Robust Compressed Sensing using Generative Models

Ajil Jalal *
ECE, UT Austin
ajiljalal@utexas.edu

Liu Liu
ECE, UT Austin
liuliu@utexas.edu

Alexandros G. Dimakis
ECE, UT Austin
dimakis@austin.utexas.edu

Constantine Caramanis
ECE, UT Austin
constantine@utexas.edu

Abstract

The goal of compressed sensing is to estimate a high dimensional vector from an underdetermined system of noisy linear equations. In analogy to classical compressed sensing, here we assume a generative model as a prior, that is, we assume the vector is represented by a deep generative model $G : \mathbb{R}^k \to \mathbb{R}^n$. Classical recovery approaches such as empirical risk minimization (ERM) are guaranteed to succeed when the measurement matrix is sub-Gaussian. However, when the measurement matrix and measurements are heavy-tailed or have outliers, recovery may fail dramatically. In this paper we propose an algorithm inspired by the Median-of-Means (MOM). Our algorithm guarantees recovery for heavy-tailed data, even in the presence of outliers. Theoretically, our results show our novel MOM-based algorithm enjoys the same sample complexity guarantees as ERM under sub-Gaussian assumptions. Our experiments validate both aspects of our claims: other algorithms are indeed fragile and fail under heavy-tailed and/or corrupted data, while our approach exhibits the predicted robustness.

1 Introduction

Compressive or compressed sensing is the problem of reconstructing an unknown vector $x^* \in \mathbb{R}^n$ after observing $m < n$ linear measurements of its entries, possibly with added noise: $y = Ax^* + \eta$, where $A \in \mathbb{R}^{m \times n}$ is called the measurement matrix and $\eta \in \mathbb{R}^m$ is noise. Even without noise, this is an underdetermined system of linear equations, so recovery is impossible without a structural assumption on the unknown vector x^*. The vast literature [84, 37, 72, 9, 18, 27, 2, 86, 11] on this subject typically assumes that the unknown vector is "natural," or "simple," in some application-dependent way.

Compressed sensing has been studied on a wide variety of structures such as sparse vectors [19], trees [20], graphs [90], manifolds [21, 89] or deep generative models [15]. In this paper, we concentrate on deep generative models, which were explored by [15] as priors for sample-efficient reconstruction. Theoretical results in [15] showed that if x^* lies close to the range of a generative model $G : \mathbb{R}^k \to \mathbb{R}^n$ with $d-$layers, a variant of ERM can recover x^* with $m = O(kd \log n)$ measurements. Empirically, [15] shows that generative models require $5 - 10\times$ fewer measurements to obtain the same reconstruction accuracy as Lasso. This impressive empirical performance has motivated significant recent research to better understand the behaviour and theoretical limits of compressed sensing using generative priors [36, 50, 62]

A key technical condition for recovery is the Set Restricted Eigenvalue Condition (S-REC) [15], which is a generalization of the Restricted Eigenvalue Condition [14, 17] in sparse recovery. This

*Link to our code: https://github.com/ajiljalal/csgm-robust-neurips

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

condition is satisfied if A is a sub-Gaussian matrix and the measurements satisfy $y = Ax^* + \eta$. This leads to the question: can the conditions on A be weakened, and can we allow for outliers in y and A? This has significance in applications such as MRI and astronomical imaging, where data is often very noisy and requires significant pruning/cleansing.

As we show in this paper, the analysis and algorithm proposed by [15] are quite fragile in the presence of heavy-tailed noise or corruptions in the measurements. In the statistics literature, it is well known that algorithms such as empirical risk minimization (ERM) and its variants are not robust to even a *single* outlier. Since the algorithm in [15] is a variant of ERM, it is susceptible to the same failures in the presence of heavy-tails and outliers. Indeed, as we show empirically in Section 6, precisely this occurs.

Importantly, recovery failure in the setting of [15] (which is also the focus of this paper) can be pernicious, precisely because generative models (by design) output images in their range space, and for well-designed models, these have high perceptual quality. In contrast, when a classical algorithm like LASSO [84] fails, the typical failure mode is the output of a non-sparse vector. Thus in the context of generative models, resilience to outliers and heavy-tails is especially critical. This motivates the need for algorithms that do not require strong assumptions on the measurements.

In this paper, we propose an algorithm for compressed sensing using generative models, which is robust to heavy-tailed distributions and arbitrary outliers. We study its theoretical recovery guarantees as well as empirical performance, and show that it succeeds in scenarios where other existing recovery procedures fail, without additional cost in sample complexity or computation.

1.1 Contributions

We propose a new reconstruction algorithm in place of ERM. Our algorithm uses a Median-of-Means (MOM) loss to provide robustness to heavy-tails and arbitrary corruptions. As S-REC may no longer hold, we necessarily use a different analytical approach. We prove recovery results and sample complexity guarantees for this setting even though previous assumptions such as the S-REC [15] condition do not hold. Specifically, our main contributions are as follows.

- (Algorithm) We consider robust compressed sensing for generative models where (i) a constant fraction of the measurements and measurement matrix are arbitrarily (perhaps maliciously) corrupted and (ii) the random ensemble only satisfies a weak moment assumption.

  We propose a novel algorithm to replace ERM. Our algorithm uses a median-of-means (MOM) tournament [65, 54] i.e., a min-max optimization framework for robust reconstruction. Each iteration of our MOM-based algorithm comes at essentially no additional computational cost compared to an iteration of standard ERM. Moreover, as our code shows, it is straightforward to implement.

- (Analysis and Guarantees) We analyze the recovery guarantee and outlier-robustness of our algorithm when the generative model is a d-layer neural network using ReLU activations. Specifically, in the presence of a constant fraction of outliers in y and A, we achieve $\|G(\widehat{z}) - G(z^*)\|^2 \leq O(\sigma^2 + \tau)$ with sample size $m = O(kd \log n)$, where σ^2 is the variance of the heavy-tailed noise, and τ is the optimization accuracy. Using different analytical tools (necessarily, since we do not assume sub-Gaussianity), we show our algorithm, even under heavy-tails and corruptions, has the same sample complexity as the previous literature has achieved under much stronger sub-Gaussian assumptions. En route to our result, we also prove an interesting result for ERM: by avoiding the S-REC-based analysis, we show that the standard ERM algorithm does in fact succeed in the presence of a heavy-tailed measurement matrix, thereby strengthening the best-known recovery guarantees from [15]. This does not extend (as our empirical results demonstrate) to the setting of outliers, or of heavy-tailed measurement noise. For these settings, our new algorithm is required.

- (Empirical Support) We empirically validate the effectiveness of our robust recovery algorithm on MNIST and CelebA-HQ. Our results demonstrate that (as our theory predicts) our algorithm succeeds in the presence of heavy-tailed noise, heavy-tailed measurements, and also in the presence of arbitrary outliers. At the same time our experiments confirm that ERM can fail, and in fact fails dramatically: through an experiment on the CelebA-HQ data set, we demonstrate that the ERM recovery approach [15], as well as other natural approaches including ℓ_1 loss minimization and trimmed loss minimization [81], can recover images that have little resemblance to the original.

1.2 Related work

Compressed sensing with outliers or heavy-tails has a long history. To deal with outliers only in y, classical techniques replace the ERM with a robust loss function such as ℓ_1 loss or Huber loss [58, 74, 64, 24], and obtain the optimal statistical rates. Much less is known for outliers in y and A for robust compressed sensing. Recent progress on robust sparse regression [22, 10, 23, 26, 78, 60, 59, 81] can handle outliers in y and A, but their techniques cannot be directly extended to arbitrary generative models G. Another line of research [43, 70, 65, 54] considers compressed sensing where the measurement matrix A and y have heavy-tailed distributions. Their techniques leverage variants of Median-of-Means (MOM) estimators on the loss function under weak moment assumptions instead of sub-Gaussianity, which generalize the classical MOM mean estimator in one dimension [73, 48, 3, 70].

[88] deals with compressed sensing of generative models when the measurements and the responses are non-Gaussian. However, the distribution model in [88] requires more stringent conditions compared to the weak moment assumption as will be specified in Definition 1, and their algorithm cannot tolerate arbitrary corruptions.

Generative priors have shown great promise in compressed sensing and other inverse problems, starting with [15], who generalized the theoretical framework of compressive sensing and restricted eigenvalue conditions [84, 27, 14, 17, 41, 13, 12, 28] for signals lying on the range of a deep generative model [33, 53]. Results in [50, 62, 47] established that the sample complexities in [15] are order optimal. The approach in [15] has been generalized to tackle different inverse problems [35, 8, 6, 71, 7, 79, 8, 61, 5, 46, 34, 4]. Alternate algorithms for reconstruction include [16, 25, 49, 30, 29, 82, 66, 25, 77, 38, 39]. The complexity of optimization algorithms using generative models have been analyzed in [32, 40, 57, 36]. See [75] for a more detailed survey on deep learning techniques for compressed sensing. A related line of work has explored learning-based approaches to tackle classical problems in algorithms and signal processing [1, 45, 69, 42].

2 Notation

For functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ to denote that there exists a universal constant $c_1 > 0$ such that $f(n) \leq c_1 g(n)$. Similarly, we write $f(n) \gtrsim g(n)$ to denote that there exists a universal constant $c_2 > 0$ such that $f(n) \geq c_2 g(n)$. We write $f(n) = O(g(n))$ to imply that there exists a positive constant c_3 and a natural number n_0 such that for all $n \geq n_0$, we have $|f(n)| \leq c_3 g(n)$. Similarly, we write $f(n) = \Omega(g(n))$ to imply that there exists a positive constant c_4 and a natural number n_1 such that for all $n \geq n_1$, we have $|f(n)| \geq c_4 g(n)$.

3 Problem formulation

Let $x^* = G(z^*) \in \mathbb{R}^n$ be the fixed vector of interest. The deep generative model $G : \mathbb{R}^k \to \mathbb{R}^n$ ($k \ll n$) maps from a low dimensional latent space to a higher dimensional space. In this paper, G is a feedforward neural network with ReLU activations and d layers.

Our definition of heavy-tailed samples assumes that the measurement matrix A only has bounded fourth moment. Our corruption model is Huber's ϵ-contamination model [44]. This model allows corruption in the measurement matrix A and measurements y. Precisely, these are:

Definition 1 (Heavy-tailed samples)**.** *We say that a random vector a is heavy-tailed if for a universal constant $C > 0$, the 4^{th} moment of a satisfies*

$$\left(\mathbb{E}\left[\langle a, u\rangle^4\right]\right)^{\frac{1}{4}} \leq C\left(\mathbb{E}\left[\langle a, u\rangle^2\right]\right)^{\frac{1}{2}}, \qquad \forall u \in \mathbb{R}^n.$$

For all $\delta > 0$, the $(4+\delta)^{th}$ moment of a need not exist, and we make no assumptions on them.

Definition 2 (ϵ-corrupted samples)**.** *We say that a collection of samples $\{y_i, a_i\}$ is ϵ-corrupted if they are i.i.d. observations drawn from the mixture*

$$\{y_i, a_i\} \sim (1-\epsilon)P + \epsilon Q,$$

where P is the uncorrupted distribution, Q is an arbitrary distribution.

Thus we assume that samples $\{y_i, a_i\}_{i=1}^m$ are generated from $(1-\epsilon)P + \epsilon Q$, where Q is an adversary, and P satisfies the following:

Assumption 1. *Samples $(y_i, a_i) \sim P$ satisfy $y_i = a_i^\top G(z^*) + \eta_i$, where the random vector a_i is isotropic and heavy-tailed as in Definition 2, and the noise term η_i is independent of a_i, i.i.d. with zero mean and bounded variance σ^2.*

4 Our Algorithm

$\|\cdot\|$ refers to ℓ_2 unless specified otherwise. The procedure proposed by [15] finds a reconstruction $\hat{x} = G(\hat{z})$, where $\hat{z}$ solves:

$$\widehat{z} := \arg\min_{z \in \mathbb{R}^k} \|AG(z) - y\|^2.$$

This is essentially an ERM-based approach. As is well known from the classical statistics literature, ERM's success relies on strong concentration properties, guaranteed, e.g., if the data are all sub-Gaussian. ERM may fail, however, in the presence of corruption or heavy-tails. Indeed, our experiments demonstrate that in the presence of outliers in y or A, or heavy-tailed noise in y, [15] fails to recover $G(z^*)$.

