Efficient algorithms for linear regression with heteroskedastic errors

Siddhant Chaudhary
Kahlert School of Computing
University of Utah
Salt Lake City, UT, USA
urssidd@gmail.com

Aditya Bhaskara
Kahlert School of Computing
University of Utah
Salt Lake City, UT, USA
bhaskaraaditya@gmail.com

Silvio Lattanzi
Google Research
Barcelona, Spain
silviol@google.com

Sergei Vassilvitskii
Google Research
New York, NY, USA
sergeiv@google.com

Abstract

We study the classical problem of linear regression under heteroskedastic noise, where observations arise from multiple sources with unknown and potentially unequal variances. Specifically, we consider a setting with n sources, each providing a linear measurement of an unknown d-dimensional parameter vector β^*, where each source is associated with an unknown noise variance. We focus on the classic "subset-of-signals" model, where one assumes that m of the sources have a relatively low noise, formally, with variance of at most 1.

Our goal is to accurately estimate β^* despite the presence of high-variance sources. We show that if m is sufficiently large, it is possible to recover β^* with sub-constant error. Specifically, we give efficient algorithms with a recovery error bound of $\tilde{O}\left(\frac{\sqrt{nd}}{m}\right)$ in a natural data-dependent norm (and $\tilde{O}\left(\frac{d\sqrt{n}}{m}\right)$ in the ℓ_2 norm for well-spread designs), significantly improving upon previously known results for this setting.

1 Introduction

One of the central challenges in modern machine learning is dealing with datasets where individual data points may vary significantly in quality. In the statistical literature, this phenomenon is called *heteroskedasticity*, where different observations are corrupted by noise at different (and possibly unknown) levels. Even for the most basic statistical tasks—such as mean estimation and linear regression—heteroskedasticity can severely degrade the performance of standard estimators like the empirical mean and ordinary least squares (OLS), which implicitly assume uniform noise levels across samples White [1980], Politis and Poulis [2014].

Beyond classical statistics, heteroskedasticity has been extensively studied in applied domains such as crowdsourcing and sensor networks, where data is collected from multiple agents or devices of varying reliability. In such settings, different workers or sensors may contribute measurements with widely differing levels of noise or bias Ghosh et al. [2011], Chierichetti et al. [2014]. As machine learning systems continue to scale and rely increasingly on large, heterogeneous, and weakly curated datasets, it becomes crucial to design algorithms that are robust to this variability in "quality".

A common practical approach to handling heteroskedastic data is to reweigh observations based on their perceived reliability. From a statistical standpoint, this idea is formalized by the weighted

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

least squares (WLS) estimator, which is known to be optimal when the noise variances are known White [1980]. However, in many modern applications, these variances are unknown and may even be difficult to estimate reliably owing to the fact that we may observe only one observation per "source" or sensor. This makes the design of principled algorithms for learning under heteroskedastic noise a challenging and largely unsolved problem.

In this paper, we study the classic problem of linear regression in the presence of heteroskedastic noise. Formally, we are given n observations (x_i, y_i), where

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i,$$

and the noise variables ϵ_i are independent but have different and unknown variances. This problem is known to be challenging when no assumptions are made on the noise distribution. Prior works have typically addressed this challenge by either using heuristics to estimate the variances (see Davidian and Carroll [1987]), or by imposing structural assumptions or parametric models on the noise, such as assuming specific functional forms or variance structures (see Das et al. [2023], Harvey [1976], Xu and Shimada [2000]). While these approaches yield meaningful guarantees under their respective models, they do not address the fully agnostic setting where the noise levels are arbitrary and unknown.

Our goal is to study regression in this fully heteroskedastic setting without making modeling assumptions on the noise. Inspired by recent advances in heteroskedastic mean estimation, we adopt a *subset-of-signals* perspective. Specifically, we assume that there exists an *unknown* subset $S \subseteq [n]$ of size at least m such that the corresponding noise levels are bounded, i.e., $\sigma_i \leq 1$ for all $i \in S$, while no assumptions are made on the noise level of the remaining samples. The identity of this subset is unknown to the algorithm, and we are particularly interested in the regime where $m \ll n$, so that only a small fraction of the data is reliable.

The subset-of-signals model has been extensively studied in the context of mean estimation. Here, one assumes that we have n Gaussians $\{\mathcal{N}(\mu, \sigma_i^2)\}_{i=1}^n$ that have a common mean, and for an unknown $S \subset [n]$ of size m, we have $\sigma_i \leq 1$ for $i \in S$. Early work by Chierichetti et al. [2014], motivated by applications in crowdsourcing (e.g., Ghosh et al. [2011]) and by classical statistical ideas such as density-based estimation (e.g., Chernoff [1964]), initiated the study of mean estimation under heteroskedastic errors in observation. (The problem is also commonly referred to as learning of *entangled Gaussians*.) Pensia et al. [2019] formalized the subset-of-signals model and developed algorithms with provable guarantees, as well as lower bounds. Subsequent works, including those of Liang and Yuan [2020], Compton and Valiant [2024], and Diakonikolas et al. [2025a], have significantly advanced our understanding of this problem, achieving near-optimal error rates. In one dimension, these works were able to show that μ can be recovered to a *sub-constant* error as long as $m > n^{1/4}$. (For comparison, note that if S were known, we can always achieve error $m^{-1/2}$.) In their recent work, Han et al. [2026] showed that an empirical Bayes approach achieves minimax error bounds for this problem, for all ranges of m.

Despite this progress for mean estimation, much less was known for regression at the time of writing this paper. In particular, obtaining *sub-constant error guarantees* for regression in the subset-of-signals model has remained challenging, even under distributional assumptions on the covariate vectors $\{x_i\}$ (see Pensia et al. [2019]). A closely related and well-studied model for regression is the *oblivious noise model*, wherein we have an adversarial noise vector η with $m = \alpha n$ entries that are ≤ 1 [d'Orsi et al., 2021, Diakonikolas et al., 2025b]. d'Orsi et al. [2021] studied the Huber loss estimator for this setting, and provided error bounds that are qualitatively similar to our bounds. Shortly after the time of writing this paper, the work of Compton and Schramm [2026] extended their balance-finding estimator from mean estimation to regression, and proved that their algorithm achieves minimax error bounds for regression. Notably, they prove their results for symmetric and unimodal errors, and show that, up to a factor involving the dimension d, the minimax bound for regression is equivalent to that of mean estimation. However, most of these works have distributional assumptions on *all* covariates x_i (even for $i \notin S$).

In this work, we extend and introduce analysis techniques for two known algorithms for regression: L1 regression [Pollard, 1991] and iterative truncation [Liang and Yuan, 2020], and show that they achieve sub-constant error bounds for the subset-of-signals model.

1.1 Our Results

Setup and Notation. In what follows, let $\beta^* \in \mathbb{R}^d$ be the vector we wish to estimate. We have n noisy linear observations of this vector. Specifically, for each $i \in [n]$, there is a distribution $\mathcal{D}_i$ over $\mathbb{R}^d$ and a noise level σ_i (both of which are unknown). Each observation has the form (x_i, y_i), where $x_i \sim \mathcal{D}_i, \epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, and

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i. \tag{1}$$

We assume that x_i and ϵ_i are independent. We also assume that the support of $\mathcal{D}_i$ is on vectors x with $\|x\| \leq 1$; we note that such scaling is essentially without loss of generality. Once we choose the scale of $\|x_i\|$, note that we fix the scaling of β^* and ϵ_i.

Subset of Signals model. As a stepping stone for studying heteroskedastic noise, Pensia et al. [2019] introduced the "subset-of-signals" model. We have a parameter $1 \leq m \leq n$, and we assume that for m of the indices $[n]$, we have $\sigma_i \leq 1$, and we denote this set of m indices by S (Note that mean estimation is a special case of regression, where we have $d = 1$ and $x_i = 1$ for all i). We assume that the set S is chosen before the covariates x_i are sampled. For regression, we give two main results. First, we study one of the most classic estimators for noisy data, $L1$ regression [Pollard, 1991]. We show that, interestingly, it yields non-trivial guarantees in a heteroskedastic noise model, as long as $m \gtrsim d\sqrt{n \log n}$. Our second result is an algorithmic contribution, where we adapt an iterative truncation algorithm that was used earlier for mean estimation. The key idea is that iterative truncation can be applied with "residual errors", and this leads to recovery guarantees. We now discuss our results in detail.

