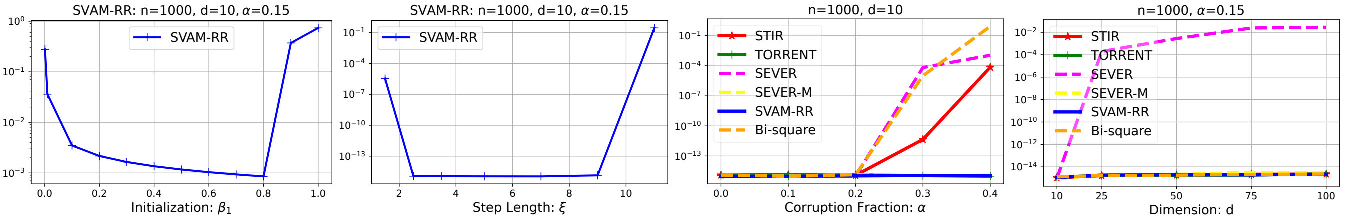
Robust Least Squares Regression. We have n data points $(\mathbf{x}^i, y_i)$, $\mathbf{x}^i \in \mathbb{R}^d$ sampled from a subGaussian distribution $\mathcal{D}$ over $\mathbb{R}^d$. We consider the **hybrid corruption** setting where on the $G = (1-\alpha)\cdot n$ "good" data points, we get labels $y_i = \left\langle\mathbf{w}^*, \mathbf{x}^i\right\rangle + \epsilon_i$ with Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \frac{1}{\beta^*})$ with variance $\frac{1}{\beta^*}$ added. On the remaining $B = \alpha\cdot n$ "bad" points, we get adversarially corrupted labels $\tilde{y}_i = \left\langle\mathbf{w}^*, \mathbf{x}^i\right\rangle + b_i$ where $b_i \in \mathbb{R}$ is chosen by the adversary. Note that b_i can be unbounded. We also consider the **pure corruption** setting where clean points receive no Gaussian noise (note that this corresponds to $\beta^* = \infty$). SVAM-RR (Alg. 2) adapts SVAM to the task of robust regression.

Theorem 2 (Partially Adaptive Adversary)**.** *For* hybrid corruptions *by a partially adaptive adversary with corruption rate* $\alpha \leq 0.18$*, there exist* $\xi > 1$ *s.t. with probability at least* $1 - \exp(-\Omega\left(d\right))$*, LWSC/LWLC properties are satisfied for the* $\tilde{Q}_\beta$ *function for* β *values as large as* $\beta_{\max} = \mathcal{O}\left(\beta^* \min\left\{\frac{1}{\alpha^{2/3}}, \sqrt{\frac{n}{d\log(n)}}\right\}\right)$*. If initialized with* $\hat{\mathbf{w}}^1, \beta^1$ *s.t.* $\beta_1 \cdot \left\Vert\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\Vert_2^2 \leq 1$, SVAM-RR *assures* $\left\Vert\hat{\mathbf{w}}^T - \mathbf{w}^*\right\Vert_2^2 \leq \mathcal{O}\left(\frac{1}{\beta^*}\max\left\{\alpha^{2/3}, \sqrt{\frac{d\log(n)}{n}}\right\}\right)$ *within* $T \leq \mathcal{O}\left(\log\frac{n}{\beta^1}\right)$ *iterations. For* pure corruptions *by a partially adaptive adversary, we have* $\beta_{\max} = \infty$ *and thus, for any* $\epsilon > 0$, SVAM-RR *assures* $\left\Vert\hat{\mathbf{w}}^T - \mathbf{w}^*\right\Vert_2^2 \leq \epsilon$ *within* $T \leq \mathcal{O}\left(\log\frac{1}{\epsilon\beta^1}\right)$ *iterations.*

Note that in the pure corruption setting, SVAM assures exact recovery of $\mathbf{w}^*$ simply by running the algorithm long enough. This is not a contradiction since in this case, the LWSC/LWSS properties can be shown to hold for all values of $\beta < \infty$ since we effectively have $\beta^* = \infty$ in this case. Thm 2 holds against a partially adaptive adversary but can be extended to a fully adaptive adversary as well but at the cost of a worse breakdown point (see Thm 3 below). Note that SVAM continues to assure exact recovery of $\mathbf{w}^*$.