Remark *Unlike typical problems in M-estimation and high dimensional statistics, the optimization problem that defines the recovery procedure here is non-convex, and thus in the worst case, computationally intractable. Interestingly, despite non-convexity, as demonstrated in [15], (some appropriate version of) gradient descent is empirically very successful. In this paper, we take this as a computational primitive, thus sidestepping the challenge of proving whether a gradient-descent based method can efficiently provide guaranteed inversion of a generative model. Our theoretical guarantees are therefore statistical but our experiments show empirically excellent performance.*

4.1 MOM objective

It is well known that the median of means estimator achieves nearly sub-Gaussian concentration for one dimensional mean estimation of variance bounded random variables [73, 48, 3]. Inspired by the median-of-means algorithm, we propose the following algorithm to handle heavy-tails and outliers in y and A. We partition the set $[m]$ into M disjoint batches $\{B_1, B_2, \cdots, B_M\}$ such that each batch has cardinality $b = \frac{m}{M}$. Without loss of generality, we assume that M exactly divides m, so that b is an integer. For the j^{th} batch B_j, define the function

$$\ell_j(z) := \frac{1}{b}\|A_{B_j} G(z) - y_{B_j}\|^2, \tag{1}$$

where $A_{B_j} \in \mathbb{R}^{b \times n}$ denotes the submatrix of A corresponding to the rows in batch B_j. Similarly, $y_{B_j} \in \mathbb{R}^b$ denotes the entries of y corresponding to the batch B_j. Our workhorse is a novel variant of median-of-means (MOM) tournament procedure [65, 54] using the loss function eq. (1):

$$\widehat{z} = \arg\min_{z \in \mathbb{R}^k} \max_{z' \in \mathbb{R}^k} \underset{1 \le j \le M}{\text{median}} (\ell_j(z) - \ell_j(z')). \tag{2}$$

We do not assume that the minimizer is unique, since we only require a reconstruction $G(\widehat{z})$ which is close to $G(z^*)$. Any value of z in the set of minimizers will suffice. The intuition behind this aggregation of batches is that if the inner player z' chooses a point close to z^*, then the outer player z must also choose a point close to z^* in order to minimize the objective. Once this happens, there is no better option for z'. Hence a neighborhood around z^* is almost an equilibrium, and in fact there can be no neighborhood far from z^* with such an equilibrium.

Computational considerations. The objective function eq. (2) is not convex and we use Algorithm 1 as a heuristic to solve eq. (2). In Section 6, we empirically observe that gradient-based methods are able to minimize this objective and have good convergence properties. Our main theorem guarantees that a small value of the objective implies a good reconstruction and hence we can certify reconstruction quality using the obtained final value of the objective.

5 Theoretical results

We begin with a brief review of the Restricted Eigenvalue Condition in standard compressed sensing and show that S-REC is satisfied by heavy-tailed distributions.

Algorithm 1 Robust compressed sensing of generative models

Input: Data samples $\{y_j, a_j\}_{j=1}^m$.
Output: $G(\widehat{z})$.
Parameters: Number of batches M.

Initialize z and z'.
for $t = 0$ to $T-1$, **do**
  For each batch $j \in [M]$, calculate $\frac{1}{|B_j|}(\ell_j(z) - \ell_j(z'))$ by eq. (1).
  Pick the batch with the median loss $\underset{1\le j\le M}{\text{median}}(\ell_j(z) - \ell_j(z'))$, and evaluate the gradient for z and z' using backpropagation on that batch.
  (i) perform gradient descent for z;
  (ii) perform gradient ascent for z'.
end for
Output the $G(\widehat{z}) = G(z)$.

5.1 Set-Restricted Eigenvalue Condition for heavy-tailed distributions

Most theoretical guarantees for compressed sensing rely on variants of the Restricted Eigenvalue Condition(REC) [14, 17] and the closest to our setting is the Set Restricted Eigenvalue Condition [15](S-REC). Formally, $A \in \mathbb{R}^{m\times n}$ satisfies S-REC(S, γ, δ) on a set $S \subseteq \mathbb{R}^n$ if for all $x_1, x_2 \in S$,

$$\|Ax_1 - Ax_2\| \ge \gamma\|x_1 - x_2\| - \delta.$$

While we can prove many powerful results using the REC condition, proving that a matrix satisfies REC typically involves sub-Gaussian entries in A. If we don't have sub-Gaussianity, proving REC requires a finer analysis. A recent technique called the *small-ball method* [67] requires significantly weaker assumptions on A, and can be used to show REC [67, 85] for A satisfying Assumption 1. While this technique can be used for sparse vectors, we do not have a general understanding of what structures it can handle, since existing proofs make heavy use of sparsity.

We now show that a random matrix whose rows satisfy Assumption 1 will satisfy S-REC over the range of a generator $G : \mathbb{R}^k \to \mathbb{R}^n$ with high probability. This generalizes Lemma 4.2 in [15]– the original lemma required i.i.d. sub-Gaussian entries in the matrix A, whereas the following lemma only needs the rows to have bounded fourth moments.

Lemma 5.1. *Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a $d-$layered neural network with ReLU activations. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d. rows satisfying Definition 1. For any $\gamma < 1$, if $m = \Omega\left(\frac{1}{1-\gamma^2} kd \log n\right)$, then with probability $1 - e^{-\Omega(m)}$, for all $z_1, z_2 \in \mathbb{R}^k$, we have*

$$\frac{1}{m}\|AG(z_1) - AG(z_2)\|^2 \ge \gamma^2\|G(z_1) - G(z_2)\|^2.$$

This implies that the ERM approach of [15] still works when we only have a heavy-tailed measurement matrix A. However, as we show in our experiments, heavy-tailed noise in y and outliers in y, A will make ERM fail catastrophically. In order to solve this problem, we leverage the median-of-means tournament defined in eq. (2), and we will now show it is robust to heavy-tails and outliers in y, A.

5.2 Main results

We now present our main result. Theorem 5.5 provides recovery guarantees in terms of the error in reconstruction in the presence of heavy-tails and outliers, where $\widehat{z}$ is the (approximate) minimizer of eq. (2). First we show that the minimum value of the objective in eq. (2) is indeed small if there are no outliers.

Lemma 5.2. *Let M denote the number of batches. Assume that the measurements y and measurement matrix A are drawn from the uncorrupted distribution satisfying Assumption 1. Then with probability $1 - e^{-\Omega(M)}$, the objective in Equation (2) satisfies*

$$\min_{z\in\mathbb{R}^k} \max_{z'\in\mathbb{R}^k} \underset{1\le j\le M}{\text{median}}(\ell_{B_j}(z) - \ell_{B_j}(z')) \le 4\sigma^2. \tag{3}$$

We now introduce Lemma 5.3 and Lemma 5.4, which control two stochastic processes that appear in eq. (2). We show that minimizing the objective in eq. (2) implies that you are close to the unknown

vector $G(z^*)$. Notice that since z^* is one feasible solution of the inner maximization step of z', we can consider $z' = z^*$. Now consider the difference of square losses in eq. (2), which is given by:

$$\begin{aligned}\ell_j(\widehat{z}) - \ell_j(z^*) &= \frac{1}{b}\|A_{B_j}G(\widehat{z}) - y_{B_j}\|^2 - \frac{1}{b}\|A_{B_j}G(z^*) - y_{B_j}\|^2, \\ &= \frac{1}{b}\|A_{B_j}(G(\widehat{z}) - G(z^*))\|^2 - \frac{2}{b}\eta_{B_j}^\top(A_{B_j}(G(\widehat{z}) - G(z^*))),\end{aligned}$$

where the last line follows from an elementary arithmetic manipulation.

Assume we have the following bounds on a majority of batches:

$$\frac{1}{b}\|A_{B_j}(G(\widehat{z}) - G(z^*))\|^2 \gtrsim \|G(\widehat{z}) - G(z^*)\|^2, \tag{4}$$

$$-\frac{2}{b}\eta_{B_j}^\top(A_{B_j}(G(\widehat{z}) - G(z^*))) \gtrsim -\|G(\widehat{z}) - G(z^*)\|. \tag{5}$$

Since the objective is the median of the sum of the above terms, a small value of the objective implies that $\|G(\widehat{z}) - G(z^*)\|$ is small. We formally show these bounds in Lemma 5.3, Lemma 5.4.

Lemma 5.3. *Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a generative model from a d-layer neural network using ReLU activations. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d. uncorrupted rows satisfying Definition 1. Let the batch size $b = \Theta\left(C^4\right)$, let the number of batches satisfy $M = \Omega(kd\log n)$, and let γ be a constant which depends on the moment constant C. Then with probability at least $1 - e^{-\Omega(m)}$, for all $z_1, z_2 \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that*

$$\frac{1}{b}\|A_{B_j}(G(z_1) - G(z_2))\|^2 \geq \gamma^2\|G(z_1) - G(z_2)\|^2 \ , \forall j \in J.$$

Lemma 5.4. *Consider the setting of Lemma 5.3 with measurements satisfying $y = AG(z^*) + \eta$, where y, A, η satisfy Assumption 1 with noise variance σ^2. For a constant batch size b and number of batches $M = \Omega(kd\log n)$, with probability at least $1 - e^{-\Omega(m)}$, for all $z \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that*

$$\frac{1}{b}|\eta_{B_j}^T A_{B_j}(G(z) - G(z^*))| \leq \sigma\|G(z) - G(z^*)\| \ , \forall j \in J.$$

The above lemmas do not account for the $\epsilon-$corrupted samples in Definition 2. However, since the batch size is constant in both the lemmas, there exists a value of ϵ such that sufficiently many batches have no corruptions. Hence we can apply Lemma 5.3, Lemma 5.4 to these uncorrupted batches. Using these lemmas with a constant batch size b, we obtain Theorem 5.5. We defer its proof to Appendix E.

Theorem 5.5. *Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a generative model from a d-layer neural network using ReLU activations. There exists a (sufficiently small) constant fraction ϵ which depends on the moment constant C in Definition 1 such that the following is true. We observe $m = O(kd\log n)$ ϵ-corrupted samples from Definition 2, under Assumption 1. For any $z^* \in \mathbb{R}^k$, let $\widehat{z}$ minimize the objective function given by eq. (2) to within additive τ of the optimum. Then there exists a (sufficiently large) constant c, such that with probability at least $1 - e^{-\Omega(m)}$, the reconstruction $G(\hat{z})$ satisfies*

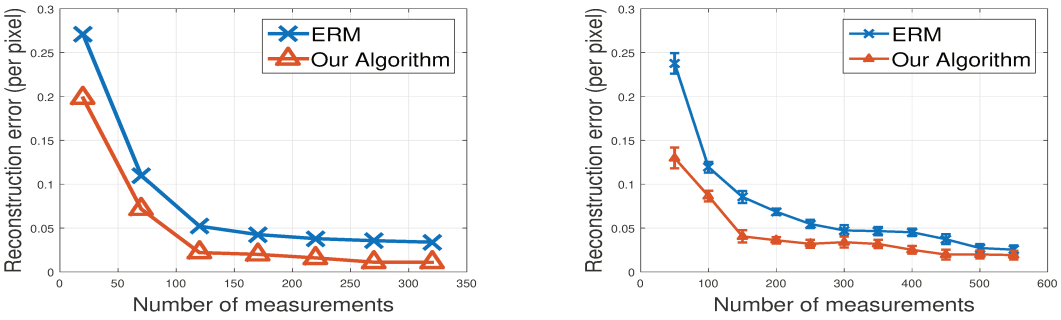
$$\|G(\widehat{z}) - G(z^*)\|^2 \leq c(\sigma^2 + \tau),$$

where σ^2 is the variance of noise under Assumption 1.

We briefly discuss the implications of Theorem 5.5, with regards to sample complexity and error in reconstruction.

Sample Complexity. Our sample complexity matches that of [15] up to constant factors. This shows that the minimizer of eq. (2) in the presence of heavy-tails and outliers provides the same guarantees as in the case of ERM with sub-Gaussian measurements.