$L1$ Error Minimization. Given the setup of (1), a natural estimator for β^* is

$$\widehat{\beta} := \text{argmin}_\beta \sum_{i \in [n]} |y_i - x_i^T \beta|. \tag{2}$$

The estimator is very well-studied, and is known to be computable efficiently (e.g., see van den Brand et al. [2021]). We prove the following, assuming a "well spread" property (Definition 3.1):

Theorem 1.1. *Consider $\{(x_i, y_i)\}_{i=1}^n$ satisfying the subset of signals model, and suppose that the design distributions $\{\mathcal{D}_i\}_{i \in S}$ satisfy the well-spread property from Definition 3.1. Let $\hat{\beta}$ be the $L1$ regression estimator. Then, there exist constants $C, C' > 0$, depending only on the constant c from Definition 3.1, such that if $n \geq C \cdot d \log n$ and $m \geq C\sqrt{nd \log n}$, with probability at least $1 - 2n^{-2}$,*

$$\|\hat{\beta} - \beta^*\| \leq C' \frac{d\sqrt{n \log n}}{m} = \widetilde{O}\left(\frac{d\sqrt{n}}{m}\right).$$

As we discuss in Section 3, Theorem 3.2 makes relatively minimal assumptions: we make no assumptions on the distributions $\mathcal{D}_i$ for $i \notin S$: they could be fully adversarial. In addition, no assumptions are made on the values of σ_i^2 for $i \notin S$. It is, however, important that the $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ (we use the symmetry of the Gaussian). The theorem also shows that we obtain sub-constant error in recovery as long as $m \gtrsim d\sqrt{n \log n}$.[1] In spite of $L1$ regression being a ubiquitous tool, such guarantees were not previously known in the heteroskedastic error model.

Remark. The choice of $\sigma_i \leq 1$ in the definition of the subset of signals model is somewhat arbitrary. If we define "small noise" using a parameter σ_τ, our error bound changes to $\widetilde{O}\left(\frac{d\sigma_\tau\sqrt{n}}{m(\sigma_\tau)}\right)$, where $m(\sigma_\tau)$ is the number of indices i such that $\sigma_i \leq \sigma_\tau$. Since the algorithm itself is oblivious of σ_τ, the bound we obtain is actually the minimum over all such bounds. For notational simplicity, we work with the case $\sigma_\tau = 1$.

Iterative Truncation. Our second result is an algorithm that is based on the idea of iterative truncation [Liang and Yuan, 2020, Yuan and Liang, 2020]. The key here is to adapt the algorithm to the case of regression, where we perform iterative trimming on the "residual" values. We then prove the following. We use the fairly standard notation $\|\cdot\|_\Sigma$, defined in Section 2.

[1] We use the notation $x \gtrsim y$ to mean that $x \geq cy$ for a sufficiently large constant c.

Theorem 1.2 (Iterative truncation algorithm)**.** *Consider $\{(x_i, y_i)\}_{i=1}^n$ satisfying the subset of signals model for regression. Assume that the x_i also satisfy $\|x_i\| \le 1$ for all $i \in [n]$, and additionally, suppose that the covariances satisfy*

$$\Sigma_S := \frac{1}{m}\sum_{i\in S} x_i x_i^\top \succeq c_1 \Sigma_n, \qquad \Sigma_n := \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \succ 0, \qquad \frac{1}{m}\sum_{i\in S}\langle x_i, u\rangle^4 \le C\left(\frac{1}{n}\sum_{i\in[n]}\langle x_i, u\rangle^2\right)^2 \quad \forall u \in \mathbb{R}^d$$

for some absolute constants $c_1 \in (0,1]$ and $C > 0$. Finally, suppose we are given an initial estimate β_0 that satisfies $2\|\beta_0 - \beta^*\|_{\Sigma_n} \le B$, for some known parameter $B \ge 1$. Then, there is an efficient algorithm that, for $m \gtrsim \sqrt{dn \ln n}$, outputs $\widehat{\beta}$ such that with probability at least $1 - O\big(\frac{KT}{n^3}\big)$ over the noise $\{\epsilon_i\}$,

$$\|\widehat{\beta} - \beta^*\|_{\Sigma_n} \le \widetilde{O}\left(\frac{\sqrt{dn}}{m}\right).$$

Above, K is the number of "stages" of the algorithm, and T the number of iterations within each stage. For more details, see Section 4.

For a full statement (including an ℓ_2 error bound), see Theorem 4.1 and the remarks that follow it. The key distinction between Theorems 1.1 and 1.2 is in (i) the assumptions, and (ii) the algorithmic complexity. The first result has the advantage of being based on a classic algorithm (which has been heavily optimized), and additionally, of being completely oblivious to the vectors $\{x_i\}_{i\notin S}$. The second result involves conditions on *all* the x_i, but in principle, there is *no distributional assumption* on the $\{x_i\}$; the bound holds as long as the realized vectors satisfy a (fairly mild) condition number guarantee! So in essence, only the $\{\epsilon_i\}$ are random, and $\{x_i\}$ can be any set of vectors satisfying the conditions. The algorithm for Theorem 1.2 is also a novel modification of the ideas of Liang and Yuan [2020]. It is interesting to note that the two approaches obtain similar guarantees under a different set of hypotheses.

2 Notation and Organization

We will use standard matrix and vector notation throughout the paper. For a vector x, we denote by $\|x\|_2$ the $L2$ norm of the vector. When it is understood from the context, we sometimes write it as simply $\|x\|$. We also consider norms with respect to positive definite matrices. For a matrix $\Sigma \succ 0$ (i.e., positive definite), we write $\|x\|_\Sigma := \sqrt{x^T \Sigma x} = \left\|\Sigma^{1/2} x\right\|_2$.

We also denote by $\mathcal{B}(x, r)$ the ℓ_2 norm ball of radius r around x. We also use $\mathcal{B}_\Sigma(x, r)$ to denote $\{z : \|z - x\|_\Sigma \le r\}$. Finally, in Section 4, we use the notion of the projection of an $x \in \mathbb{R}$ onto an interval $[a, b]$. The projection $\phi(x; [a, b])$ is defined simply as a if $x < a$ as b if $x > b$ and as x otherwise.

Organization. For reasons of space, Additional Related Work is deferred to Appendix A. Sections 3 and 4 will present the two main results of the paper respectively.

3 The L1 Regression Estimator

To recap, we are given n observations $\{(x_i, y_i)\}_{i\in[n]}$ generated according to the linear model $y_i = \langle x_i, \beta^*\rangle + \epsilon_i$, where $\beta^* \in \mathbb{R}^d$ is the unknown target vector, and $x_i \sim \mathcal{D}_i$. Each $\mathcal{D}_i$ has support restricted to the unit ball in $\mathbb{R}^d$. The noise terms $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent of x_i, and also of each other. We consider the natural estimate for β^* using L_1 regression. I.e., define

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^d} \sum_{i=1}^n |y_i - \langle x_i, \beta\rangle|. \tag{3}$$

The estimator $\hat{\beta}$ is well known to be computable in polynomial time, e.g., via Linear Programming. There is extensive literature on nearly-linear time algorithms for L1 regression (see van den Brand et al. [2021], Cohen et al. [2019], Durfee et al. [2018] and references therein).