Theorem 3 (Fully Adaptive Adversary)**.** *For pure corruptions by a fully adaptive adversary with corruption rate* $\alpha \leq 0.0036$*, LWSC/LWLC are satisfied for all* $\beta \in (0, \infty)$ *i.e.* $\beta_{\max} = \infty$ *and for any* $\epsilon > 0$, SVAM-RR *assures* $\left\Vert\hat{\mathbf{w}}^T - \mathbf{w}^*\right\Vert_2^2 \leq \epsilon$ *within* $T \leq \mathcal{O}\left(\log\frac{1}{\epsilon\beta^1}\right)$ *iterations if initialized as described in the statement of Theorem 2.*

Establishing LWSC/LWLC: In the appendices, Lem-



(a) β_1 Sensitivity

(b) ξ Sensitivity

(c) α Sensitivity

(d) d Sensitivity

Figure 1: SVAM offers stable convergence and recovery, $\left\|\hat{\mathbf{w}}^T - \mathbf{w}^*\right\|_2$, superior to competitor algorithms for a wide range of hyperparameters β_1, ξ, corruption rates $\alpha = k/n$ and feature dimensionality d.

mata 15 and 16 establish LWSC/LWLC properties for robust least squares regression while Theorems 14 and 21 establish the breakdown points and existence of suitable increments $\xi > 1$. Handling a fully adaptive adversary requires making mild modifications to the notions of LWSC/LWLC, details of which are present in Appendix G.1.

Model Recovery and Breakdown Point: For pure corruption, SVAM-RR offers exact model recovery against partially and fully adaptive adversaries as it assures $\left\|\hat{\mathbf{w}}^T - \mathbf{w}^*\right\|_2^2 \leq \epsilon$ for any $\epsilon > 0$ if SVAM-RR is executed long enough. For hybrid corruption where even "clean" points receive Gaussian noise with variance $\frac{1}{\beta^*}$, SVAM-RR assures $\left\|\hat{\mathbf{w}}^T - \mathbf{w}^*\right\|_2^2 \leq \mathcal{O}\left(\frac{1}{\beta^*}\sqrt{\frac{d\log(n)}{n}}\right)$ as $\alpha \to 0$ i.e. $\left\|\hat{\mathbf{w}}^T - \mathbf{w}^*\right\|_2^2 \to 0$ as $n \to \infty$ assuring consistent recovery. This significantly improves previous results by (Bhatia, Jain, and Kar 2015; Mukhoty et al. 2019) which offer $\mathcal{O}\left(\frac{1}{\beta^*}\right)$ error even if $\alpha \to 0$ and $n \to \infty$. Note that SVAM-RR has a superior breakdown point (allowing upto 18% corruption rate) against an oblivious adversary. The breakdown point deteriorates as expected (still allowing upto 0.36% corruption rate) against a fully adaptive adversary. We now present analyses for other GLM problems.

Robust Gamma Regression. The data generation and corruption model for gamma regression are slightly different given that the gamma distribution has support only over positive reals. First, the canonical parameter is calculated as $\eta_i = \exp(\langle\mathbf{w}^*, \mathbf{x}^i\rangle)$ using which a clean label y^i is generated. To simplify the analysis, we assume that $\|\mathbf{w}^*\|_2 = 1$, $\phi = 0.5$, $\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, I)$. For the $G = (1-\alpha)\cdot n$ "good" points, labels are generated as $y_i = \exp(\langle\mathbf{w}^*, \mathbf{x}^i\rangle)(1-\phi)$ i.e. the *no-noise* model. For the remaining $B = \alpha \cdot n$ "bad" points, the label is corrupted as $\tilde{y}^i = y^i \cdot b_i$ where $b_i > 0$ is a positive real number (but otherwise arbitrary and unbounded). A multiplicative corruption makes more sense since the final label must be positive. SVAM-GAMMA (Algorithm 4) adapts SVAM to robust gamma regression. Due to the alternate canonical parameter used in gamma regression, the initialization requirement also needs to be modified to $\beta_1 \cdot \left(\exp\left(\left\|\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\|_2\right) - 1\right)^2 \leq 1$. However, the hyperparameter tuning strategy discussed in § continues to apply.