Statistical accuracy and robustness. Let us analyze the error terms in our theorem. The term τ is a consequence of the minimization algorithm not being perfect, since it only reaches within τ of the true minimum. Hence it cannot be avoided. The term σ^2 is due to the noise in measurements. In the



(a) Results on MNIST.

(b) Results on CelebA-HQ

Figure 1: We compare Algorithm 1 with the baseline ERM [15] under the heavy-tailed setting *without* arbitrary outliers. We fix $k = 100$ for the MNIST dataset and $k = 512$ for the CelebA-HQ dataset. We vary the number of measurements, and plot the reconstruction error per pixel averaged over multiple trials. With increasing number of measurements, we observe the reconstruction error decreases. For heavy-tailed y and A *without* arbitrary outliers, our method obtains significantly smaller reconstruction error in comparison to ERM.

main result of [15], the reconstruction $G(\widehat{z})$ has error bounded by $\|G(\widehat{z}) - G(z^*)\|^2 \lesssim \|\eta\|^2/m + \tau$.[2] This gives the following conditions:

- If η is sub-Gaussian with variance σ^2, then $\|\eta\|^2/m \approx \sigma^2$ with high probability. Hence our bounds match up to constants.
- If higher order moments of η do not exist, an application of Chebyshev's inequality says that with probability $1 - \delta$, [15] has $\|G(z^*) - G(\widehat{z})\|^2 \approx \sigma^2/(m\delta)$, and this can be extremely large if we want $\delta = e^{-\Omega(m)}$.

Hence our method is clearly superior if η only has bounded variance, and if η is sub-Gaussian, then our bounds match up to constants. In the presence of corruptions, [15] has no provable guarantee.

6 Experiments

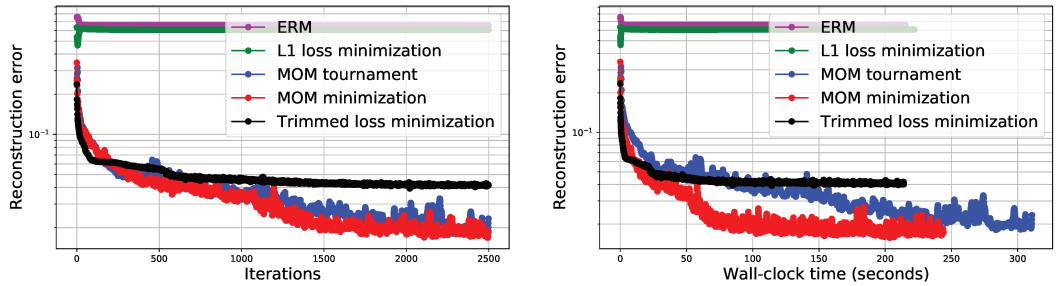


Figure 2: Plot of the reconstruction error versus the iteration number (left) and plot of the reconstruction error versus wall-clock time (right). ERM [15] and ℓ_1 minimization fail to converge. Our two proposed methods, MOM tournament(blue) and MOM minimization(red), have the smallest reconstruction error. We provide a theoretical analysis for the MOM tournament algorithm, and observe that direct minimization of the MOM objective also works in practice. The computation time of our algorithms is nearly the same as the baselines.

In this section, we study the empirical performance of our algorithm on generative models trained on real image datasets. We show that we can reconstruct images under heavy-tailed samples and arbitrary outliers. For additional experiments and experimental setup details, see Appendix F.

Heavy-Tailed Samples In this experiment, we deal with the *uncorrupted* compressed sensing model P, which has heavy-tailed measurement matrix and stochastic noise: $y = AG(z^*) + \eta$. We use

[2] In [15], the bound is stated as $\|\eta\|^2$, but our A has a different scaling, and hence the correct bound in our setting is $\|\eta\|^2/m$.

a Student's $t-$distribution (a typical example of heavy-tails) for A and η. We compare Algorithm 1 with the baseline ERM [15] for heavy-tailed data *without* arbitrary corruptions on MNIST [55] and CelebA-HQ [51, 63]. We trained a DCGAN [80] with $k = 100$ and $d = 5$ layers to produce 64×64 MNIST images. For CelebA-HQ, we used a PG-GAN [51] with $k = 512$ to produce images of size $256 \times 256 \times 3 = 196,608$.

We vary the number of measurements m and obtain the reconstruction error $\|G(\widehat{z}) - G(z^*)\|^2/n$ for Algorithm 1 and ERM, where $G(z^*)$ is the ground truth image. In Figure 1, Algorithm 1 and ERM both have decreasing reconstruction error per pixel with increasing number of measurements. To conclude, even for heavy-tailed noise *without* arbitrary outliers, Algorithm 1 obtains significantly smaller reconstruction error when compared to ERM.

Arbitrary corruptions. In this experiment, we use the same heavy-tailed samples as above, and we add $\epsilon = 0.02$-fraction of arbitrary corruption. We set the outliers of measurement matrix A as random sign matrix, and the outliers of y are fixed to be -1. We note that we don't use any targeted attack to simulate the outliers. We perform our experiments on the CelebA-HQ dataset using a PG-GAN of latent dimension $k = 512$, and fix the number of measurements to $m = 1000$.

We compare our algorithm to a number of natural baselines. Our first baseline is ERM [15] which is not designed to deal with outliers. While its fragility is interesting to note, in this sense it is not unexpected. For outliers in y, classical robust methods replace the loss function by an ℓ_1 loss function or Huber loss function. This is done in order to avoid the squared loss, which makes recovery algorithms very sensitive to outliers. In this case, we have $\widehat{z} := \arg\min \|y - AG(z)\|_1$.

We also investigate the performance of trimmed loss minimization, which is a recent algorithm proposed by [81]. This algorithm picks the $t-$fraction of samples with smallest empirical loss for each update step, where t is a hyper-parameter.

We run Algorithm 1 and its variant MOM minimization. The MOM minimization directly minimizes

$$\widehat{z} = \arg\min_{z \in \mathbb{R}^k} \text{median}_{1 \le j \le M}(\ell_j(z)), \tag{6}$$

and we use gradient-based methods similar to Algorithm 1 to solve it. Since Algorithm 1 optimizes z and z' in one iteration, the actual computation time of MOM tournament is twice that of MOM minimization. As shown in Figure 2, Figure 3, ERM [15] and ℓ_1 loss minimization fail to converge to the ground truth and in particular, they may recover a completely different person. Trimmed loss minimization [81] only succeeds on occasion, and when it fails, it obtains a visibly different person. The convergence of the MOM minimization per iteration is very similar to the MOM tournament, and they both achieve much smaller reconstruction error compared to trimmed loss minimization. The right panel of Figure 2 plots the reconstruction error versus the actual computation time, showing our algorithms match baselines. We plot the MSE vs. number of measurements in Figure 4b, where the fraction of corruptions is set to $\epsilon = 0.02$.

Miscellaneous Experiments *Is ERM ever better than MOM?* So far we have analyzed cases where MOM performs better than ERM. Since ERM is known to be optimal in linear regression when dealing with uncorrupted sub-Gaussian data, we expect it to be superior to MOM when our measurements are all sub-Gaussian. We evaluate this in Fig. 4a and observe that ERM obtains smaller MSE in this setting. Notice that as we reduce the number of batches in MOM, it approaches ERM.

How sensitive is MOM to the number of batches? In Figure 4c we study the MSE of MOM tournaments and MOM minimization as we vary the number of batches.

In order to select the optimal number of batches (M), we keep a set of validation measurements that we do not use in the optimization routines for estimating x. We can run MOM for different value of M to get multiple reconstructions, and then evaluate each reconstruction using the validation measurements to pick the best reconstruction. Note that one should use the median-of-means loss while evaluating the validation error as well.

7 Conclusion

The phenomenon observed in Figure 3 highlights the importance of our method. Our work raises several questions about why the objective we consider can be minimized, and suggests we need a new

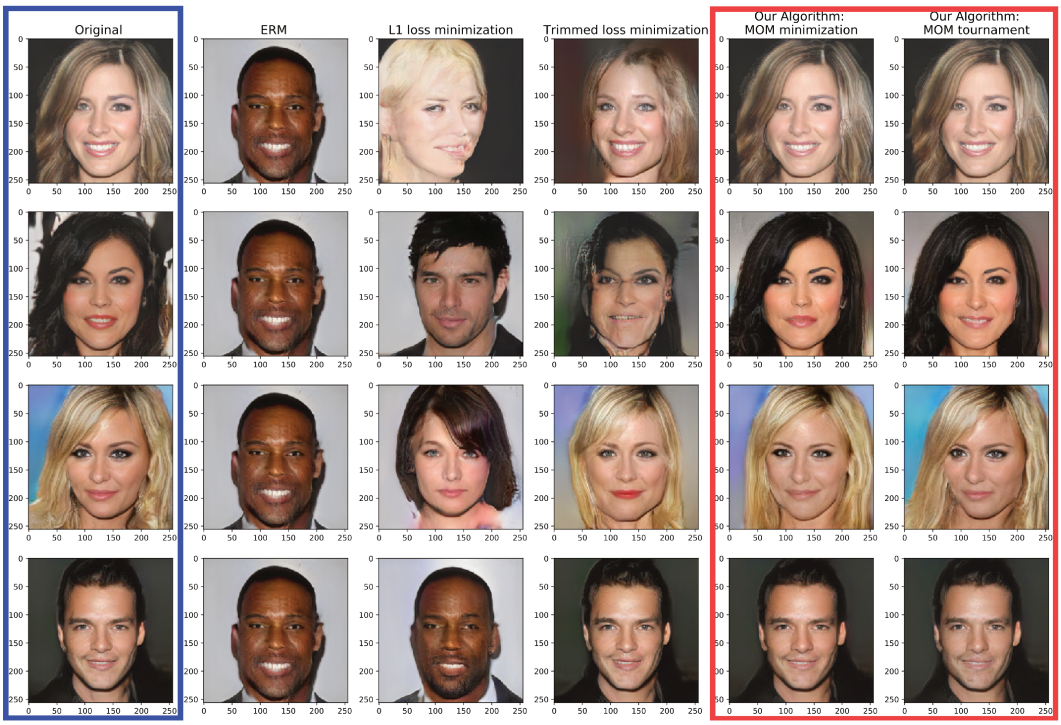
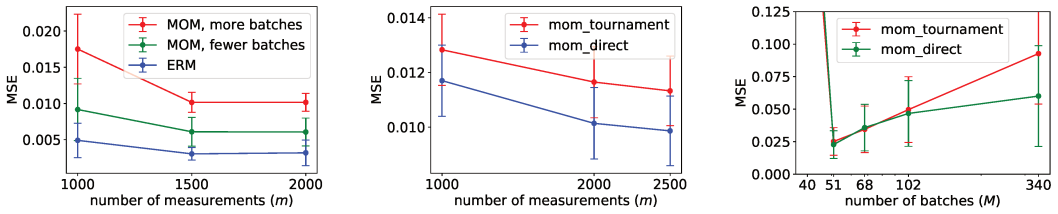


Figure 3: Reconstruction results on CelebA-HQ for $m = 1000$ measurements with 20 corrupted measurements. For each row, the first column is ground truth from a generative model. Subsequent columns show reconstructions by ERM [15], ℓ_1-minimization, trimmed loss minimization [81]. In particular, vanilla ERM, ℓ_1-minimization obtain completely different faces. Since we use the same outlier for different rows, vanilla ERM produces the same reconstruction irrespective of the ground truth. Trimmed loss minimization only succeeds on occasion (the last row), and when it fails, it obtains a similar but still different face. The last two columns show reconstructions by our proposed algorithms. The second to last one is directly minimizing the MOM objective eq. (6), and the last column minimizes the MOM tournament objective eq. (2). We provide a theoretical analysis for the MOM tournaments algorithm, and observe that direct minimization of the MOM objective also works in practice. We observe the last two columns have much better reconstruction performance – we get a high quality reconstruction under heavy-tailed measurements and arbitrary outliers.

paradigm for analysis that accounts for similar instances that enjoy empirical success, even though they can be provably hard in the worst case.



(a) ERM vs MOM (b) MOM on arbitrary corruptions (c) MSE vs. Number of batches

Figure 4: (a) We compare ERM and MOM by plotting MSE vs number of measurements when the measurements are *sub-Gaussian without corruptions*. (b) Aggregate statistics for MOM in the presence of corruptions. (c) MSE vs number of batches for MOM on 1000 heavy-tailed measurements and 20 corruptions. All error bars indicate 95% confidence intervals. Plots use a PGGAN on CelebA-HQ.