Recall that in the subset of signals model, there exists an (unknown) $S \subset [n]$ such that $|S| \geq m$ and $\sigma_i \leq 1$ for all $i \in S$. For $i \in [n] \setminus S$, the σ_i can be completely arbitrary. We also assume that for every $i \in S$, the design distribution $\mathcal{D}_i$ satisfies the following property:

Definition 3.1 (Well-Spread Covariate Distributions)**.** Let $\mathcal{D}_i$ be a distribution over $\mathbb{R}^d$ with support within the unit ball. We say that $\mathcal{D}_i$ is *well-spread* if there is an absolute constant $c > 0$ such that for every $\theta \in \mathbb{S}^{d-1}$,

$$\Pr_{x_i \sim \mathcal{D}_i}\left[\langle x_i, \theta\rangle \geq \frac{c}{\sqrt{d}}\right] \geq \frac{1}{4}.$$

Remark. When x_i is chosen uniformly at random from the unit sphere, we expect $\langle x_i, \theta\rangle$ to have magnitude $1/\sqrt{d}$ (for any θ). The definition states that with a constant probability, we get approximately this correlation. We treat c as a constant for simplicity, but our analysis tracks the dependence on c explicitly (the final error bound scales as $1/c^2$), so it can also be viewed as a parameter. Note that such an assumption on $\mathcal{D}_i$ is *necessary* in some sense. As an example in 2D, suppose that all the "small variance" $\mathcal{D}_i$ only had support along the x-axis, then no algorithm could accurately recover the y-coordinate of β. The well-spread assumption thus guarantees that $i \in S$ (i.e., the high-quality data) provides a "spanning view" of the entire parameter space. Our main result is then the following.

Theorem 3.2. *Consider the subset of signals model for regression described above, and suppose that the design distributions $\{\mathcal{D}_i\}_{i\in S}$ satisfy the well-spread property. Let $\hat{\beta}$ be the L1 regression estimator. Then there exist constants $C, C' > 0$, depending only on the constant c from Definition 3.1, such that if $n \geq Cd\log n$ and $m \geq C\sqrt{nd\log n}$, then with probability at least $1 - 2n^{-2}$,*

$$\|\hat{\beta} - \beta^*\| \leq C'\frac{d\sqrt{n\log n}}{m} = \widetilde{O}\left(\frac{d\sqrt{n}}{m}\right).$$

Analysis Outline. The remainder of the section proves Theorem 3.2; we start with an outline of the argument. Denote $\Delta := \hat{\beta} - \beta^*$. We also write $\tau = \|\Delta\|$ and let $\overline{\Delta} = \frac{\Delta}{\tau}$ denote the unit vector along Δ. We study the following one-dimensional function, defined on the line joining β^* and $\hat{\beta}$:

$$\psi(t) := F(\beta^* + t\cdot\overline{\Delta}), \quad \text{where} \quad F(\beta) = \sum_{i=1}^{n} |y_i - \langle x_i, \beta\rangle|. \tag{4}$$

By definition, $\psi(t)$ is minimized at $t = \tau$.

The main steps of the analysis are as follows: we look at the magnitude of the derivative of $\psi(t)$ at the points $t = 0$ and $t = \tau$. First, since τ is a minimizer, we will be able to conclude that the left derivative at τ (defined below) must be ≤ 0 (Lemma 3.3). Then, using a standard Hoeffding bound, we argue that the magnitude of the derivative at the origin is bounded by $O(\sqrt{n\log n})$ with high probability (Lemma 3.4). These two lemmas imply that the difference between the two derivative values is $O(\sqrt{n\log n})$.

Finally, we argue that if there are m indices such that $\sigma_i \leq 1$, then the difference between the two derivatives above must be *lower bounded* by a term that scales roughly as $m\cdot\tau/\mathrm{poly}(d)$ (Lemma 3.6). The challenge here turns out to be that we do not know the direction of $\overline{\Delta}$; due to this we use an ϵ-net argument along with the well-spread condition. Once we have this lower bound on the gap, we can use the earlier lemmas to conclude that $\frac{m\cdot\tau}{\mathrm{poly}(d)} \lesssim \sqrt{n\log n}$, which then lets us conclude the result.

We now make the argument formal, and consider $\psi(t)$ from (4). For the technical reason of non-differentiability of the L_1 norm, we consider the derivatives $\psi'_+(0) := \lim_{\delta\to 0}\psi'(\delta)$, and $\psi'_-(\tau) := \lim_{\delta\to 0}\psi'(\tau - \delta)$. (Here, $\delta \to 0$ is taken to mean approaching from the positive side.)

Lemma 3.3. *For the function $\psi(t)$ defined above, we have*

$$\psi'_-(\tau) \leq 0.$$

Proof. By definition, $\psi(t) = F(\beta^* + t\overline{\Delta})$, and ψ is a convex function of the scalar variable t, since it is the composition of a convex function (F) with an affine map. Moreover, ψ is minimized at $t = \tau$, because $\beta^* + \tau\overline{\Delta} = \hat{\beta}$ and $\hat{\beta}$ minimizes F. Now for any convex function, the left derivative at a minimizer is nonpositive, and this completes the proof. □

Lemma 3.4. *Consider the function $\psi(t)$ defined above, and let $n \geq 2$. With probability at least $1 - n^{-3}$,*

$$\left|\psi'_+(0)\right| \leq C_1\sqrt{n \log n}, \qquad \textit{where } C_1 := 2\sqrt{2}.$$

Proof. By the definition of ψ, we can also write it as $\psi(t) = \sum_{i=1}^n \left| \epsilon_i - t\langle x_i, \overline{\Delta}\rangle \right|$. For each i, define $r_i := \operatorname{sign}(-\epsilon_i)$. Since the ϵ_i are independent centered Gaussians, the variables r_i are i.i.d. Rademacher random variables, and they are independent of the x_i.

For all i such that $\epsilon_i \neq 0$ (which holds almost surely), the right derivative at 0 is

$$\psi'_+(0) = \sum_{i=1}^n \langle x_i, \overline{\Delta}\rangle \operatorname{sign}(-\epsilon_i) = \sum_{i=1}^n r_i \langle x_i, \overline{\Delta}\rangle = \left\langle \sum_{i=1}^n r_i x_i, \overline{\Delta} \right\rangle.$$

Hence since $\overline{\Delta}$ is a unit vector, by Cauchy–Schwarz, $\left|\psi'_+(0)\right| \leq \left\|\sum_{i=1}^n r_i x_i\right\|_2$.

Now apply a standard Hilbert-space (i.e., vector valued) Hoeffding inequality for Rademacher sums (e.g., see Pinelis and Sakhanenko [1986]): since $\|r_i x_i\|_2 \leq 1$ and the summands are independent and mean zero,

$$\Pr\left[\left\|\sum_{i=1}^n r_i x_i\right\|_2 \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \|x_i\|_2^2}\right) \leq 2\exp\left(-\frac{t^2}{2n}\right).$$

Taking $t = \sqrt{2n\log(2n^3)}$ shows that with probability at least $1 - n^{-3}$,

$$\left\|\sum_{i=1}^n r_i x_i\right\|_2 \leq \sqrt{2n\log(2n^3)} \leq 2\sqrt{2}\sqrt{n\log n},$$

where the last inequality uses $n \geq 2$. Therefore the same bound holds for $|\psi'_+(0)|$. □

Before stating the next lemmas, let us define $a_i := \left\langle x_i, \overline{\Delta}\right\rangle$, for convenience, and $\phi_i(t) := |\epsilon_i - t a_i|$. Then we have

$$\psi(t) = \sum_{i=1}^n |\epsilon_i - t a_i| = \sum_{i=1}^n \phi_i(t).$$

We first have the following simple technical lemma:

Lemma 3.5. *Let $a, \epsilon \neq 0$ and $\tau > 0$ be given. Define $\phi(t) := |\epsilon - t \cdot a|$. Then:*

1. *For any choice of the parameters, we have $\phi'_+(0) \leq \phi'_-(\tau)$.*
2. *When ϵ lies between 0 and τa, $\phi'_+(0) \leq \phi'_-(\tau) - 2|a|$.*

Proof. Since $a \neq 0$, we can write $\phi(t) = |a|\left|t - \frac{\epsilon}{a}\right|$. Thus $\phi'(t)$ is $-|a|$ for $t < \frac{\epsilon}{a}$, and $|a|$ for $t > \frac{\epsilon}{a}$.