Theorem 4. *For data corrupted by a partially adaptive adversary with* $\alpha \leq \frac{0.002}{\sqrt{d}}$, *there exist* $\xi > 1$ *s.t. with probability at least* $1 - \exp(-\Omega(d))$, *LWSC/LWLC conditions are satisfied for the* $\tilde{Q}_\beta$ *function for* β *values as large as* $\beta_{\max} = \mathcal{O}\left(1/\left(\exp\left(\mathcal{O}\left(\alpha\sqrt{d}\right)\right) - 1\right)\right)$. *If initialized at* $\hat{\mathbf{w}}^1, \beta_1$ *s.t.* $\beta_1 \cdot \left(\exp\left(\left\|\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\|_2\right) - 1\right)^2 \leq 1$ *and* $\beta \geq 1$, SVAM-GAMMA *assures* $\left\|\hat{\mathbf{w}}^T - \mathbf{w}^*\right\|_2 \leq \epsilon$ *for any* $\epsilon \geq \mathcal{O}\left(\alpha\sqrt{d}\right)$ *within* $T \leq \mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ *steps.*

Model recovery, Consistency, Breakdown pt. It is notable that prior results in literature do not offer any breakdown point results for gamma regression. We find that Thm 4 requires $\beta_1 \cdot \left(\exp\left(\left\|\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\|_2\right) - 1\right)^2 \leq 1$ and $\beta \geq 1$ which imply $\left\|\hat{\mathbf{w}}^1 - \mathbf{w}^*\right\|_2 \leq \ln 2$. This is in contrast to Thms 2 and 3 that allow any initial $\hat{\mathbf{w}}^1$ so long as β_1, ξ are sufficiently small. SVAM-GAMMA guarantees convergence to a region of radius $\mathcal{O}\left(\alpha\sqrt{d}\right)$ around $\mathbf{w}^*$ whereas Thms 2 and 3 assure exact recovery. However, these do not seem to be artifacts of the proof technique. In experiments, SVAM-GAMMA did not offer vanishingly small recovery errors and did indeed struggle if initialized with $\beta_1 \ll 1$. It may be the case that there exist lower bounds preventing exact recovery for gamma regression similar to mean estimation.

Robust Mean Estimation. We have n data points of which the set G of $(1-\alpha)\cdot n$ "good" points are generated from a d-dimensional spherical Gaussian $\mathbf{x}^i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ i.e. $\mathbf{x}^i = \boldsymbol{\mu} + \boldsymbol{\epsilon}^i$ where $\boldsymbol{\epsilon}^i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and $\Sigma = \frac{1}{\beta^*} \cdot I$ for some $\beta^* > 0$. The rest are the set B of $\alpha \cdot n$ "bad" points that are corrupted by an adversary i.e. $\tilde{\mathbf{x}}^i = \boldsymbol{\mu}^* + \mathbf{b}^i$ where $\mathbf{b}^i \in \mathbb{R}^d$ can be unbounded. SVAM-ME (Algorithm 3) adapts SVAM to the robust mean estimation problem. For notational clarity we use, $\boldsymbol{\eta} = \boldsymbol{\mu}$, in this problem.