8 Acknowledgments

Ajil Jalal and Alex Dimakis have been supported by NSF Grants CCF 1763702,1934932, AF 1901292, 2008710, 2019844 research gifts by NVIDIA, Western Digital, WNCG IAP, computing resources from TACC and the Archie Straiton Fellowship. Constantine Caramanis and Liu Liu have been supported by NSF Award 1704778 and a grant from the Army Futures Command.

9 Broader Impact

Sparsity has played an important role across many areas of statistics, engineering and computer science, as a regularizing prior that captures important structure in many applications. Recent work has illustrated that given enough data, deep generative models are poised to play a revolutionary role, as a modern, data-driven replacement for sparsity. Much work remains to bring this agenda to fruition, but we believe that, as a variety of recent works have suggested, this direction can revolutionize imaging in a number of different important domains, not least of all, medical imaging.

This work addresses the robustness, and hence the trustworthiness and reliability of GAN-inversion-based techniques. As mentioned, this is especially critical, since high quality GANs will always produce perceptually high quality images, hence recovery failures may not be readily detectable by inspection.

Still, many significant issues remain that this work does not address. This includes understanding when and how sufficiently powerful and expressive GANs can be trained, since the scope of high quality GANs still appears to be limited. Another important consideration includes the core computational issue: the GAN inversion problem, which this work also faces, is intractable in the worst case, yet in practice appears to not pose a significant challenge. Understanding this dichotomy is very important.

References

[1] Anders Aamand, Piotr Indyk, and Ali Vakilian. (learned) frequency estimation algorithms under zipfian distribution. *arXiv preprint arXiv:1908.05198*, 2019.

[2] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 10 2012.

[3] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

[4] Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, and Timo Bremer. Mimicgan: Robust projection onto image manifolds with corruption mimicking. *arXiv preprint arXiv:1912.07748*, 2019.

[5] Muhammad Asim, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *arXiv preprint arXiv:1905.11672*, 2019.

[6] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Blind image deconvolution using deep generative priors. *arXiv preprint arXiv:1802.04073*, 2018.

[7] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Solving bilinear inverse problems using deep generative priors. *CoRR, abs/1802.04073*, 3(4):8, 2018.

[8] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. *arXiv preprint arXiv:1912.02008*, 2019.

[9] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[10] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

[11] Richard G Baraniuk. Compressive sensing [lecture notes]. *IEEE signal processing magazine*, 24(4):118–121, 2007.

[12] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[13] Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[14] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[15] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[16] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. *ICLR*, 2:5, 2018.

[17] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[18] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[19] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[20] Chen Chen and Junzhou Huang. Compressive sensing mri with wavelet tree sparsity. In *Advances in neural information processing systems*, pages 1115–1123, 2012.

[21] Minhua Chen, Jorge Silva, John Paisley, Chunping Wang, David Dunson, and Lawrence Carin. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 58(12):6140–6155, 2010.

[22] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.

[23] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

[24] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

[25] Manik Dhar, Aditya Grover, and Stefano Ermon. Modeling sparse deviations for compressed sensing using generative models. *arXiv preprint arXiv:1807.01442*, 2018.

[26] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[27] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[28] Yonina C Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[29] Alyson K Fletcher, Parthe Pandit, Sundeep Rangan, Subrata Sarkar, and Philip Schniter. Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis. In *Advances in Neural Information Processing Systems*, pages 7440–7449, 2018.

[30] Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1884–1888. IEEE, 2018.

[31] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

[32] Fabian Latorre Gómez, Armin Eftekhari, and Volkan Cevher. Fast and provable admm for learning with generative priors. *arXiv preprint arXiv:1907.03343*, 2019.

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[34] Paul Hand and Babhru Joshi. Global guarantees for blind demodulation with generative priors. In *Advances in Neural Information Processing Systems*, pages 11531–11541, 2019.

[35] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pages 9136–9146, 2018.

[36] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *arXiv preprint arXiv:1705.07576*, 2017.

[37] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[38] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. *arXiv preprint arXiv:1810.03982*, 2018.

[39] Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. *arXiv preprint arXiv:2005.03991*, 2020.

[40] Chinmay Hegde. Algorithmic aspects of inverse problems using generative models. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 166–172. IEEE, 2018.

[41] Chinmay Hegde, Michael Wakin, and Richard G Baraniuk. Random projections for manifold learning. In *Advances in neural information processing systems*, pages 641–648, 2008.

[42] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. 2018.

[43] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[44] Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

[45] Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. In *Advances in Neural Information Processing Systems*, pages 7400–7410, 2019.

[46] Gauri Jagatap and Chinmay Hegde. Phase retrieval using untrained neural network priors. 2019.

[47] Shirin Jalali and Xin Yuan. Solving linear inverse problems using generative models. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 512–516. IEEE, 2019.

[48] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[49] Maya Kabkab, Pouya Samangouei, and Rama Chellappa. Task-aware compressed sensing with generative adversarial networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[50] Akshay Kamath, Sushrut Karmalkar, and Eric Price. Lower bounds for compressed sensing with generative models. *arXiv preprint arXiv:1912.02938*, 2019.

[51] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[52] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[53] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[54] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.

[55] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[56] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[57] Qi Lei, Ajil Jalal, Inderjit S Dhillon, and Alexandros G Dimakis. Inverting deep generative models, one layer at a time. In *Advances in Neural Information Processing Systems*, pages 13910–13919, 2019.

[58] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013.

[59] Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*, 2019.

[60] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

[61] Zhaoqiang Liu, Selwyn Gomes, Avtansh Tiwari, and Jonathan Scarlett. Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors. *arXiv preprint arXiv:2002.01697*, 2020.

[62] Zhaoqiang Liu and Jonathan Scarlett. Information-theoretic lower bounds for compressive sensing with generative models. *arXiv preprint arXiv:1908.10744*, 2019.

[63] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.

[64] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

[65] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.

[66] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[67] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[68] Shahar Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3-4):641–674, 2017.

[69] Chris Metzler, Ali Mousavi, and Richard Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems*, pages 1772–1783, 2017.

[70] Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[71] Lukas Mosser, Olivier Dubrule, and Martin J Blunt. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. *Mathematical Geosciences*, 52(1):53–79, 2020.

[72] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[73] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[74] Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via l1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

[75] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *arXiv preprint arXiv:2005.06001*, 2020.

[76] REAC Paley and Antoni Zygmund. A note on analytic functions in the unit circle. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 266–272. Cambridge University Press, 1932.

[77] Parthe Pandit, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Inference with deep generative priors in high dimensions. *arXiv preprint arXiv:1911.03409*, 2019.

[78] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[79] Shuang Qiu, Xiaohan Wei, and Zhuoran Yang. Robust one-bit recovery via relu generative networks: Improved statistical rates and global landscape analysis. *arXiv preprint arXiv:1908.05368*, 2019.

[80] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[81] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. *arXiv preprint arXiv:1810.11874*, 2018.

[82] Ganlin Song, Zhou Fan, and John Lafferty. Surfing: Iterative optimization over incrementally trained deep networks. In *Advances in Neural Information Processing Systems*, pages 15008–15017, 2019.

[83] Michel Talagrand. A new isoperimetric inequality and the concentration of measure phenomenon. In *Geometric Aspects of Functional Analysis*, pages 94–124. Springer, 1991.

[84] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[85] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[86] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[87] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[88] Xiaohan Wei, Zhuoran Yang, and Zhaoran Wang. On the statistical rate of nonlinear recovery in generative models with heavy-tailed data. In *International Conference on Machine Learning*, pages 6697–6706, 2019.

[89] Weiyu Xu and Babak Hassibi. Compressed sensing over the grassmann manifold: A unified analytical framework. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 562–567. IEEE, 2008.

[90] Weiyu Xu, Enrique Mallada, and Ao Tang. Compressive sensing over graphs. In *2011 Proceedings IEEE INFOCOM*, pages 2087–2095. IEEE, 2011.

[91] Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.

A Proof of Lemma 5.1

Lemma (Lemma 5.1)**.** *Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a $d-$layered neural network with ReLU activations. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d rows satisfying Assumption 1. If $m = \Omega\left(\frac{1}{1-\gamma^2} kd \log n\right)$, then with probability $1 - e^{-\Omega(m)}$, A satisfies*

$$\frac{1}{m}\|AG(z_1) - AG(z_2)\|^2 \geq \gamma^2 \|G(z_1) - G(z_2)\|^2$$

for all $z_1, z_2 \in \mathbb{R}^k$.

Proof. The proof is based on Proposition A.1 and Proposition A.2, which will be introduced as follows. Proposition A.1 shows that the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the range of $e^{O(kd\log n)}$ different $2k-$dimensional subspaces.

Proposition A.2 guarantees the result for a single subspace with probability $1 - e^{-m}$. Since $m = \Omega(kd \log n)$, the proof follows from a union bound over the $e^{O(kd\log n)}$ subspaces in Proposition A.1. □

Proposition A.1. *If $G : \mathbb{R}^k \to \mathbb{R}^n$ is a $d-$layered neural network with ReLU activations, then the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the union of $O\left(n^{2kd}\right)$ different $2k-$dimensional subspaces.*

Proof of Proposition (A.1). From Lemma 8.3 in [15], the set $\{G(z) : z \in \mathbb{R}^k\}$ lies in the union of $O(n^{kd})$ different $k-$dimensional subspaces.

This implies that the set

$$\{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$$

lies in the union of $M = O(n^{2kd})$ different $2k-$dimensional subspaces.

□

Proposition A.2. *Consider a single $2k-$dimensional subspace given by $S_1 = \{Wz : W \in \mathbb{R}^{n\times 2k}, W^TW = I_{2k}, z \in \mathbb{R}^{2k}\}$. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d rows drawn from a distribution satisfying Assumption* (1)*. If $m = O(\frac{C^2k}{\frac{3}{4}-\gamma^2})$, with probability $1 - e^{-\Omega(m)}$, A satisfies*

$$\frac{1}{m}\|Av\|^2 \geq \gamma^2\|v\|^2, \forall v \in S_1.$$

Proof. The proof follows Theorem 14.12 in [87], with non-trivial modifications for our setting.

We want to show that for all vectors $v \in S_1$,

$$\frac{1}{m}||Av||^2 \geq \gamma^2||v||^2.$$

For $u, \tau \in \mathbb{R}$, define the truncated quadratic function

$$\phi_\tau(u) = \begin{cases} u^2 & \text{if } |u| \leq \tau, \\ \tau^2 & \text{otherwise.} \end{cases} \tag{7}$$

By construction, $\phi_\tau(\langle a_i, v\rangle) \leq \langle a_i, v\rangle^2$.