Now consider the following cases:

Case 1. $\epsilon/a < 0$. In this case, both the derivatives of interest are $|a|$, and thus $\phi'_+(0) = \phi'_-(\tau)$.

Case 2. $\epsilon/a > 0$. In this case $\phi'_+(0) = -|a|$. Now if $\epsilon/a < \tau$, we have $\phi'_-(\tau) = |a|$ and so we are in the second case of the Lemma. If $\epsilon/a \geq \tau$, then $\phi'_-(\tau) = -|a|$, and we are once again in the first case.

Since $\epsilon \neq 0$, these are the only two cases, and this completes the proof. □

Next we come to the crucial part of analysis: showing that if there are sufficiently many indices i such that σ_i is small (i.e., $|S| \geq m$), then the difference $\psi'_+(0) - \psi'_-(\tau)$ is large.

Lemma 3.6. *Suppose we are in the subset-of-signals model, i.e., there exists $S \subseteq [n]$ with $|S| \geq m \geq C_m d\log(nd)$, for a sufficiently large constant C_m depending only on c, such that $\sigma_i \leq 1$ for all $i \in S$, and suppose that the corresponding $\mathcal{D}_i$ satisfy the well-spread property. Then there is an*

absolute constant $C_0 > 0$ such that with probability at least $1 - n^{-2}$ (over the choice of $\{x_i, \epsilon_i\}_{i \in S}$), we have

$$\psi'_-(\tau) - \psi'_+(0) \geq \frac{cm}{64\sqrt{d}} \min\left(1, \frac{c\tau}{2\sqrt{d}}\right) \quad \text{for all } \tau \geq C_0 \frac{d^{3/2}(\log d + \log n)}{m}.$$

(The constant c here comes from the definition of the well-spread property.)

Proof. The rough idea is to argue that for sufficiently many indices $i \in S$, for the function $\phi_i(t)$, we fall into part (2) of Lemma 3.5. But the challenge in the analysis is that Δ, and therefore the parameters a_i, are *dependent* on the realized x_i, ϵ_i, and thus we cannot compute probabilities easily. We thus use an ϵ-net argument and prove our desired claim for *all* directions, not just $\overline{\Delta}$.

First, let us fix some $\tau \geq C_0 \frac{d^{3/2} \log(nd)}{m}$, as in the hypothesis of the lemma. Then define

$$\eta := \frac{c}{2\sqrt{d}}, \qquad Y_i := \mathbf{1}[\, 0 < \epsilon_i < \eta\tau \,],$$

To calculate $\Pr[Y_i = 1]$, we use the following standard lemma (see Appendix B.1 for a proof).

Lemma 3.7. *Let $Y \sim \mathcal{N}(0, \sigma^2)$. Then for all $t \geq 0$,*

$$\Pr(0 < Y < t) \geq \min\left(\frac{t}{4\sigma}, \frac{1}{4}\right)$$

Now, let us consider any *fixed* vector $\theta \in \mathbb{S}^{d-1}$ and define $h(\theta) := \sum_{i \in S} Y_i \cdot \mathbf{1}\Big[\langle x_i, \theta \rangle \geq \frac{c}{\sqrt{d}}\Big]$.

Every term in the summation is clearly 0 or 1. Now consider a random choice of ϵ_i, x_i. Since Y_i depends only on ϵ_i and is independent of x_i, we have for every $i \in S$, using the well-spread property along with Lemma 3.7 with $\sigma_i \leq 1$, $\Pr\Big[Y_i = 1 \text{ and } \langle x_i, \theta \rangle \geq \frac{c}{\sqrt{d}}\Big] \geq \frac{1}{4} \cdot \frac{\min(\eta\tau, 1)}{4} =: p_0$.

Therefore the terms in $h(\theta)$ are independent Bernoulli random variables, each with mean at least p_0, and by a standard Chernoff bound (applied to a dominated sum of Bernoulli(p_0) variables),

$$\Pr\Big[h(\theta) < \frac{mp_0}{2}\Big] \leq \exp\Big(-\frac{mp_0}{8}\Big).$$

Using the same $\eta = c/(2\sqrt{d})$ as above, let N_η be a η-net of $\mathbb{S}^{d-1}$. Standard covering number bounds (see, e.g., Rudelson and Vershynin [2009]) give $|N_\eta| \leq \left(\frac{3}{\eta}\right)^d = \left(\frac{6\sqrt{d}}{c}\right)^d$.

A union bound over all $\theta' \in N_\eta$ yields $\Pr\big[\min_{\theta' \in N_\eta} h(\theta') < \frac{mp_0}{2}\big] \leq \left(\frac{6\sqrt{d}}{c}\right)^d \exp\big(-\frac{mp_0}{8}\big)$. This probability is at most n^{-3}, as long as we have $\frac{mp_0}{8} \geq d \ln\Big(\frac{6\sqrt{d}}{c}\Big) + 3 \ln n$.

By definition, $p_0 = \frac{\min(\eta\tau, 1)}{16}$. If $\eta\tau \geq 1$, then $p_0 = \frac{1}{16}$, and the inequality holds because $m \geq C_m d \log(nd)$ with C_m large enough. If $\eta\tau < 1$, then $\frac{mp_0}{8} = \frac{m\eta\tau}{128} \geq \frac{c\, C_0\, d \log(nd)}{256}$ by our lower bound on τ, which again suffices when C_0 is large enough (depending only on c). Thus in either case, the inequality above holds, and therefore the desired probability guarantee holds as well.

We can thus complete the net argument as follows: consider any $\theta \in \mathbb{S}^{d-1}$ (potentially not in the η-net), choose $\theta' \in N_\eta$ with $\|\theta - \theta'\| \leq \eta$. For any i such that $\langle x_i, \theta' \rangle \geq \frac{c}{\sqrt{d}}$, we have, using the fact that $\|x_i\| \leq 1$,

$$\langle x_i, \theta \rangle \geq \langle x_i, \theta' \rangle - |\langle x_i, \theta - \theta' \rangle| \geq \frac{c}{\sqrt{d}} - \eta = \frac{c}{2\sqrt{d}}.$$

This implies that with probability at least $1 - \frac{1}{n^3}$ over the choice of $\{x_i, \epsilon_i\}_{i \in S}$, we have:

$$\forall \theta \in \mathbb{S}^{d-1}, \quad \sum_{i \in S} Y_i \cdot \mathbf{1}\left[\langle x_i, \theta \rangle \geq \frac{c}{2\sqrt{d}}\right] \geq \frac{mp_0}{2} = \frac{m \cdot \min(\eta\tau, 1)}{32}. \tag{5}$$

Next, applying this with $\theta = \overline{\Delta}$, we will complete the proof of the lemma. Let us consider any $i \in S$ with $Y_i = 1$ and $\langle x_i, \overline{\Delta} \rangle \geq \frac{c}{2\sqrt{d}}$. For such an i, we have

$$0 < \epsilon_i < \eta\tau \leq \tau \langle x_i, \overline{\Delta} \rangle,$$

by using the value of η. Thus if we denote $a_i = \langle x_i, \overline{\Delta} \rangle$, then ϵ_i lies strictly between 0 and τa_i. Thus we can use Lemma 3.5 to conclude that the ith term contributes at least $2|a_i| \geq \frac{c}{\sqrt{d}}$ to the derivative gap $\psi'_-(\tau) - \psi'_+(0)$. Therefore,

$$\psi'_-(\tau) - \psi'_+(0) \geq \frac{mp_0}{2} \cdot \frac{c}{\sqrt{d}}.$$

Plugging in the value of p_0, this gives $\psi'_-(\tau) - \psi'_+(0) \geq \frac{cm}{32\sqrt{d}} \min\left(1, \frac{c\tau}{2\sqrt{d}}\right)$ for the chosen (fixed) value of τ.