Theorem 5. *For data corrupted by a partially adaptive adversary with corruption rate* $\alpha \leq 0.26$, *there exists* $\xi > 1$ *s.t. with probability at least* $1 - \exp(-\Omega(d))$, *LWSC/LWLC conditions are satisfied for the* $\tilde{Q}_\beta$ *function for* β *upto* $\beta_{\max} = \mathcal{O}\left(\frac{\beta^*}{d}\min\left\{\log\frac{1}{\alpha}, \sqrt{nd}\right\}\right)$. *If initialized with* $\hat{\boldsymbol{\mu}}^1, \beta^1$ *s.t.* $\beta_1 \cdot \left\|\hat{\boldsymbol{\mu}}^1 - \boldsymbol{\mu}^*\right\|_2^2 \leq 1$, SVAM-ME *assures* $\left\|\hat{\boldsymbol{\mu}}^T - \boldsymbol{\mu}^*\right\|_2^2 \leq \epsilon$ *for any* $\epsilon \geq \mathcal{O}\left(\textit{trace}^2(\Sigma) \cdot \max\left\{\frac{1}{\ln(1/\alpha)}, \frac{1}{\sqrt{nd}}\right\}\right)$ *within* $T \leq \mathcal{O}\left(\log\frac{n}{\beta^1}\right)$ *iterations.*

Model recovery, Consistency, Breakdown pt. Note that

Algorithm 2: SVAM-RR: Robust Least Squares Regression

```
Input: Data $\{(\mathbf{x}^i, y_i)\}_{i=1}^n$, initial scale $\beta_1$, initial model $\hat{\mathbf{w}}^1$, $\xi$
Output: A model estimate $\hat{\mathbf{w}} \approx \mathbf{w}^*$
for $t = 1, 2, \ldots, T-1$ do
    $s_i \leftarrow \exp\left(-\frac{\beta_t}{2}(y_i - \left\langle \mathbf{x}^i, \hat{\mathbf{w}}^t \right\rangle)^2\right)$
    $S \leftarrow \mathrm{diag}(s_1, \ldots, s_n)$
    $\hat{\mathbf{w}}^{t+1} \leftarrow (XSX^\top)^{-1}(XS\mathbf{y})$
    $\beta_{t+1} \leftarrow \xi \cdot \beta_t$
end for
return $\hat{\mathbf{w}}^T$
```

Algorithm 3: SVAM-ME: Robust Mean Estimation

```
Input: Data $\{\mathbf{x}^i\}_{i=1}^n$, initial scale $\beta_1$, initial model $\hat{\boldsymbol{\mu}}^1$, $\xi$
Output: A mean estimate $\hat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}^*$
for $t = 1, 2, \ldots, T-1$ do
    $s_i \leftarrow \exp\left(-\frac{\beta_t}{2}\left\|\mathbf{x}^i - \hat{\boldsymbol{\mu}}^t\right\|_2^2\right)$
    $\hat{\boldsymbol{\mu}}^{t+1} \leftarrow \left(\sum_{i=1}^n s_i\right)^{-1}\left(\sum_{i=1}^n s_i \mathbf{x}^i\right)$
    $\beta_{t+1} \leftarrow \xi \cdot \beta_t$
end for
return $\hat{\boldsymbol{\mu}}^T$
```

Algorithm 4: SVAM-GAMMA: Robust Gamma Regression

```
Input: Data $\{(\mathbf{x}^i, y_i)\}_{i=1}^n$, initial scale $\beta_1$, initial model $\hat{\mathbf{w}}^1$, $\xi$
Output: A model estimate $\hat{\mathbf{w}} \approx \mathbf{w}^*$
for $t = 1, 2, \ldots, T-1$ do
    $s_i \leftarrow \mathcal{G}(y_i \,|\, \tilde{\eta}_{\beta_t}, \tilde{\phi}_{\beta_t})$                //see Table 1
    $\hat{\mathbf{w}}^{t+1} \leftarrow \arg\min \sum_{i=1}^n s_i \cdot \ell(\mathbf{w}, \mathbf{x}^i, y_i)$ where
    $\ell(\mathbf{w}, \mathbf{x}, y) = (1-\phi)^{-1} y \exp(\langle \mathbf{w}, \mathbf{x} \rangle) - \langle \mathbf{w}, \mathbf{x} \rangle$
    $\beta_{t+1} \leftarrow \xi \cdot \beta_t$
end for
return $\hat{\mathbf{w}}^T$
```