This implies that

$$\frac{1}{m}||Av||^2 = \frac{1}{m}\sum_{i=1}^m \langle a_i, v\rangle^2 = \frac{\|v\|^2}{m}\sum_{i=1}^m \langle a_i, \tfrac{v}{\|v\|}\rangle^2 \tag{8}$$

$$\geq \frac{\|v\|^2}{m}\sum_{i=1}^m \phi_\tau(\langle a_i, \tfrac{v}{\|v\|}\rangle) \tag{9}$$

$$\geq \|v\|^2 \mathbb{E}\left[\frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|}\rangle)}{m}\right] - \|v\|^2 \left|\frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|}\rangle)}{m} - \mathbb{E}\left[\frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|}\rangle)}{m}\right]\right| \tag{10}$$

$$= \|v\|^2 \mathbb{E}\left[\phi_\tau(\langle a, \frac{v}{\|v\|}\rangle)\right] - \|v\|^2 \left|\frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|}\rangle)}{m} - \mathbb{E}\left[\phi_\tau(\langle a, \frac{v}{\|v\|}\rangle)\right]\right| \tag{11}$$

$$\geq \|v\|^2 \mathbb{E}\left[\phi_\tau(\langle a, \frac{v}{\|v\|}\rangle)\right] - \|v\|^2 \sup_{v\in S_1}\left|\frac{1}{m}\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|}\rangle) - \mathbb{E}\left[\phi_\tau(\langle a, \frac{v}{\|v\|}\rangle)\right]\right| \tag{12}$$

In Claim A.3 we will show that for $\tau^2 = \frac{C^4}{\frac{3}{4}-\gamma^2}$, we have

$$\mathbb{E}\left[\phi_\tau(\langle a, \frac{v}{\|v\|}\rangle)\right] \geq (\gamma^2 + \frac{1}{4}).$$

In Claim A.4 we will show that with overwhelming probability in m,

$$\sup_{v:\|v\|\leq 1}\left|\frac{1}{m}\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|}\rangle) - \mathbb{E}\left[\phi_\tau(\langle a, \frac{v}{\|v\|}\rangle)\right]\right| \leq \frac{1}{4}.$$

These two results together imply that

$$\frac{1}{m}\|Av\|^2 \geq \gamma^2\|v\|^2.$$

with overwhelming probability in m. □

Claim A.3. *Assume that the random vector a satisfies Assumption (1) with constant C. Let ϕ_τ be the thresholded quadratic function defined in Eqn (7). For all $v \in \mathbb{R}^n$, $\|v\| \leq 1$, we have*

$$\mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right] \geq \left(1 - \frac{C^4}{\tau^2}\right)\|v\|^2.$$

Proof.

$$\|v\|^2 - \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right] = \mathbb{E}\left[\langle a, v\rangle^2\right] - \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right] \tag{13}$$

$$= \mathbb{E}\left[(\langle a, v\rangle^2 - \tau^2)1_{\{|\langle a,v\rangle|\geq\tau\}}\right] \tag{14}$$

$$\leq \mathbb{E}\left[\langle a, v\rangle^2 1_{\{|\langle a,v\rangle|\geq\tau\}}\right] \tag{15}$$

By the Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\langle a, v\rangle^2 1_{\{|\langle a,v\rangle|\geq\tau\}}\right] \leq \left(\mathbb{E}\left[\langle a, v\rangle^4\right]\right)^{\frac{1}{2}} \left(\Pr\left[|\langle a, v\rangle| \geq \tau\right]\right)^{\frac{1}{2}} \tag{16}$$

From Assumption (1), we have

$$\left(\mathbb{E}\left[\langle a, v\rangle^4\right]\right)^{\frac{1}{2}} \leq C^2\mathbb{E}\left[\langle a, v\rangle^2\right].$$

From Chebyshev's inequality and Assumption (1), we have

$$\left(\Pr\left[|\langle a, v\rangle| \geq \tau\right]\right)^{\frac{1}{2}} \leq \left(\frac{\mathbb{E}\left[|\langle a, v\rangle|^4\right]}{\tau^4}\right)^{\frac{1}{2}} \leq \left(\frac{C^4\mathbb{E}\left[|\langle a, v\rangle|^2\right]^2}{\tau^4}\right)^{\frac{1}{2}} = \frac{C^2\mathbb{E}\left[|\langle a, v\rangle|^2\right]}{\tau^2}. \tag{17}$$

Substituting the above two inequalities into eq. (16), we get

$$\mathbb{E}\left[\langle a, v\rangle^2 1_{\{|\langle a,v\rangle|\geq\tau\}}\right] \leq \frac{C^4\mathbb{E}\left[\langle a, v\rangle^2\right]^2}{\tau^2} \tag{18}$$

$$= \frac{C^4\|v\|^4}{\tau^2} \leq \frac{C^4\|v\|^2}{\tau^2}. \tag{19}$$

Substituting into Eqn (13),

$$\|v\|^2 - \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right] \leq \frac{C^4\|v\|^2}{\tau^2}, \tag{20}$$

which completes the proof. □

Claim A.4. *For an orthonormal matrix $U \in \mathbb{R}^{n\times 2k}$, let $S := \{v : v = Uz, \|v\| = 1\}$. Let ϕ_τ be the function defined in Proposition A.2. For $m = \Omega\left(\tau^2 k\right)$, we have*

$$\sup_{v\in S}\left|\frac{1}{m}\sum_{i=1}^{m}\phi_\tau(\langle a_i, v\rangle) - \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right]\right| \leq \frac{1}{4}.$$

with probability $1 - e^{-\Omega(m)}$.

Proof. Define

$$Z_m = \sup_{v\in S}\left|\frac{1}{m}\sum_{i=1}^{m}\phi_\tau(\langle a_i, v\rangle) - \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right]\right|.$$

We will first show that

$$\mathbb{E}_A\left[Z_m\right] \leq \frac{1}{8}$$

for large enough m. Then we use Talagrand's inequality [83] to show that

$$\Pr\left[Z_m \geq \mathbb{E}\left[Z_m\right] + \frac{1}{8}\right] \leq e^{-\Omega(m)},$$

using which we can conclude that $Z_m \leq \frac{1}{4}$ with probability $1 - e^{-\Omega(m)}$.

By the symmetrization inequality, we have

$$\mathbb{E}_A\left[Z_m\right] \leq 2\mathbb{E}_{\epsilon,A}\left[\sup_{v\in S}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\phi_\tau(\langle a_i, v\rangle)\right|\right]$$

where $\{\epsilon_i\}_{i=1}^m$ are i.i.d Bernoulli ± 1 random variables.

Since ϕ_τ is a Lipschitz function with Lipschitz constant 2τ, we can apply the Ledoux-Talagrand contraction inequality [56] (refer to Appendix G for the sake of completeness) to get

$$\begin{aligned}
&2\mathbb{E}_{\epsilon,A}\left[\sup_{v\in S}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\phi_\tau(\langle a_i, v\rangle)\right|\right] \\
\leq &8\tau\mathbb{E}_{\epsilon,A}\left[\sup_{v\in S}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\langle a_i, v\rangle\right|\right] &(21)\\
= &8\tau\mathbb{E}_{\epsilon,A}\left[\sup_{v\in S}\left|\frac{1}{m}\epsilon^T Av\right|\right]. &(22)
\end{aligned}$$

Since $S := \{v : v = Uz, \|v\| = 1\}$, we have

$$\begin{aligned}
&8\tau\mathbb{E}_{\epsilon,A}\left[\sup_{v\in S}\left|\frac{1}{m}\epsilon^T Av\right|\right] &(23)\\
= &8\tau\mathbb{E}_{\epsilon,A}\left[\sup_{z:\|z\|=1}\left|\frac{8\tau}{m}\epsilon^T AUz\right|\right] &(24)\\
\leq &\frac{8\tau}{m}\mathbb{E}_{\epsilon,A}\left[\|\epsilon^T AU\|_2\right] &(25)\\
\leq &\frac{8\tau}{m}\sqrt{\mathbb{E}_{\epsilon,A}\left[\|\epsilon^T AU\|_2^2\right]} &(26)
\end{aligned}$$

The third line follows from the Cauchy-Schwartz inequality, and the fourth line follows from Jensen's inequality.

Notice that

$$\mathbb{E}_\epsilon \left[\|\epsilon^T AU\|_2^2\right] = \text{trace}(AUU^T A^T) = \text{trace}(U^T A^T AU)$$

Since $U^T U = I_{2k}$, we have

$$\mathbb{E}_{\epsilon,A} \left[\|\epsilon^T AU\|_2^2\right] = \mathbb{E}_A \left[\text{trace}(U^T A^T AU)\right] \tag{27}$$

$$= \sum_{i=1}^{m} \mathbb{E}_{a_i} \text{trace}(U^T a_i a_i^T U) \tag{28}$$

$$= \sum_{i=1}^{m} \text{trace}(U^T I_n U) = m \,\text{trace}(I_{2k}) = 2km. \tag{29}$$

Putting this together, and choosing $m = \Omega(\tau^2 k)$, we have

$$\mathbb{E}_A \left[Z_m\right] \leq 8\tau\sqrt{\frac{2k}{m}} \leq \frac{1}{8}.$$

We now need to show that

$$\Pr\left[Z_m \geq \mathbb{E}\left[Z_m\right] + \frac{1}{8}\right] \leq e^{-\Omega(m)}.$$

By construction, $\phi_\tau(\langle a_i, v\rangle) \leq \tau^2$ for all $v \in S$.

In order to apply Talagrand's inequality, we need to bound

$$\sigma^2 = \sup_{v \in S} \mathbb{E}\left[\left(\phi_\tau(\langle a, v\rangle) - \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right]\right)^2\right].$$

We can bound this by

$$\text{var}(\phi_\tau(\langle a, v\rangle)) \leq \mathbb{E}\left[\phi_\tau^2(\langle a, v\rangle)\right] \tag{30}$$

$$\leq \tau^2 \mathbb{E}\left[\phi_\tau(\langle a, v\rangle)\right] \leq \tau^2 \tag{31}$$

Applying Talagrand's inequality, we have

$$\Pr\left[Z_m \geq \mathbb{E}\left[Z_m\right] + t\right] \leq C_1 \exp\left(-\frac{C_2 m t^2}{\tau^2 + \tau^2 t}\right).$$

Setting $t = \frac{1}{8}, m = \Omega(\tau^2 k)$ we get

$$\Pr[Z_m \geq \frac{1}{4}] \leq \Pr\left[Z_m \geq \mathbb{E}\left[Z_m\right] + \frac{1}{8}\right] \leq C_1 e^{-\frac{C_2 m}{\tau^2}} = e^{-\Omega(m)}.$$

This concludes the proof. □

B Proof of Lemma 5.2

Lemma B.1. *Let M denote the number of batches. Then with probability $1 - e^{-\Omega(M)}$, the objective in Equation (2) satisfies*

$$\min_{z \in \mathbb{R}^k} \max_{z' \in R^k} \underset{1 \leq j \leq M}{\text{median}} \ell_{B_j}(z) - \ell_{B_j}(z') \leq 4\sigma^2. \tag{32}$$

Proof. By setting $z \leftarrow z^*$, for all $z' \in \mathbb{R}^k$, for any $j \in [M]$, we have

$$\ell_{B_j}(z^*) - \ell_{B_j}(z') \leq \ell_{B_j}(z^*) = \frac{1}{b}\|\eta_{B_j}\|^2. \tag{33}$$

Since the noise is i.i.d. and has variance σ^2, we have $\mathbb{E}\left[\ell_{B_j}(z^*)\right] = \mathbb{E}\frac{1}{b}\|\eta_{B_j}\|^2 = \sigma^2$.

For batch $j \in [M]$, define the indicator random variable

$$Y_j = \mathbf{1}\left\{\ell_{B_j}(z^*) \geq 4\sigma^2\right\}.$$

By Markov's inequality, since $\mathbb{E}[\ell_{B_j}(z^*)] = \sigma^2$, we have

$$\Pr\left[Y_j = 1\right] \leq \frac{1}{4} \Rightarrow \mathbb{E}\left[\sum_{j=1}^{M} Y_j\right] \leq \frac{M}{4}. \tag{34}$$

By the Chernoff bound,

$$\Pr\left[\sum_{j=1}^{M} Y_j \geq \frac{M}{2}\right] \leq \Pr\left[\sum_{j=1}^{M} Y_j \geq 2\mathbb{E}[\sum_{j=1}^{M} Y_j]\right] \leq e^{-\Omega(M)}. \tag{35}$$

The above inequality implies that with probability $1 - e^{-\Omega(M)}$, for all $z' \in \mathbb{R}^k$, at least $\frac{M}{2}$ batches satisfy

$$\ell_{B_j}(z^*) - \ell_{B_j}(z') \leq 4\sigma^2.$$

This gives

$$\min_{z\in\mathbb{R}^k} \max_{z'\in R^k} \underset{1\leq j\leq M}{\text{median}}(\ell_{B_j}(z) - \ell_{B_j}(z')) \leq 4\sigma^2. \tag{36}$$

□

C Proof of Lemma 5.3

Lemma (Lemma 5.3)**.** *Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a generative model from a d-layer neural network using ReLU activations. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d rows satisfying Assumption 1. Let the batch size $b = \Theta\left(C^4\right)$, let the number of batches satisfy $M = \Omega(kd\log n)$, and let γ be a constant which depends on the moment constant C. Then with probability at least $1 - e^{-\Omega(m)}$, for all $z_1, z_2 \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that*

$$\frac{1}{b}\|A_{B_j}(G(z_1) - G(z_2))\|^2 \geq \gamma^2\|G(z_1) - G(z_2)\|^2 \ , \forall j \in J.$$

Proof. Proposition A.1 shows that the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the range of $e^{O(kd\log n)}$ different $2k-$dimensional subspaces.