As the final step, we handle the fact that $\tau = \|\Delta\|$ is itself random. Let $\tau_{\max} := \frac{2\sqrt{d}}{c}$ (the value at which $\eta\tau = 1$), and consider the geometric grid $\tau_j := 2^j \cdot C_0 \frac{d^{3/2}\log(nd)}{m}$ for $j = 0, 1, \ldots, J$, where J is the smallest index with $\tau_J \geq \tau_{\max}$. Since $m \leq n$, we have $J = O(\log n)$. We apply the argument above to every τ_j and take a union bound, so the failure probability is at most $(J+1)n^{-3} < n^{-2}$ (for n larger than an absolute constant). Now consider the actual value $\tau \geq \tau_0$, and let j be the largest index with $\tau_j \leq \tau$. Every $i \in S$ counted in (5) (with τ_j in place of τ, and $\theta = \overline{\Delta}$) satisfies $0 < \epsilon_i < \eta\tau_j \leq \eta\tau \leq \tau\langle x_i, \overline{\Delta} \rangle$, so the argument above gives

$$\psi'_-(\tau) - \psi'_+(0) \geq \frac{cm}{32\sqrt{d}} \min\left(1, \frac{c\tau_j}{2\sqrt{d}}\right).$$

If $j < J$, then $\tau_j > \tau/2$; if $j = J$, then $\tau_j \geq \tau_{\max}$ and the minimum equals 1. In either case, the right-hand side is at least $\frac{cm}{64\sqrt{d}} \min\left(1, \frac{c\tau}{2\sqrt{d}}\right)$, which proves the lemma. □

We are now ready to prove our main result, Theorem 3.2. We define:

$$L := C_0 \frac{d^{3/2}(\log d + \log n)}{m}, \qquad U := C_1 \frac{d\sqrt{n \log n}}{m},$$

where C_0 is the absolute constant from Lemma 3.6 and C_1 is the absolute constant from Lemma 3.4. Now if $\tau \leq L$, the conclusion of the Theorem already follows by the choice of our parameters. If not, we use the conclusions of Lemmas 3.3, 3.4 to conclude an upper bound of $O(\sqrt{n \log n})$ on $\psi'_-(\tau) - \psi'_+(0)$, with high probability. Then by using Lemma 3.6, we can derive an upper bound on τ, thus completing the proof. The detailed proof is deferred to Appendix B.2.

4 Iterative Truncation on Residuals

In this section, we present a novel variant of the *iterative truncation* algorithm used in Liang and Yuan [2020] to apply to regression. The main result is the following:

Theorem 4.1 (Iterative truncation algorithm)**.** *Consider the subset-of-signals model wherein there exists an unknown subset $S \subseteq [n]$ with $|S| \geq m$ such that $\sigma_i \leq 1$ for all $i \in S$. Assume that the x_i satisfy $\|x_i\| \leq 1$ for all $i \in [n]$, and additionally, suppose that the covariances satisfy*

$$\Sigma_S := \frac{1}{m}\sum_{i \in S} x_i x_i^\top \succeq c_1 \Sigma_n, \qquad \Sigma_n := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \succ 0, \qquad \frac{1}{m}\sum_{i \in S} \langle x_i, u \rangle^4 \leq C \left(\frac{1}{n}\sum_{i \in [n]} \langle x_i, u \rangle^2\right)^2$$

for some absolute constants $c_1 \in (0, 1]$ and $C > 0$. Finally, suppose we are given an initial estimate β_0 that satisfies $2\|\beta_0 - \beta^*\|_{\Sigma_n} \leq B$, for some known parameter $B \geq 1$ (this is without loss of generality, since B can always be increased).

Then consider the output $\widehat{\beta}$ of Algorithm 1 with the setting $\delta^{(0)} := B$, $K := \lceil \log_2 B \rceil$, and $T := \left\lceil \frac{\sqrt{n} \ln n}{\kappa} \right\rceil$, where $\kappa := \frac{c_1}{40} \min\left(1, \sqrt{\frac{c_1}{2C}}\right)$ (any larger value of T also works). As long as $m \geq$

$C_m\sqrt{dn\ln n}$ for a sufficiently large constant C_m depending only on c_1 and C, with probability at least $1 - O\big(\frac{KT}{n^3}\big)$ over the noise $\{\epsilon_i\}$,

$$\|\widehat{\beta} - \beta^*\|_{\Sigma_n} \le C' \frac{\sqrt{dn\ln n}}{m} = \widetilde{O}\left(\frac{\sqrt{dn}}{m}\right),$$

for a constant $C' > 0$ depending only on c_1 and C. Furthermore if $\lambda_{\min}(\Sigma_n)$ denotes the smallest eigenvalue of Σ_n,

$$\|\widehat{\beta} - \beta^*\|_2 \le \widetilde{O}\left(\frac{1}{(\lambda_{\min}(\Sigma_n))^{1/2}} \cdot \frac{\sqrt{dn\ln n}}{m}\right).$$

Remarks. As discussed in the introduction, this result is in some ways weaker than Theorem 3.2: first, we require initial estimates β_0, B. While very loose bounds also suffice, this is still an extra assumption. Second, we need the matrix Σ_n to have a well-behaved spectrum (especially for the ℓ_2 error bound). This said, the main strength of this result is that the conditions on $\{x_i\}$ are purely deterministic (given by the spectral constraints on Σ_n and Σ_S), and there is no distributional assumption on $\mathcal{D}_i$ as in Theorem 3.2. We also remark that the condition $\Sigma_S \succeq c_1\Sigma_n$ can be viewed as a deterministic analog of the well-spread assumption (Definition 3.1). Finally, we note that if the $\{x_i\}$ are indeed drawn from a well-spread distribution (such as the uniform distribution over the sphere, or a re-normalized multi-variate normal distribution), we expect Σ_n to have all of the eigenvalues close to $\frac{1}{d}$; specifically, they will be in the interval $[\frac{1}{2d}, \frac{2}{d}]$ with high probability. In this case, the bound on $\|\widehat{\beta} - \beta^*\|_2$ matches bound from Theorem 3.2 up to constants.

The Algorithm. We now discuss how to adapt the iterative algorithm of Liang and Yuan [2020] to the regression setting. In Liang and Yuan [2020], we are given observations $u_1, u_2, \ldots, u_n$, where $u_i \sim \mathcal{N}(\mu^*, \sigma_i^2)$ and the goal is to estimate μ^*. The algorithm is iterative; at each step it maintains a guess μ and an interval length δ. For each u_i, they consider the projection $\tilde{u}_i$ onto the interval $I := [\mu - \delta, \mu + \delta]$.[2] The next guess for μ^* then becomes the average of $\tilde{u}_i$.

As such, the algorithm above is tailored to mean estimation and relies heavily on intervals in $\mathbb{R}$, so it is not clear how to adapt it for estimating a vector β^* in regression (where our input consists of (x_i, y_i) pairs). Our algorithm is as follows: we have a current guess β and an interval length δ as before. The key idea is to pass to the "residuals", $r_i = y_i - \langle\beta, x_i\rangle$, obtained from the current guess β. Then we project these residuals onto the interval $[-\delta, \delta]$, obtaining $\tilde{r}_i$. Using this, we obtain the "corrected" versions of y_i, that we call $\tilde{y}_i$, defined as $\tilde{r}_i + \langle\beta, x_i\rangle$. Using these $\tilde{y}_i$, we use ordinary least squares to obtain the new guess β. (One can verify that if we are in 1D and x_i are all equal to 1, this procedure is identical to the one above.) As in Liang and Yuan [2020], we need to perform this iterative update carefully in multiple phases. Interestingly, we show that we can largely mimic their analysis, but with some important technical additions. For the details, please see Algorithm 1.