Algorithm 5: SVAM-LR: Robust Classification

```
Input: Data $\{(\mathbf{x}^i, y_i)\}_{i=1}^n$, initial scale $\beta_1$, initial model $\hat{\mathbf{w}}^1$, $\xi$
Output: A model estimate $\hat{\mathbf{w}} \approx \mathbf{w}^*$
for $t = 1, 2, \ldots, T-1$ do
    $s_i \leftarrow (1 + \exp(-\beta_t y_i \left\langle \mathbf{x}^i, \hat{\mathbf{w}}^t \right\rangle))^{-1}$
    $\hat{\mathbf{w}}^{t+1} \leftarrow \arg\min \sum_{i=1}^n s_i \cdot \ell(\mathbf{w}, \mathbf{x}^i, y_i)$ where
    $\ell(\mathbf{w}, \mathbf{x}, y) = \log(1 + \exp(-y \langle \mathbf{x}, \mathbf{w} \rangle))$
    $\beta_{t+1} \leftarrow \xi \cdot \beta_t$
end for
return $\hat{\mathbf{w}}^T$
```

for any constant $\alpha > 0$, the estimation error does not go to zero as $n \to \infty$. As mentioned in §, an error of $\Omega(\alpha)$ is unavoidable no matter how large n gets. Thus, the best hope we have is of the estimation error going to zero as $\alpha \to 0$ and $n \to \infty$. The error in Theorem 5 does indeed go to zero in this setting. Also, note that the error depends only on the trace of the covariance matrix of the clean points, and thus for trace$(\Sigma) = \mathcal{O}(1)$, the result offers an estimation error independent of dimension. SVAM-ME offers a large breakdown point (allowing upto 26% corruption rate).

Establishing LWSC/LWLC for Gamma Regression and Mean Estimation: In the appendices, Lemmata 28, 29, Lemmata 23, 24 establish the LWSC/LWLC properties for the $\tilde{Q}_\beta$ function for gamma regression and mean estimation and Theorems 27 and Theorem 22 establish the breakdown points and existence of increments $\xi > 1$.

Robust Classification. In this case the labels are generated as $y_i = \mathrm{sign}(\left\langle \mathbf{w}^*, \mathbf{x}^i \right\rangle)$ and the bad points in the set B get their labels flipped $\tilde{y}_i = -\mathrm{sign}(\left\langle \mathbf{w}^*, \mathbf{x}^i \right\rangle)$. SVAM-LR (Algorithm 5) adapts SVAM to robust logistic regression.

Experiments

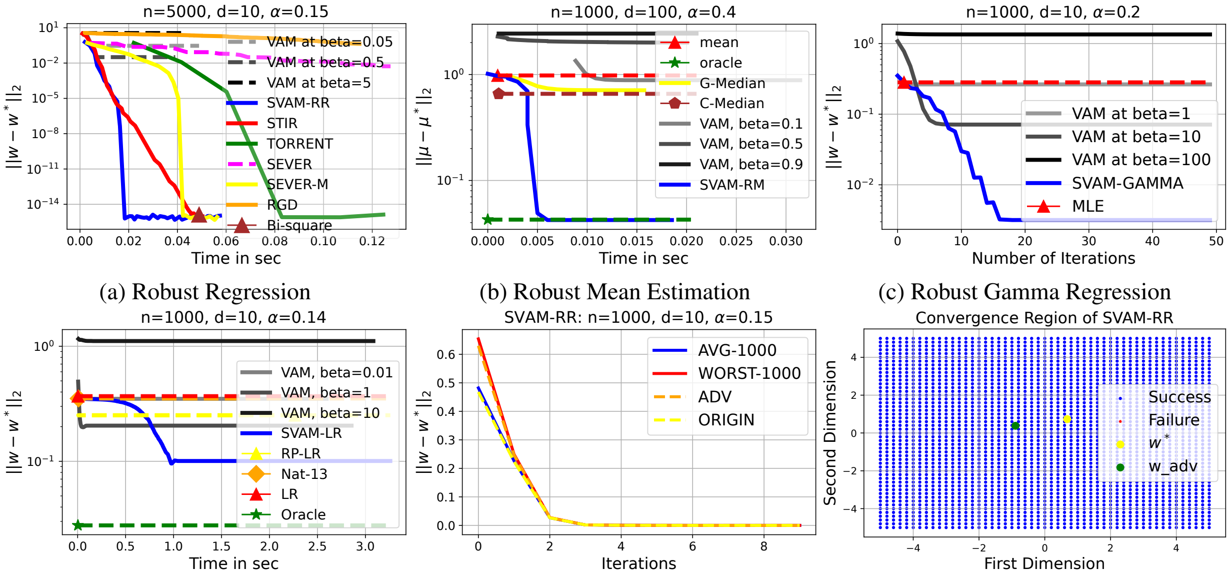
We used a 64-bit machine with Intel® Core™ i7-6500U CPU @ 2.50GHz, 4 cores, 16 GB RAM, Ubuntu 16.04 OS.