Proposition C.1 guarantees the result for a single subspace with probability $1 - e^{-\Omega(M)}$. Since $M = \Omega(kd\log n)$ and the batch size is constant which depends on the moment constant C, the lemma follows from a union bound over the $e^{O(kd\log n)}$ subspaces in Proposition A.1. □

Proposition C.1. *Consider a single $2k-$dimensional subspace given by $S = \{Wz : W \in \mathbb{R}^{n\times 2k}, W^TW = I_{2k}, z \in \mathbb{R}^{2k}\}$. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d rows drawn from a distribution satisfying Assumption (1) with constant C. If the batch size $b = O(C^4)$ and the number of batches satisfies $M = \Omega\left(k\log\frac{1}{\epsilon}\right)$, with probability $1 - e^{-\Omega(M)}$, for all $x \in S$, there exist a subset of batches $J_x \subseteq [M]$ with $|J_x| \geq 0.90M$ such that*

$$\frac{1}{b}\|A_{B_j}x\|^2 \geq \gamma^2\|x\|^2 \, \forall j \in J_x,$$

where $\gamma = \Theta\left(\frac{1}{C^2}\right)$ is a constant that depends on the moment constant C.

Proof. Since the bound we want to prove is homogeneous, it suffices to show it for all vectors in S that have unit norm. Let $W \in \mathbb{R}^{n\times 2k}$ be the orthonormal matrix spanning S, and S_1 denote the set of unit norm vectors in its span. That is,

$$S_1 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| = 1, W \in \mathbb{R}^{n\times 2k}, W^TW = I_{2k}\}.$$

For a fixed $x \in S_1$ and $0 < t < 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\langle a, x\rangle^2\right] =& \mathbb{E}\left[\langle a, x\rangle^2 \mathbf{1}\{\langle a, x\rangle \le t^2\|x\|^2\}\right] \mathbb{E}\left[\langle a, x\rangle^2 \mathbf{1}\{\langle a, x\rangle > t^2\|x\|^2\}\right] && (37)\\
&\le t^2\|x\|^2 + \mathbb{E}\left[\langle a, x\rangle^4\right]^{\frac{1}{2}} \left(\Pr\left[\langle a, x\rangle^2 \ge t^2\|x\|^2\right]\right)^{\frac{1}{2}} && (38)\\
&\le t^2\|x\|^2 + C^2\|x\|^2 \left(\Pr\left[\langle a, x\rangle^2 \ge t^2\|x\|^2\right]\right)^{\frac{1}{2}} && (39)\\
\Rightarrow \Pr\left[\langle a, x\rangle^2 \ge t^2\|x\|^2\right] \ge& \frac{\left(1-t^2\right)^2\|x\|^4}{C^4\|x\|^4} = \frac{\left(1-t^2\right)^2}{C^4} = C_1. && (40)
\end{aligned}$$

This is essentially a modified version of the Paley-Zigmund inequality [76].

Consider a batch B_j, which has b samples. By the concentration of Bernoulli random variables, with probability $1 - 2e^{-\Omega(C_1 b)}$, we have

$$\sum_{i\in B_j} \mathbf{1}\left\{\langle a_i, x\rangle^2 \ge t^2\|x\|^2\right\} \ge \frac{bC_1}{2}$$

This implies that if we set b such that $1 - 2e^{-\Omega(C_1 b)} = 0.975$, then with probability 0.975, B_j has $\frac{bC_1}{2}$ samples $\langle a_i, x\rangle$ whose magnitude is at least $t\|x\|$. This implies that the average square magnitude over the batch satisfies

$$\frac{1}{b}\|A_{B_j}x\|^2 = \frac{1}{b}\sum_{i\in B_j}\langle a_i, x\rangle^2 \ge t^2\|x\|^2\frac{bC_1}{2b} = \frac{C_1t^2\|x\|^2}{2}, \tag{41}$$

with probability 0.975.

Consider the indicator random variable associated with the complement of the above event. That is,

$$Y_j(x) = \left\{\frac{1}{b}\|A_{B_j}x\|^2 \le \frac{C_1t^2}{2}\|x\|^2.\right\}$$

From (41) we have that $\mathbb{E}\left[Y_j(x)\right] \le 0.025$.

Consider the sum of indicator random variables over M batches. By standard concentrations of Bernoulli random variables, we have with probabibility $1 - e^{-\Omega(M)}$,

$$\sum_{j=1}^{M} Y_j(x) \le 2\mathbb{E}\left[\sum_{j=1}^{M} Y_j(x)\right] \le 0.05.$$

This implies that there exist a subset of batches $J \subseteq [M]$ with $|J| \ge 0.95M$ such that

$$\frac{1}{b}\|A_{B_j}x\|^2 \ge \frac{C_1t^2\|x\|^2}{2} \;\forall\, j \in J,$$

with probability $1 - e^{-\Omega(M)}$. This shows that we have the statement of the proposition for a fixed vector in S_1.

We now show that this holds true for an $\epsilon-$cover of S_1. Let S_ϵ denote a minimial $\epsilon-$covering of S_1. That is, S_ϵ is a finite subset of S_1 such that for all $x \in S_1$, there exists $\tilde{x} \in S_\epsilon$ such that $\|x - \tilde{x}\| \le \epsilon$. Since S_1 has dimension $2k$ and diameter 1, we can find a set S_ϵ whose cardinality is at most $\left(O\left(\frac{1}{\epsilon}\right)\right)^{2k}$.

By a union bound, with probability $1 - e^{-\Omega(M)}|S_\epsilon|$, for all $\tilde{x} \in S_\epsilon$ there exists a subset of batches $J_{\tilde{x}} \subset [M]$ with $|J_{\tilde{x}}| \ge 0.95M$ such that

$$\frac{1}{b}\|A_{B_j}\tilde{x}\|^2 \ge \frac{C_1t^2}{2} \;\forall\, j \in J_{\tilde{x}} \tag{42}$$

Since $|S|_\epsilon \leq e^{O(k \log \frac{1}{\epsilon})}$, if $M = \Omega\left(k \log \frac{1}{\epsilon}\right)$, the above statement holds with probability $1 - e^{-\Omega(M)}$.

We now show that the statement of the proposition is true for all vectors in S_1. Since the proposition statement holds for an $\epsilon-$cover of S_1, we now only need to consider the effect of A at a scale of ϵ.

Now consider the set

$$S_2 = \{x - \tilde{x} : x \in S_1, \tilde{x} \in S_\epsilon, \|x - \tilde{x}\| \leq \epsilon\}.$$

Note that this a subset of all vectors in the span of W that have norm at most ϵ. That is, if

$$S_3 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| \leq \epsilon\},$$

we have $S_2 \subseteq S_3$.

For a vector $v \in \mathbb{R}^n$, consider the random variable

$$Z_i(v) = \mathbf{1}\left[\langle a_i, v\rangle \geq \frac{\sqrt{C_1}t}{2\sqrt{2}}\right].$$

Define the random process

$$\Psi\left(a_1, a_2, \cdots, a_m\right) = \sup_{v \in S_2} \frac{1}{m} \sum_{i=1}^m \mathbf{1}\left[|\langle a_i, v\rangle| \geq \frac{\sqrt{C_1}t}{2\sqrt{2}}\right].$$

By the bounded difference inequality, with probability $1 - 2e^{-C_2\delta^2}$,

$$\Psi(a_1, a_2, \cdots, a_m) \leq \mathbb{E}\left[\Psi(a_1, a_2, \cdots, a_m)\right] + \frac{\delta}{\sqrt{m}}$$

Since $S_2 \subseteq S_3$, we can bound the expectation of Ψ by

$$\mathbb{E}\left[\Psi(a_1, \cdots, a_m)\right] \leq \mathbb{E} \sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^m \mathbf{1}\left[|\langle a_i, v\rangle| \geq \frac{\sqrt{C_1}t}{2\sqrt{2}}\right] \tag{43}$$

$$\leq \mathbb{E} \sup_{v \in S_3} \sum_{i=1}^m \frac{|\langle a_i, v\rangle|}{mt\sqrt{C_1}/2\sqrt{2}} \tag{44}$$

$$= \mathbb{E} \sup_{v \in S_3} \sum_{i=1}^m \frac{2\sqrt{2}|\langle a_i, v\rangle|}{mt\sqrt{C_1}} \tag{45}$$

$$\leq \mathbb{E} \sup_{v \in S_3} \left|\sum_{i=1}^m 2\sqrt{2} \frac{|\langle a_i, v\rangle| - \mathbb{E}\left[|\langle a, v\rangle|\right]}{mt\sqrt{C_1}}\right| + \sup_{v \in S_3} \sum_{i=1}^m \frac{2\sqrt{2}\mathbb{E}\left[|\langle a, v\rangle|\right]}{mt\sqrt{C_1}} \tag{46}$$

Since a is isotropic and v has norm at most ϵ, by Jensen's inequality, we can bound the second term in the RHS by

$$\mathbb{E} \sup_{v \in S_3} \sum_{i=1}^m \frac{2\sqrt{2}\mathbb{E}\left[|\langle a, v\rangle|\right]}{mt\sqrt{C_1}} \lesssim \frac{\epsilon}{t\sqrt{C_1}}. \tag{47}$$

To bound the first term in the RHS, we use the Gine-Zinn symmetrization inequality [31, 68, 56]

$$\mathbb{E} \sup_{v \in S_3} \left|\sum_{i=1}^m 2\sqrt{2} \frac{|\langle a_i, v\rangle| - \mathbb{E}\left[|\langle a, v\rangle|\right]}{mt\sqrt{C_1}}\right| \lesssim \mathbb{E} \sup_{v \in S_3} \left|\sum_{i=1}^m \frac{\xi_i \langle a_i, v\rangle}{mt\sqrt{C_1}}\right| \tag{48}$$

where $\xi_i, i \in [m]$ are i.i.d ± 1 Bernoulli variables.

We can bound this by

$$\mathbb{E} \sup_{v \in S_3} \left|\sum_{i=1}^m \frac{\xi_i \langle a_i, v\rangle}{mt\sqrt{C_1}}\right| = \mathbb{E}_{\xi, A}\left[\sup_{v \in S_3} \left|\frac{\xi^T A v}{mt\sqrt{C_1}}\right|\right], \tag{49}$$

$$= \mathbb{E}_{\xi, A}\left[\sup_{z : \|z\| \leq \epsilon} \left|\frac{\xi^T A W z}{mt\sqrt{C_1}}\right|\right] \tag{50}$$

$$\leq \mathbb{E}_{\xi,A}\left[\frac{\epsilon\|\xi^T AW\|}{mt\sqrt{C_1}}\right] \tag{51}$$

$$\leq \frac{\epsilon\sqrt{\mathbb{E}_{\xi,A}\|\xi^T AW\|^2}}{mt\sqrt{C_1}} \tag{52}$$

$$= \frac{\epsilon\sqrt{\mathbb{E}_A \text{trace}(AWW^T A^T)}}{mt\sqrt{C_1}} \tag{53}$$

$$= \frac{\epsilon\sqrt{2km}}{mt\sqrt{C_1}} \lesssim \frac{\epsilon}{t}\sqrt{\frac{k}{mC_1}} \tag{54}$$

The third line follows from the Cauchy-Schwartz inequality, and the fourth line follows from Jensen's inequality.