Proof Overview. We now outline the proof of Theorem 4.1. The idea is to keep track of $e_t^{(k)}$, the error vector in the tth step of the kth stage, i.e, $e_t^{(k)} := \beta_t^{(k)} - \beta^*$, and to measure the reduction in the quantity $\|e_t^{(k)}\|_{\Sigma_n}$. The goal will be to show that this norm contracts over stages, as well as over the iterations in a given stage, similar to the mean estimation analysis from Liang and Yuan [2020]. We define the auxiliary functions (for $\beta \in \mathbb{R}^d$ and $\delta > 0$):

$$\begin{aligned} r_i(\beta) &:= y_i - x_i^\top \beta \\ z_i(\beta, \delta) &:= x_i^\top(\beta - \beta^*) + \phi(r_i(\beta); [-\delta, \delta]) \\ \widetilde{z}_i(\beta, \delta) &:= z_i(\beta, \delta) - \mathbb{E} z_i(\beta, \delta) \end{aligned}$$

One can then decompose $\|e_t^{(k)}\|_{\Sigma_n}$ in terms of the *deviation* $\widetilde{z}_i$ and the *expected* value $\mathbb{E}[z_i]$. For the former, we run into a dependency issue: the current iterate $\beta_t^{(k)}$ is a function of the ϵ_i, so we cannot use concentration inequalities that leverage randomness in ϵ_i. To overcome this, we use a *uniform* bound over *all vectors* β in an appropriate net. (See Lemma C.2.) The expectation term, as it turns out, can be bounded following the recipe from Liang and Yuan [2020], see Lemma C.3. The lemma shows that when $i \in S$, we can expect to obtain a contraction in the corresponding term. Using this,

[2] Recall the definition of the projection function $\phi(x; [a, b])$ from Section 2.

Algorithm 1: Iterative Truncation for large d

Input: Input-output pairs $(x_i, y_i)_{i\in[n]}$ such that $y_i = x_i^\top \beta^* + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, initial estimate β_0 of β^*, parameter $B \geq 1$ such that $2\left\|\beta^* - \beta_0\right\|_{\Sigma_n} \leq B$, parameter m such that $\sigma_i \leq 1$ for at least m indices i.

Set initial estimate $\beta_0^{(0)} \leftarrow \beta_0$;

Set $\delta^{(0)} \leftarrow B$, $K \leftarrow \lceil \log_2 B \rceil$ and $T \leftarrow \lceil \sqrt{n} \ln n/\kappa \rceil$, with κ as in Theorem 4.1;

for *stage* $k \in [0, K]$ **do**

  for *step* $t \in \{0, 1, \ldots, T-1\}$ **do**

    Compute residuals $r_{i,t} := y_i - x_i^\top \beta_t^{(k)}$;

    Truncate residuals $\tilde{r}_{i,t} := \phi(r_{i,t}; [-\delta^{(k)}, \delta^{(k)}])$;

    Compute truncated responses $\tilde{y}_{i,t}^{(k)} := x_i^\top \beta_t^{(k)} + \tilde{r}_{i,t}$;

    Compute next estimate via OLS over the truncated responses:

$$\beta_{t+1}^{(k)} := \text{argmin}_\beta \sum_{i=1}^{n} (\tilde{y}_{i,t}^{(k)} - x_i^\top \beta)^2 = (X^\top X)^{-1} X^\top \tilde{Y}_t^{(k)}$$

    where above $\tilde{Y}_t^{(k)} \in \mathbb{R}^{n\times 1}$ is the vector of all truncated responses, and $X \in \mathbb{R}^{n\times d}$ is the design matrix;

  Set estimate for next stage: $\beta_0^{(k+1)} \leftarrow \beta_T^{(k)}$;

  Halve the truncation width for next stage: $\delta^{(k+1)} \leftarrow \frac{\delta^{(k)}}{2}$;

Output: $\beta_T^{(K)}$

together with our assumption that $\Sigma_S \succeq c_1 \Sigma_n$ for some constant c_1 and our fourth-moment bound, we can argue that the error undergoes a geometric decay (Lemma C.4).

Putting these results together, we can show that the error drops by a constant in every stage (i.e., outer loop of the algorithm), with high probability. Furthermore, in the last stage (by which time $\delta^{(k)}$ reaches 1), we can prove that the error magnitude is strictly constrained, yielding our target error bound of $\widetilde{O}(\frac{\sqrt{dn \ln n}}{m})$.

The full analysis of the algorithm is presented in Appendix C.

Conclusion and Future Directions

We present two algorithms for the problem of linear regression under heteroskedastic noise. Both the algorithms are analyzed in the so-called subset-of-signals model. Our algorithms provide novel theoretical guarantees for this fundamental problem. In our first result, the algorithm is well-known, but our analysis is the key novelty. In the latter result, the algorithm requires a careful adaptation of previously known ideas.

An important future direction of this work is to study whether similar algorithmic techniques can be used to solve more general learning problems under heteroskedastic errors. An example of this could be structured estimation problems where the unknown parameter lies in a low dimensional subspace of an ambient space, e.g. sparse linear regression and low-rank matrix sensing. In these settings, a natural goal is to design efficient algorithms whose recovery guarantees scale with the dimension of the low-dimensional subspace, instead of the dimension of the ambient space. Finally, it will be interesting to see if a more general theory of learning under heteroskedastic noise can be developed using these algorithmic techniques.

References

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980. URL https://onlinelibrary.wiley.com/doi/abs/0012-9682(198005)48:4<817:AHCMEA>2.0.CO;2-K.

Dimitris N. Politis and Stefanos Poulis. Heteroskedastic linear regression: Steps towards adaptivity, efficiency, and robustness. In Michael G. Akritas, S. N. Lahiri, and Dimitris N. Politis, editors, *Topics in Nonparametric Statistics*, pages 283–297, New York, NY, 2014. Springer New York.

Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, page 167–176, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302616. doi: 10.1145/1993574.1993599. URL https://doi.org/10.1145/1993574.1993599.

Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. Learning entangled single-sample gaussians. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, page 511–522, USA, 2014. Society for Industrial and Applied Mathematics. ISBN 9781611973389.

Marie Davidian and Raymond J Carroll. Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091, 1987. doi: 10.2307/2289384.

Aniket Das, Dheeraj M. Nagaraj, Praneeth Netrapalli, and Dheeraj Baby. Near optimal heteroscedastic regression with symbiotic learning. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3696–3757. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/das23a.html.

A. C. Harvey. Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, 44(3):461–465, 1976. URL https://onlinelibrary.wiley.com/doi/abs/0012-9682(197605)44:3<461:ERMWMH>2.0.CO;2-B.

P. Xu and Seiichi Shimada. Least squares parameter estimation in multiplicative noise models. *Communications in Statistics Part B: Simulation and Computation*, 29:83–96, 02 2000.

Herman Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16: 31–41, 1964. URL https://api.semanticscholar.org/CorpusID:121030566.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Estimating location parameters in entangled single-sample distributions, 2019. URL https://arxiv.org/abs/1907.03087.

Yingyu Liang and Hui Yuan. Learning entangled single-sample gaussians in the subset-of-signals model. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2712–2737. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/liang20b.html.

Spencer Compton and Gregory Valiant. Near-optimal mean estimation with unknown, heteroskedastic variances. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 194–200, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi: 10.1145/3618260.3649754. URL https://doi.org/10.1145/3618260.3649754.

Ilias Diakonikolas, Daniel M. Kane, Sihan Liu, and Thanasis Pittas. Entangled mean estimation in high dimensions. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 1680–1688, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718162. URL https://doi.org/10.1145/3717823.3718162.

Yanjun Han, Abhishek Shetty, and Jacob Shkrob. An empirical Bayes perspective on heteroskedastic mean estimation. In Steve Hanneke and Tor Lattimore, editors, *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3076–3108. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/han26a.html.

Tommaso d'Orsi, Gleb Novikov, and David Steurer. Consistent regression when oblivious outliers overwhelm, 2021. URL https://arxiv.org/abs/2009.14774.

Ilias Diakonikolas, Chao Gao, Daniel Kane, John Lafferty, and Ankit Pensia. Information-computation tradeoffs for noiseless linear regression with oblivious contamination. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=XqHrG8lBai.

Spencer Compton and Tselil Schramm. Algorithms for adaptive and heteroskedastic linear regression at the computational threshold, 2026. URL https://arxiv.org/abs/2608.18402.