Benchmarks. SVAM was benchmarked against several baselines **(a) VAM**: this is SVAM executed with a fixed value of the scale β by setting the scale increment to $\xi = 1$ to investigate any benefits of varying the scale β, **(b) MLE**: likelihood maximization on all points (clean + corrupted) without any weights assigned to data points – this checks for benefits of performing weighted MLE, **(c) Oracle**: an execution of the MLE only on the clean points – this is the gold standard in robust learning and offers the best possible outcome. In addition, several **problem-specific competitors** were also considered. For robust regression, STIR (Mukhoty et al. 2019), TORRENT (Bhatia, Jain, and Kar 2015), SEVER (Diakonikolas et al. 2019b), RGD (Prasad et al. 2018), and the classical robust M-estimator of Tukey's bisquare loss were included. Note that TORRENT already outperforms L_1 regularization methods while achieving better or competitive recovery errors (see (Bhatia, Jain, and Kar 2015, Fig 2(b))). Since SVAM-RR was faster than TORRENT itself, L_1 regularized methods such as (Nguyen and Tran 2013; Wright and Ma 2010) were not considered. For robust mean estimation, popular baselines such as coordinate-wise median and geometric median were taken. For robust classification, the rank-pruning method RP-LR (Northcutt, Wu, and Chuang 2017) and the method from (Natarajan et al. 2013) were used.

Experimental Setting and Reproducibility. Due to lack of space, details of experimental setup, data generation, how adversaries were simulated etc are presented in Appendix C. SVAM also offered superior robustness than competitors against a wide range of ways to simulate adversarial corruption (see Appendix D for details). Code for SVAM is available at https://github.com/purushottamkar/svam/.

Experimental Observations

Robust Regression. Fig 2(a) shows that SVAM-RR, SEVER, RGD, STIR, and TORRENT are competitive and achieve oracle-level error. However, SVAM-RR can be twice as fast in terms of execution time. Since TORRENT itself outperforms L_1 regularization methods while achieving better or competitive recovery errors (see Fig 2(b) in (Bhatia, Jain, and Kar 2015)), we do not compare against L_1 methods. SVAM-RR is several times faster than classical robust M-estimators such as Tukey's bisquare loss. Also, no single value of β can offer the performance of SVAM, as is indicated by the poor performance of VAM. Fig 4 in the appendix shows that this is true even if very large or very small values of β are used with VAM. We note that SEVER chooses a threshold in each iteration to eliminate specific points as corrupted. This threshold is chosen randomly (possibly for ease of proof) but causes SEVER to of-