Since $m = Mb$, from the above inequality and Eqn (47) we can now bound $\mathbb{E}\Psi$ as

$$\mathbb{E}\left[\Psi(a_1, \cdots, a_m)\right] \lesssim \frac{\epsilon}{t}\sqrt{\frac{k}{MbC_1}} + \frac{\epsilon}{t\sqrt{C_1}} \tag{55}$$

Substituting the above inequality into the bounded difference inequality, we have with probability at least $1 - e^{-\Omega(\delta^2)}$,

$$\Psi(a_1, a_2, \cdots, a_m) \lesssim \frac{\epsilon}{t}\sqrt{\frac{k}{MbC_1}} + \frac{\epsilon}{t\sqrt{C_1}} + \frac{\delta}{\sqrt{Mb}} \tag{56}$$

Setting $M = \Omega(k), \delta = O\left(\sqrt{\frac{M}{b}}\right), \epsilon = O\left(\frac{t}{b}\sqrt{C_1}\right)$, we can reduce the terms in the above inequality to

$$\frac{\epsilon}{t}\sqrt{\frac{k}{MbC_1}} \leq O\left(\frac{1}{b^{\frac{3}{2}}}\right), \tag{57}$$

$$\frac{\epsilon}{t\sqrt{C_1}} \leq O\left(\frac{1}{b}\right), \tag{58}$$

$$\frac{\delta}{\sqrt{Mb}} \leq O\left(\frac{1}{b}\right), \tag{59}$$

Since $b > 1$, the sum of these three terms is dominated by $O\left(\frac{1}{b}\right)$. From this, we can conclude that for small enough ϵ, δ, with probability $1 - e^{-\Omega\left(\frac{M}{b}\right)}$,

$$\Psi(a_1, a_2, \cdots, a_m) \leq \frac{0.05}{b} \tag{60}$$

$$\Rightarrow \sup_{v \in S_3} \sum_{i=1}^{m} \mathbf{1}\left[|\langle a_i, v\rangle| \geq \frac{t\sqrt{C_1}}{2\sqrt{2}}\right] \leq 0.05M. \tag{61}$$

This allows us to control the effect of A at a scale of ϵ. It says that there at most $0.05M$ samples on which vectors with magnitude at most ϵ have a magnitude greater than $\frac{t\sqrt{C_1}}{2\sqrt{2}}$ after interacting with A. This implies that there at least $0.95M$ batches in which all samples are well behaved.

Since we have control over an $\epsilon-$cover of S_1 as well as vectors at a scale of ϵ in S_1, we can now prove our result for all vectors in S_1.

For any $x \in S_1$, let $\tilde{x} \in S_\epsilon$ be the point in the $\epsilon-$cover which is closest to x. For a batch B_j, we can express $\|A_{B_j}x\|$ as

$$\frac{1}{\sqrt{b}}\|A_{B_j}x\| \geq \frac{1}{\sqrt{b}}\|A_{B_j}\tilde{x}\| - \frac{1}{\sqrt{b}}\|A_{B_j}(x - \tilde{x})\|. \tag{62}$$

From (42), there exists a subset of batches $J_{\tilde{x}} \subseteq [M]$ with $|J_{\tilde{x}}| \geq 0.95M$ such that

$$\frac{1}{\sqrt{b}}\|A_{B_j}\tilde{x}\| \geq \frac{\sqrt{C_1}t}{\sqrt{2}} \;\forall\; j \in J_{\tilde{x}}. \tag{63}$$

From (61), there exists a subset of batches $J_{x-\tilde{x}} \subseteq [M]$ with $|J_{x-\tilde{x}}| \geq 0.95M$ such that for all $j \in J_{x-\tilde{x}}$,

$$|\langle a_i, x - \tilde{x}\rangle| \leq \frac{\sqrt{C_1}t}{2\sqrt{2}} \ \forall \ i \in B_j \tag{64}$$

$$\Rightarrow \frac{1}{\sqrt{b}}\|A_{B_j}(x - \tilde{x})\| \leq \frac{\sqrt{C_1}t}{2\sqrt{2}}, \tag{65}$$

$$\Rightarrow -\frac{1}{\sqrt{b}}\|A_{B_j}(x - \tilde{x})\| \geq -\frac{\sqrt{C_1}t}{2\sqrt{2}}. \tag{66}$$

From the bounds on $\|A_{B_j}\tilde{x}\|$ and the bound on $\|A_{B_j}(x - \tilde{x}\|$, we can conclude that for all $x \in S_1$ there exist a subset of batches $J_x = J_{\tilde{x}} \cap J_{x-\tilde{x}}$ with cardinality at least $0.9M$ such that

$$\frac{1}{\sqrt{b}}\|A_{B_j}x\| \geq \frac{\sqrt{C_1}t}{2\sqrt{2}}, \ \forall \ j \in J_x. \tag{67}$$

This completes the proof, with $\gamma = \frac{\sqrt{C_1}t}{2\sqrt{2}} = \frac{t(1-t^2)}{C^2 2\sqrt{2}}$. □

D Proof of Lemma 5.4

Lemma (Lemma 5.4)**.** *Consider the setting of Lemma 5.3 with measurements satisfying $y = AG(z^*) + \eta$. For any $t > 0$ and noise variance σ^2, let the batch size b and number of batches M satisfy $b = \Theta(\frac{\sigma^2}{t^2})$ and $M = \Omega(kd \log n)$. Then with probability at least $1 - e^{-\Omega(m)}$, for all $z \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that*

$$\frac{1}{b}|\eta_{B_j}^T A_{B_j}(G(z) - G(z^*))| \leq t\|G(z) - G(z^*)\| \ , \forall j \in J.$$

Proof. Proposition A.1 shows that the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the range of $e^{O(kd \log n)}$ different $2k-$dimensional subspaces. This trivially implies that for a fixed $z^* \in \mathbb{R}^k$, the set $\{G(z) - G(z^*) : z \in \mathbb{R}^k\}$ also lies in the range of $e^{O(kd \log n)}$ different $2k-$dimensional subspaces.

Proposition D.1 guarantees the result for a single subspace with probability $1 - e^{-\Omega(M)}$. Since $M = \Omega(kd \log n)$ and the batch size is constant which depends on the noise variance σ^2 and t^2, the lemma follows from a union bound over the $e^{O(kd \log n)}$ subspaces. □

Proposition D.1. *Consider a single $2k-$dimensional subspace given by $S = \{Wz : W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}, z \in \mathbb{R}^{2k}\}$. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d rows drawn from a distribution satisfying Assumption (1) with constant C. If the batch size $b = \Theta\left(\frac{\sigma^2}{t^2}\right)$ and the number of batches satisfies $M = \Omega\left(k \log \frac{1}{\epsilon}\right)$, with probability $1 - e^{-\Omega(M)}$, for all $x \in S$, there exist a subset of batches $J_x \subseteq [M]$ with $|J_x| \geq 0.90M$ such that*

$$\frac{1}{b}|\eta_{B_j}^T A_{B_j}x| \leq t\|x\| \ , \forall j \in J.$$

Proof. Since the bound we want to prove is homogeneous, it suffices to show it for all vectors in S that have unit norm. Let $W \in \mathbb{R}^{n \times 2k}$ be the orthonormal matrix spanning S, and S_1 denote the set of unit norm vectors in its span. That is,

$$S_1 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| = 1, W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}\}.$$

Consider the set S_ϵ, which is a minimal $\epsilon-$covering of S_1. That is, for every $x \in S_1$, there exists $\tilde{x} \in S_\epsilon$ such that $\|\tilde{x} - x\| \leq \epsilon$.

For a fixed $\tilde{x} \in S_\epsilon$, and $t > 0$, by Chebyshev's inequality,

$$\Pr\left[\frac{1}{b}|\eta^T A_{B_j}\tilde{x}| \geq \frac{t}{2}\right] \leq \frac{\sum_{i \in B_j}\left(\eta_i^2 \langle a_i, \tilde{x}\rangle^2\right)}{b^2 t^2/4} \tag{68}$$

$$= \frac{b\sigma^2\|\tilde{x}\|^2}{b^2t^2/4} \tag{69}$$

$$= \frac{\sigma^2 4}{bt^2} \leq \frac{1}{40}, \tag{70}$$

if $b \geq \frac{160\sigma^2}{t^2}$.

Define the indicator random variable

$$Y_i(x) = \mathbf{1}\left\{\frac{1}{b}|\eta^T A_{B_i} x| \geq \frac{t}{2}\right\}.$$

From Eqn (70) we have

$$\mathbb{E}\left[Y_i(\tilde{x})\right] \leq \frac{1}{40}.$$

By concentration of Bernoulli variables, with probability $1 - e^{-\Omega(M)}$,

$$\sum_{j=1}^{M} Y_i(\tilde{x}) \leq 2\mathbb{E}\left[Y_1(\tilde{x})\right] \leq \frac{1}{20}.$$

This implies that for a fixed $\tilde{x} \in S_\epsilon$, with probability $1 - e^{-\Omega(M)}$, there exist a subset of batches $J_{\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that

$$\frac{1}{b}|\eta^T A_{B_j} \tilde{x}| \leq \frac{t}{2} \; \forall \; j \in J_{\tilde{x}}. \tag{71}$$

Since the size of S_ϵ is at most $\left(O\left(\frac{1}{\epsilon}\right)\right)^{2k}$, we can union bound over all $\tilde{x}$ in S_ϵ. Hence, if $M = \Omega\left(k \log \frac{1}{\epsilon}\right)$, then with probability $1 - e^{-\Omega(M)}$, for all $\tilde{x} \in S_\epsilon$, there exist a subset $J_{\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that

$$\frac{1}{b}|\eta^T A_{B_j} \tilde{x}| \leq \frac{t}{2} \; \forall \; j \in J_{\tilde{x}}. \tag{72}$$

This shows that the multiplier component is well behaved on a large fraction of the batches for an $\epsilon-$cover of S_1. Now we need to extend the argument to all vectors in S_1.

Now consider the set

$$S_2 = \{x - \tilde{x} : x \in S_1, \tilde{x} \in S_\epsilon, \|x - \tilde{x}\| \leq \epsilon\}.$$

Note that this a subset of all vectors in the span of W that have norm at most ϵ. That is, if

$$S_3 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| \leq \epsilon\},$$

we have $S_2 \subseteq S_3$.

For any $v \in \mathbb{R}^n$, define the random variable

$$Z_j(v) = \mathbf{1}\left\{|\eta_i a_i^T v| \geq \frac{t}{2}\right\}. \tag{73}$$

Now define the random process

$$\Psi(a_1, \cdots, a_m) = \sup_{v \in S_2} \frac{1}{m} \sum_{i=1}^{m} Z_i(v) \tag{74}$$

Since $S_2 \subseteq S_3$, we can bound $\mathbb{E}\left[\Psi\right]$ via

$$\mathbb{E}\left[\Psi\right] \leq \mathbb{E}\left[\sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^{m} Z_i(v)\right] \tag{75}$$

$$\leq \mathbb{E}\left[\sup_{v\in S_3}\frac{1}{m}\sum_{i=1}^{m}\frac{|\eta_i a_i^T v|}{t/2}\right] \tag{76}$$

$$\leq \mathbb{E}\left[\sup_{v\in S_3}\left|\frac{1}{m}\sum_{i=1}^{m}\frac{|\eta_i a_i^T v| - \mathbb{E}|\eta_i a_i^T v|}{t/2}\right|\right] + \mathbb{E}\left[\sup_{v\in S_3}\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{E}|\eta_i a_i^T v|}{t/2}\right] \tag{77}$$

We can bound the term on the right by

$$\mathbb{E}\left[\sup_{v\in S_3}\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{E}|\eta_i a_i^T v|}{t/2}\right] \leq \frac{\mathbb{E}\left[\sup_{v\in S_3}\|\eta_i\|_2 \ \ |\langle a_i, v\rangle|\right]}{t/2} \tag{78}$$

$$\lesssim \frac{\sigma\epsilon}{t}, \tag{79}$$

where we have used the Cauchy Schwartz inequality, followed by the fact that η is independent noise and has variance σ^2, a is isotropic, and $v \in S_3$ has norm at most ϵ.

To bound the term on the left, we use the Gine-Zinn symmetrization inequality [31, 68, 56]

$$\mathbb{E}\left[\sup_{v\in S_3}\left|\frac{1}{m}\sum_{i=1}^{m}\frac{|\eta_i a_i^T v| - \mathbb{E}|\eta_i a_i^T v|}{t/2}\right|\right] \lesssim \mathbb{E}\left[\sup_{v\in S_3}\left|\frac{1}{m}\sum_{i=1}^{m}\frac{\xi_i\eta_i a_i^T v}{t/2}\right|\right] \tag{80}$$

where $\xi_i, i \in [m]$ are i.i.d $\pm$ Bernoulli random variables.