David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991. doi: 10.1017/S0266466600004394.

Jan van den Brand, Yin Tat Lee, Yang P. Liu, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Minimum cost flows, mdps, and l1-regression in nearly linear time for dense instances. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 859–869, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451108. URL https://doi.org/10.1145/3406325.3451108.

Hui Yuan and Yingyu Liang. Learning entangled single-sample distributions via iterative trimming. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2666–2676. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/yuan20a.html.

Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 938–942, 2019. doi: 10.1145/3313276.3316303.

David Durfee, Kevin A. Lai, and Saurabh Sawlani. ℓ_1 regression using lewis weights preconditioning and stochastic gradient descent. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1626–1656. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/durfee18a.html.

Iosif Pinelis and Alexander I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability and Its Applications*, 30:143–148, 1986. URL https://api.semanticscholar.org/CorpusID:122952302.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009. doi: https://doi.org/10.1002/cpa.20294. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20294.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1008, 1982. URL https://onlinelibrary.wiley.com/doi/abs/0012-9682(198207)50:4<987:ACHWEO>2.0.CO;2-3.

Anthony C. Atkinson, Marco Riani, and Francesca Torti. Robust methods for heteroskedastic regression. *Computational Statistics & Data Analysis*, 104:209–222, 2016. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2016.07.002. URL https://www.sciencedirect.com/science/article/pii/S0167947316301578.

Conceição Amado, Ana M. Bianco, Graciela Boente, and Isabel M. Rodrigues. Robust estimation of heteroscedastic regression models: a brief overview and new proposals, 2023. URL https://arxiv.org/abs/2311.02822.

Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962. URL https://api.semanticscholar.org/CorpusID:122932724.

Faouzi El Bantli and Marc Hallin. L1-estimation in linear models with heterogeneous white noise. *Statistics & Probability Letters*, 45(4):305–315, 1999. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(99)00072-3. URL https://www.sciencedirect.com/science/article/pii/S0167715299000723.

A Additional Related Work

Linear regression with heteroskedastic errors has been extensively studied empirically, over several decades [White, 1980, Engle, 1982, Politis and Poulis, 2014, Atkinson et al., 2016, Amado et al., 2023]. Liang and Yuan [2020] consider our same question, but they cannot obtain sub-constant error guarantees in subset-of-signals model, even for $m = n/2$.

A closely related problem is that of regression with *outliers*, where some of the errors can be arbitrarily large. The main difference between these two lines of work is that the number of outliers is often viewed as small relative to the total number of observations (whereas in our setting we think of m as much smaller than n). With some notable exceptions, much of the recent algorithmic progress in robust statistics focuses on the case where a small fraction of the input is noisy (e.g., see the recent book of Diakonikolas and Kane [2023]).

Within machine learning theory, Chierichetti et al. [2014] began the study of the simpler mean estimation problem under heteroskedastic errors, building on classical ideas, e.g., Chernoff [1964], Parzen [1962], Bantli and Hallin [1999]. Several works, including those of Pensia et al. [2019], Liang and Yuan [2020], and finally Compton and Valiant [2024], developed a nearly tight understanding of mean estimation, and introduced the subset-of-signals model, that we also study. The work of Diakonikolas et al. [2025a] extends these results to high dimensions using new ideas. Most recently, Han et al. [2026] developed an empirical Bayes approach for the mean estimation problem, and achieved minimax error bounds.

Prior work and direct comparisons. Pensia et al. [2019] proposed a *modal interval regression estimator* (see Algorithm 13 of their paper) that achieves the error bound $\left\|\hat{\beta} - \beta^*\right\|_2 \leq \frac{c' n \sigma_{(cd \log n)}}{\lambda_{\min}}$. Here, c' is a positive constant, and the covariates x_i are assumed to be normally distributed with $x_i \sim N(\mu_i', \Sigma_i')$ and $\lambda_{\min} := \min_i \lambda_{\min}(\Sigma_i')$. Notably, their algorithm runs in time $O(n^d)$ and their error bound *increases* with the number of samples n. In contrast, our algorithms are efficient, work under more general assumptions on the covariates x_i, and achieve a sub-constant error bound. Yuan and Liang [2020] also propose a regression algorithm based on *iterative trimming* (see Section 4 of their paper). They make similar assumptions as to ours on the distributions of the covariates, and their algorithm achieves the error bound $\left\|\hat{\beta} - \beta^*\right\|_2 \leq c c_1 \sigma_{\lceil \alpha n \rceil}$, where c, c_1 and α are constants. While their algorithm has a polynomial running time complexity, the error bound is still not sub-constant. Most recently, after the time of writing this paper, Compton and Schramm [2026] generalized their balance-finding estimator from the mean estimation setting to linear regression, and introduced a residual-balance based algorithm that achieves the minimax error bound for heteroskedastic regression. Their algorithm achieves an error bound of the order $\tilde{O}\left(\sqrt{d}\left(\frac{n}{m^4}\right)^{1/6}\right)$. However, their algorithm requires distributional assumptions on *all* covariates x_i; in contrast, our results only put such assumptions on the covariates x_i for the "good" indices $i \in S$.

B Technical Lemmas for Section 3

B.1 Proof of Lemma 3.7

Proof. Let $Z = Y/\sigma \sim \mathcal{N}(0, 1)$ and set $a = t/\sigma$. Then

$$\Pr(0 < Y < t) = \Pr(0 < Z < a) = \int_0^a \phi(x)dx,$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the density function of the standard normal. Since ϕ is decreasing on $[0, \infty)$, the "average" function

$$A(a) := \frac{1}{a} \int_0^a \phi(x)dx$$

is decreasing in $a > 0$. Hence, for $0 < a \leq 1$,

$$A(a) \geq A(1) = \int_0^1 \phi(x)dx = \Phi(1) - \tfrac{1}{2} \approx 0.84 - 0.5 > \tfrac{1}{4}.$$

Therefore, if $t \le \sigma$ (i.e., $a \le 1$),

$$\Pr(0 < Y < t) = aA(a) > \frac{a}{4} = \frac{t}{4\sigma}.$$

If $t \ge \sigma$ (i.e., $a \ge 1$), then

$$\Pr(0 < Y < t) \ge \Pr(0 < Z < 1) = \Phi(1) - \tfrac{1}{2} > \tfrac{1}{4}.$$

Combining the two cases gives $\Pr(0 < Y < t) \ge \frac{\min(t,\sigma)}{4\sigma}$, as claimed. □

B.2 Proof of Theorem 3.2

We now give the full proof of our first main result, Theorem 3.2.

Proof. Let

$$L := C_0 \frac{d^{3/2}(\log d + \log n)}{m}, \qquad U := C_1 \frac{d\sqrt{n \log n}}{m},$$

where C_0 is the absolute constant from Lemma 3.6 and C_1 is the absolute constant from Lemma 3.4.

If $\tau \le L$, then we are already done: since $n \ge Cd \log n$ implies $d \le n$, we have $\log(nd) \le 2 \log n$ and $\sqrt{d} \log n \le \sqrt{n \log n / C}$, so that $L \le \frac{2C_0}{\sqrt{C}} \cdot \frac{d\sqrt{n \log n}}{m}$. Moreover, the hypotheses imply $m \ge C\sqrt{nd \log n} \ge C^{3/2} d \log n \ge \frac{C^{3/2}}{2} d \log(nd)$, so for C large enough the condition $m \ge C_m d \log(nd)$ of Lemma 3.6 holds. Thus, we may assume that $\tau > L$, in which case Lemma 3.6 applies.