(a) Robust Regression

(b) Robust Mean Estimation

(c) Robust Gamma Regression

(d) Robust Classification

(e) Adversarial Initialization

(f) Global Convergence

Figure 2: (a,b,c,d) compare SVAM and various competitors. The number of data points n, dimensions d, and fraction of corruptions $\alpha = k/n$ are mentioned at the top of each figure. The figures show that VAM with a single fixed value of β cannot replace the gradual variations in β_t as done by SVAM. Figs 2(e,f) confirm that SVAM offers convergence to $\mathbf{w}^*$ irrespective of the model initialization. Corruptions are introduced using an adversarial model $\tilde{\mathbf{w}}$ i.e. for corrupted points, the label was set to $\tilde{y}_i = \langle \tilde{\mathbf{w}}, \mathbf{x}_i \rangle$. SVAM was then initialized at $\tilde{\mathbf{w}}$ itself for faulty initialization but was found to offer exact recovery regardless.

fer sluggish convergence. Thus, we also report the performance of a modification SEVER-M that was given an unfair advantage by revealing to it the actual number of corrupted points (SVAM was not given this information). This sped-up SEVER but SVAM continued to outperform SEVER-M. Fig 3 in the appendix reports repeated runs of the experiment where SVAM continues to lead.

Robust Logistic and Gamma Regression. Fig 2(c,d) report results of SVAM on robust gamma and logistic regression problems. The figures show that executing VAM with a fixed value of β cannot replace the gradual variations in β_t done by SVAM. Additionally, for robust classification, SVAM-LR achieves error, an order of magnitude smaller than all competitors except the oracle. SVAM also outperforms the RP-LR (Northcutt, Wu, and Chuang 2017) and (Natarajan et al. 2013) algorithms that were specifically designed for robust classification. A horizontal dashed line is used to indicate the final performance of algorithms for which iteration-wise performance was unavailable.

Robust Mean Estimation. Fig 2(b) reports results on robust mean estimation problems. SVAM outperforms VAM with any fixed value of β as well as the naive sample mean (the MLE in this case). Popular approaches coordinate-wise median and geometric median were fast but offered poor results. SVAM on the other hand achieved oracle error-level error by assigning proper scores to all data points.

Sensitivity to Hyperparameter Tuning. In Figs 1(a,b), SVAM-RR was offered hyperparameters in a wide range of values to study how it responded when provided mis-specified hyperparameters. SVAM offered stable convergence for a wide range of β_1, ξ indicating that it is resilient to minor mis-specifications in hyperparameters.

Sensitivity to Dimension and Corruption. Figs 1(c,d) compare the error offered by various algorithms in recovering $\mathbf{w}^*$ for robust least-squares regression when the fraction of corrupted points α and feature dimension d were varied. All values are averaged over 20 experiments with each experiment using 1000 data points. α was varied in the range $[0, 0.4]$ and d in the range $[10, 100]$ with fixed hyperparameters. STIR and Bi-square are sensitive to corruption while SEVER is sensitive to both corruption and dimension. RGD is not visible in the figures as its error exceeded the figure boundaries. Experiments for Fig 1(c) fixed $d = 10$ and vary α while Fig 1(d) fixed $\alpha = 0.15$ and vary d. Figs 1(c,d) show that SVAM-RR can tolerate large fractions of the data getting corrupted and is not sensitive to d.

Testing SVAM for Global Convergence. To test the effect of initialization, in Fig 2(e), corruptions were introduced using an adversarial model $\tilde{\mathbf{w}}$ i.e. for corrupted points, labels were set to $\tilde{y}_i = \langle \tilde{\mathbf{w}}, \mathbf{x}_i \rangle$. SVAM-RR was initialized at 1000 randomly chosen models, the origin, as well as at the adversarial model $\tilde{\mathbf{w}}$ itself. WORST-1000 (resp. AVG-1000) indicate the worst (resp. average) performance SVAM had at any of the 1000 initializations. Fig 2(f) further emphasizes this using a toy 2D problem. SVAM was initialized at all points on the grid. An initialization was called a success if SVAM got error $< 10^{-6}$ within eight or fewer iterations. In all these experiments SVAM rapidly converged to the true model irrespective of model initialization.