Let $\xi\eta = (\xi_1\eta_1, \xi_2\eta_2, \cdots, \xi_m\eta_m)$ denote the the element wise product of the vectors $\xi = (\xi_1, \xi_2, \cdots, \xi_m)$ and $\eta = (\eta_1, \eta_2, \cdots, \eta_m)$. We can bound the above inequality by

$$\mathbb{E}\sup_{v\in S_3}\left|\sum_{i=1}^{m}\frac{\xi_i\eta_i\langle a_i, v\rangle}{mt/2}\right| = \mathbb{E}_{\xi,\eta,A}\left[\sup_{v\in S_3}\left|\frac{(\xi\eta)^T A v}{mt/2}\right|\right], \tag{81}$$

$$= \mathbb{E}_{\xi,\eta,A}\left[\sup_{z:\|z\|\leq\epsilon}\left|\frac{(\xi\eta)^T A W z}{mt/2}\right|\right] \tag{82}$$

$$\leq \mathbb{E}_{\xi,\eta,A}\left[\frac{\epsilon\|(\xi\eta)^T A W\|}{mt/2}\right] \tag{83}$$

$$\leq \frac{\epsilon\sqrt{\mathbb{E}_{\xi,\eta,A}\|(\xi\eta)^T A W\|^2}}{mt/2} \tag{84}$$

$$= \frac{\epsilon\sigma\sqrt{\mathbb{E}_A \text{trace}(AWW^T A^T)}}{mt/2} \tag{85}$$

$$= \frac{\epsilon\sigma\sqrt{2km}}{mt/2} \lesssim \frac{\epsilon\sigma}{t}\sqrt{\frac{k}{m}} \tag{86}$$

The third line follows from the Cauchy-Schwartz inequality, and the fourth line follows from Jensen's inequality, and the fifth line follows from the fact that $\xi\eta$ has i.i.d coordinates that are independent of A and have variance σ^2.

From the above inequality and eq. (78), we get

$$\mathbb{E}[\Psi(a_1, a_2, \cdots, a_m)] \lesssim \frac{\sigma\epsilon}{t}\sqrt{\frac{k}{m}} + \frac{\sigma\epsilon}{t} \lesssim \frac{\sigma\epsilon}{t} \tag{87}$$

If we choose $\epsilon = c_1\frac{t}{\sigma b}$ for a small enough constant c_1, then we can bound the expectation as

$$\mathbb{E}\left[\Psi(a_1, \cdots, a_m)\right] \leq \frac{0.025}{b} \tag{88}$$

By the bounded differences inequality, with probability $1 - e^{-\Omega(\delta^2)}$,

$$\Psi(a_1, \cdots, a_m) \leq \mathbb{E}\left[\Psi(a_1, \cdots, a_m)\right] + \frac{\delta}{\sqrt{m}} \tag{89}$$

Setting $\delta = 0.025\sqrt{\frac{M}{b}}$, we get $\frac{\delta}{\sqrt{m}} = \frac{0.025}{\sqrt{Mb}}\sqrt{\frac{M}{b}} = \frac{0.025}{b}$. This gives

$$\Psi(a_1, \cdots, a_m) \leq \frac{0.025}{b} + \frac{0.025}{b} = \frac{0.05}{b}. \tag{90}$$

From which we conclude that

$$\Rightarrow \sup_{v \in S_2} \sum_{i=1}^{m} \mathbf{1}\left\{|\eta_i a_i^T v| \geq \frac{t}{2}\right\} \leq \frac{0.05m}{b} = 0.05M. \tag{91}$$

Now consider any $x \in S_1$. There exists $\tilde{x} \in S_\epsilon$ such that $\|\tilde{x} - x\| \leq \epsilon$. From eq. (72) there exist a subset $J_{\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that

$$\frac{1}{b}|\eta_{B_j}^T A_{B_j} \tilde{x}| \leq \frac{t}{2} \; \forall \; j \in J_{\tilde{x}}. \tag{92}$$

Similarly, from eq. (91), there exists a subset $J_{x-\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that for all $j \in J_{x-\tilde{x}}$, we have

$$|\eta_i a_i^T (x - \tilde{x})| \leq \frac{t}{2} \; \forall \; i \in B_j, \tag{93}$$

$$\Rightarrow \frac{1}{b}|\eta_{B_j}^T A_{B_j} (x - \tilde{x})| \leq \frac{t}{2}. \tag{94}$$

From the triangle inequality and a simple union bound, for all $x \in S_1$, there exists a subset $J_x = J_{\tilde{x}} \cap J_{x-\tilde{x}}$ with cardinality $0.9M$ such that

$$\frac{1}{b}|\eta_{B_j}^T A_{B_j} x| \leq \frac{1}{b}|\eta_{B_j}^T A_{B_j} (x - \tilde{x})| + \frac{1}{b}|\eta_{B_j}^T A_{B_j} \tilde{x}| \tag{95}$$

$$\leq \frac{t}{2} + \frac{t}{2} = t \tag{96}$$

This completes the proof.

□

E Proof of Theorem 5.5

Proof. In Theorem 5.5, we fix the batch size b to be a suitable constant, specified in Lemma 5.3, Lemma 5.4. Then for $\epsilon \leq \frac{0.01}{b}$, the number of arbitrarily corrupted samples of A and y are at most $\frac{0.01}{b} bM = 0.01M$. This implies that there exist $0.99M$ batches with uncorrupted samples of A, y. For the rest of the proof, consider only these uncorrupted batches, and ignore the corrupted batches.

For a batch j, define the following

$$\mathbb{Q}_j(\widehat{z}, z^*) := \frac{1}{b}\|A_{B_j}(G(\widehat{z}) - G(z^*))\|^2, \tag{97}$$

$$\mathbb{M}_j(\widehat{z}) := \frac{2}{b}\eta_{B_j}^\top (A_{B_j}(G(\widehat{z}) - G(z^*))). \tag{98}$$

it is easy to verify that $\ell_j(\widehat{z}) - \ell_j(z^*) = \mathbb{Q}_j(\widehat{z}, z^*) - \mathbb{M}_j(\widehat{z})$. The component $\mathbb{Q}_j(\widehat{z}, z^*)$ is commonly called the quadratic component, and $\mathbb{M}_j(\widehat{z})$ is called the multiplier component.

By Lemma 5.2, the minimum value of the MOM objective is at most $4\sigma^2$ with high probability. Since $\widehat{z}$ minimizes the objective eq. (2) to within additive τ of the optimum, it implies that the median batch satisfies

$$\mathbb{Q}_j(\widehat{z}, z^*) - \mathbb{M}_j(\widehat{z}) \leq 4\sigma^2 + \tau. \tag{99}$$

Using Lemma 5.3, Lemma 5.4 on the $0.99M$ batches that do not have corruptions, if the batch size is a large enough constant, we see that there exist $0.78M$ batches on which both the following inequalities hold

$$\gamma^2\|G(\widehat{z}) - G(z^*)\|^2 \leq \mathbb{Q}_j(\widehat{z}, z^*) \text{ and } -\sigma\|G(\widehat{z}) - G(z^*)\| \leq -\mathbb{M}_j(\widehat{z}). \tag{100}$$

Putting the above two inequalities together, the median batch satisfies

$$\gamma^2\|G(\widehat{z}) - G(z^*)\|^2 - \sigma\|G(\widehat{z}) - G(z^*)\| \leq 4\sigma^2 + \tau.$$

Solving the quadratic inequality for $\|G(\widehat{z}) - G(z^*)\|$, we have

$$\|G(\widehat{z}) - G(z^*)\|^2 \lesssim \sigma^2 + \tau. \qquad \square$$

F Experimental Setup

F.1 MNIST dataset

We first compare Algorithm 1 with the baseline ERM [15] for heavy tailed dataset *without* arbitrary corruptions on MNIST dataset [55]. We trained a DCGAN [80] to produce 64×64 MNIST images.[3] We choose the dimension of the latent space as $k = 100$, and the model has 5 layers.

Based on this generative model, the uncorrupted compressed sensing model P has heavy tailed measurement matrix and stochastic noise: $y = AG(z^*) + \eta$. We consider a Student's t distribution (a typical example of heavy tails) – the measurement matrix A is generated from a Student's t distribution with degrees of freedom 4, and η with degrees of freedom 3 with bounded variance σ^2. We vary the number of measurement m and obtain the reconstruction error $\|G(\widehat{z}) - G(z^*)\|^2$ for Algorithm 1 and ERM, where $G(z^*)$ is the ground truth image. Each curve in Figure 1a demonstrates the averaged reconstruction error for 50 trials. In Figure 1a, Algorithm 1 and ERM both have decreasing reconstruction error per pixel with increasing number of measurement. In particular, Algorithm 1 obtains significantly smaller reconstruction error comparing with the baseline ERM.

F.2 CelebA-HQ dataset

We continue the study of empirical performance of our algorithm on real image datasets with higher quality. We generate high quality RGB images with size 256×256 from CelebA-HQ[4]. Hence the dimension of each image is $256 \times 256 \times 3 = 196608$. In all of our experiments, we fix the dimension of the latent space as $k = 512$, and train a DCGAN on this dataset to obtain a generative model G.

We first compare our algorithm with the baseline ERM [15] for heavy tailed dataset without arbitrary corruptions, and then deal with the situation of outliers.

Heavy tailed samples. In this experiment, we deal with the *uncorrupted* compressed sensing model P, which has heavy tailed measurement matrix and stochastic noise: $y = AG(z^*) + \eta$. We also use a Student's t distribution for A and η – the measurement matrix A is generated from a Student's t distribution with degrees of freedom 4, and stochastic noise η with degrees of freedom 3 with a bounded variance.

We obtain the reconstruction error $\|G(\widehat{z}) - G(z^*))\|$ vs. the number of measurement m for our algorithm and ERM, where z^* is the ground truth. In Figure 1b, each curve is an average of 20 trials. For heavy tailed y and A without any corruption, both methods are consistent, and have decaying reconstruction error with increasing sample size. Our method obtains significantly smaller reconstruction error, and shows competitive results over the baseline ERM for heavy tailed data set, even without any arbitrary outliers.

[3] Code was cloned from the following repository https://github.com/pytorch/examples/tree/master/dcgan.

[4] Code was cloned from the following repository: https://github.com/facebookresearch/pytorch_GAN_zoo.

F.3 Hyperparameter selection

When using the Adam [52] optimizer, we varied the learning rate over $[0.1, 0.05, 0.01, 0.005]$ for our algorithm and baselines. When using the Yellowfin [91] optimizer, we varied our learning rates over $[10^{-4}, 5\cdot 10^{-5}, 10^{-5}, 5\cdot 10^{-6}, 10^{-6}]$. We selected the best learning rate based on fresh measurements that were not used for optimization.

G Background

Theorem G.1 (Ledoux-Talagrand Contraction Inequality)**.** *For a compact set $\mathcal{T}$, let $x_1, \cdots, x_m$ be i.i.d vectors whose real valued components are indexed by $\mathcal{T}$, i.e., $x_i = (x_{i,s})_{s\in\mathcal{T}}$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function such that $\phi(0) = 0$. Let $\epsilon_1, \cdots, \epsilon_m$ be independent Rademacher random variables. Then*

$$\mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{m}\epsilon_i\phi(x_{i,s})\right|\right] \leq 2\mathbb{E}\left[\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{m}\epsilon_i x_{i,s}\right|\right].$$

Theorem G.2 (Talagrand's Inequality for Bounded Empirical Processes)**.** *For a compact set $\mathcal{T}$, let $x_1, \cdots, x_m$ be i.i.d vectors whose real valued components are indexed by $\mathcal{T}$, i.e., $x_i = (x_{i,s})_{s\in\mathcal{T}}$. Assume that $\mathbb{E}x_{i,s} = 0$ and $|x_{i,s}| \leq b$ for all $s \in \mathcal{T}$. Let $Z = \sup_{s\in\mathcal{T}}\left|\frac{1}{m}\sum_{i=1}^{m} x_{i,s}\right|$. Let $\sigma^2 = \sup_{s\in\mathcal{T}}\mathbb{E}x_s^2$ and $\nu = 2b\mathbb{E}Z + \sigma^2$. Then*

$$\Pr\left[Z \geq \mathbb{E}Z + t\right] \leq C_1 \exp\left(-\frac{C_2 m t^2}{\nu + bt}\right).$$

where C_1, C_2 are absolute constants.