Acknowledgements

The authors thank the anonymous reviewers of this paper for suggesting illustrative experiments and pointing to relevant literature. B.M. is supported by the Technology Innovation Institute and MBZUAI joint project (NO. TII/ARRC/2073/2021): Energy-based Probing for Spiking Neural Networks. D.D. is supported by the Research-I Foundation at IIT Kanpur and acknowledges support from the Visvesvaraya PhD Scheme for Electronics & IT (FELLOW/2016-17/MLA/194). P.K. thanks Microsoft Research India and Tower Research for research grants.

References

Bhatia, K.; Jain, P.; and Kar, P. 2015. Robust Regression via Hard Thresholding. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS).*

Cantoni, E.; and Ronchetti, E. 2001. Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*, 96(455): 1022–1030.

Cheng, Y.; Diakonikolas, I.; and Ge, R. 2019. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2755–2771. SIAM.

Dalalyan, A. S.; and Minasyan, A. 2022. All-In-One Robust Estimator of the Gaussian Mean. *Annals of Statistics*, 50(2): 1193–1219.

Diakonikolas, I.; Kamath, G.; Kane, D.; Li, J.; Moitra, A.; and Stewart, A. 2019a. Robust Estimators in High-Dimensions Without the Computational Intractability. *SIAM Journal on Computing*, 48(2): 742–864.

Diakonikolas, I.; Kamath, G.; Kane, D.; Li, J.; Steinhardt, J.; and Stewart, A. 2019b. Sever: A Robust Meta-Algorithm for Stochastic Optimization. In *36th International Conference on Machine Learning (ICML).*

Diakonikolas, I.; and Kane, D. M. 2019. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911.*

Feng, J.; Xu, H.; Mannor, S.; and Yan, S. 2014. Robust Logistic Regression and Classification. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS).*

Feng, Y.; Huang, X.; Shi, L.; Yang, Y.; and Suykens, J. A. 2015. Learning with the Maximum Correntropy Criterion Induced Losses for Regression. *Journal of Machine Learning Research*, 16(30): 993–1034.

Jiang, K.; Kulis, B.; and Jordan, M. 2012. Small-Variance Asymptotics for Exponential Family Dirichlet Process Mixture Models. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS).*

Lai, K. A.; Rao, A. B.; and Vempala, S. 2016. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 665–674. IEEE.

Lecué, G.; and Lerasle, M. 2020. Robust machine learning by median-of-means: Theory and practice. *Annals of Statistics*, 48(2): 906–931.

Li, T.; Beirami, A.; Sanjabi, M.; and Smith, V. 2021. Tilted Empirical Risk Minimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR).*

McCullagh, P.; and Nelder, J. A. 1989. *Generalized Linear Models*. Chapman and Hall.

Mukhoty, B.; Gopakumar, G.; Jain, P.; and Kar, P. 2019. Globally-convergent Iteratively Reweighted Least Squares for Robust Regression Problems. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS).*

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *Advances in neural information processing systems*, 1196–1204.

Nelder, J. A.; and Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A*, 135(3): 370–384.

Nguyen, N. H.; and Tran, T. D. 2013. Exact Recoverability From Dense Corrupted Observations via ℓ_1-Minimization. *IEEE transactions on information theory*, 59(4): 2017–2035.

Northcutt, C. G.; Wu, T.; and Chuang, I. L. 2017. Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI).*

Prasad, A.; Suggala, A. S.; Balakrishnan, S.; and Ravikumar, P. 2018. Robust Estimation via Robust Gradient Estimation. ArXiv:1802.06485 [stat.ML].

Suggala, A. S.; Bhatia, K.; Ravikumar, P.; and Jain, P. 2019. Adaptive Hard Thresholding for Near-optimal Consistent Robust Regression. In *32nd Conference on Learning Theory (COLT).*

Valdora, M.; and Yohai, V. J. 2014. Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146: 31–48.

Vershynin, R. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wright, J.; and Ma, Y. 2010. Dense Error Correction via ℓ^1 Minimization. *IEEE Transactions on Information Theory*, 56(7): 3540–3560.

Yang, E.; Tewari, A.; and Ravikumar, P. 2013. On Robust Estimation of High Dimensional Generalized Linear Models. In *23rd International Joint Conference on Artificial Intelligence (IJCAI).*