Now since the conclusion of Lemma 3.3 always holds, and the conclusion of Lemmas 3.4, and 3.6 both hold with probability $> 1 - \frac{1}{n^2}$, all the conclusions hold with probability $\ge 1 - \frac{2}{n^2}$. Using Lemmas 3.3 and 3.6, $\psi'_-(\tau) \le 0$ and the gap $\psi'_-(\tau) - \psi'_+(0)$ is large enough. This implies that

$$\psi'_+(0) \le -\frac{cm}{64\sqrt{d}} \min\left(1, \frac{c\tau}{2\sqrt{d}}\right).$$

However, by Lemma 3.4,

$$\left|\psi'_+(0)\right| \le C_1 \sqrt{n \log n}.$$

Combining these two inequalities, we have

$$\frac{cm}{64\sqrt{d}} \min\left(1, \frac{c\tau}{2\sqrt{d}}\right) \le C_1 \sqrt{n \log n}.$$

Now if $\frac{c\tau}{2\sqrt{d}} > 1$, then the 1 term is the result of the $\min()$ above, and we get $m \le \frac{64C_1}{c}\sqrt{nd \log n}$, which contradicts our hypothesis $m \ge C\sqrt{nd \log n}$ as long as $C > \frac{64C_1}{c}$. Thus, we must have

$$\frac{c^2 \tau m}{128d} \le C_1 \sqrt{n \log n} \implies \tau \le \frac{128 C_1}{c^2} \cdot \frac{d\sqrt{n \log n}}{m} = O\left(\frac{d\sqrt{n \log n}}{m}\right).$$

Thus, we have that $\|\hat{\beta} - \beta^*\| = \tau \le \widetilde{O}\left(\frac{d\sqrt{n}}{m}\right)$, thus completing the proof of the theorem. □

C Analysis of the Iterative Truncation Algorithm

We now show how to extend the analysis from the mean estimation setting to the regression setting. Throughout our analysis, we condition on the design matrix X (i.e we treat X as a fixed matrix). Now, fix a stage $k \in \{0, 1, \ldots, K\}$. Let $e_t^{(k)}$ be the error vector in the tth step of the kth stage, i.e we define

$$e_t^{(k)} := \beta_t^{(k)} - \beta^*$$

Our goal will be to show that the norm of the error $e_t^{(k)}$ contracts over stages (similar to the proof from Liang and Yuan [2020] for mean estimation). We start by defining the following quantities for any fixed $\beta \in \mathbb{R}^d$ and $\delta > 0$:

$$\begin{aligned} r_i(\beta) &:= y_i - x_i^\top \beta \\ z_i(\beta, \delta) &:= x_i^\top (\beta - \beta^*) + \phi(r_i(\beta); [-\delta, \delta]) \\ \widetilde{z}_i(\beta, \delta) &:= z_i(\beta, \delta) - \mathbb{E} z_i(\beta, \delta) \\ \Sigma_n &:= \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \end{aligned}$$

Above, in the definition of $\widetilde{z}_i$, we note that the expectation $\mathbb{E} z_i$ is taken w.r.t the noise ϵ_i. We first prove a simple Lipschitz property of the z_is:

Lemma C.1 ($\widetilde{z}_i$ is Lipschitz as a function of β)**.** *Let $\delta > 0$ be fixed, and let $\beta, \beta' \in \mathbb{R}^d$. Then, for any i,*

$$|\widetilde{z}_i(\beta, \delta) - \widetilde{z}_i(\beta', \delta)| \leq 4 \left| x_i^\top (\beta - \beta') \right|$$

Proof. From the definition of $z_i(\beta, \delta)$, we have

$$\begin{aligned} |z_i(\beta, \delta) - z_i(\beta', \delta)| &= \left| x_i^\top (\beta - \beta^*) + \phi(r_i(\beta); [-\delta, \delta]) - x_i^\top (\beta' - \beta^*) - \phi(r_i(\beta'); [-\delta, \delta]) \right| \\ &= \left| x_i^\top (\beta - \beta') + \phi(r_i(\beta); [-\delta, \delta]) - \phi(r_i(\beta'); [-\delta, \delta]) \right| \end{aligned}$$

Now, from the properties of the truncation function and the definition of r_i, we have the following:

$$\begin{aligned} \phi(r_i(\beta); [-\delta, \delta]) &= \phi(y_i - x_i^\top \beta; [-\delta, \delta]) \\ &= \phi(\epsilon_i + x_i^\top (\beta^* - \beta); [-\delta, \delta]) \end{aligned}$$

A similar property holds for β' in place of β as well. It is then easily seen that

$$\begin{aligned} &\left| \phi(\epsilon_i + x_i^\top (\beta^* - \beta); [-\delta, +\delta]) - \phi(\epsilon_i + x_i^\top (\beta^* - \beta'); [-\delta, +\delta]) \right| \\ &\leq \left| x_i^\top (\beta^* - \beta) - x_i^\top (\beta^* - \beta') \right| \\ &= \left| x_i^\top (\beta - \beta') \right| \end{aligned}$$

So, combining everything, we see that

$$|z_i(\beta, \delta) - z_i(\beta', \delta)| \leq 2 \left| x_i^\top (\beta - \beta') \right|$$

Taking expectations w.r.t ϵ_is, we see that $\mathbb{E}[z_i(\cdot, \delta)]$ satisfies the same inequality. Hence, from the definition of $\widetilde{z}_i$, the claim follows. □

Now, we write the error $e_{t+1}^{(k)}$ in terms of the z_is as follows:

$$\begin{aligned} e_{t+1}^{(k)} &= \beta_{t+1}^{(k)} - \beta^* \\ &\overset{(a)}{=} (X^\top X)^{-1} X^\top \tilde{Y}_t^{(k)} - \beta^* \\ &\overset{(b)}{=} \Sigma_n^{-1} \frac{1}{n} X^\top \tilde{Y}_t^{(k)} - \Sigma_n^{-1} \Sigma_n \beta^* \\ &= \Sigma_n^{-1} \left(\frac{1}{n} X^\top \tilde{Y}_t^{(k)} - \Sigma_n \beta^* \right) \\ &\overset{(c)}{=} \Sigma_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \tilde{y}_{i,t}^{(k)} - \frac{1}{n} \sum_{i=1}^n x_i (x_i^\top \beta^*) \right) \\ &\overset{(d)}{=} \Sigma_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i z_i \left(\beta_t^{(k)}, \delta^{(k)} \right) \right) \end{aligned}$$

where above in (a), we have expanded the definition of $\beta_{t+1}^{(k)}$; in (b) we have used the definition of Σ_n; in (c) we have written the matrix products $X^\top \tilde{Y}_t^{(k)}$ and $\Sigma_n \beta^*$ in terms of the column vectors of $X^\top$, and in (d) we are using the definition of $z_i(\beta, \delta)$ with substitutions $\beta = \beta_t^{(k)}$ and $\delta = \delta^{(k)}$.

This enables us to track the norm of the error over time steps and stages. Recall that $\|v\|_{\Sigma_n} := \sqrt{v^\top \Sigma_n v} = \left\|\Sigma_n^{1/2} v\right\|_2$. Using the fact that $z_i = \widetilde{z}_i + \mathbb{E}z_i$, we can bound the error norm as follows:

$$\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n} \leq \left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i \widetilde{z}_i(\beta_t^{(k)}, \delta^{(k)})\right\|_{\Sigma_n} + \left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i \mathbb{E}z_i(\beta_t^{(k)}, \delta^{(k)})\right\|_{\Sigma_n} \tag{6}$$

We now bound each of the two terms on the RHS above separately. To bound the first term above, we use a standard ϵ-net argument along with sub-Gaussian bounds, where the net is going to be over a ball $\mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$ of fixed (deterministic) radius δ_e:

Lemma C.2 (Bounding the first term of (6))**.** *Let $\delta > 0$ and $\delta_e > 0$ be fixed (deterministic) parameters, and let $n \geq 9$. There is an absolute constant $C_3 > 0$ such that, with probability at least $1 - \frac{1}{n^3}$ (over the ϵ_is), for any β with $\|\beta - \beta^*\|_{\Sigma_n} \leq \delta_e$, we have*

$$\left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i \widetilde{z}_i(\beta, \delta)\right\|_{\Sigma_n} \leq C_3 \delta \sqrt{\frac{d \ln n}{n}} + \frac{\delta_e}{n^2}$$

Proof. First, let us fix some $\beta \in \mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$; we will first bound the norm of the corresponding vector for this β, and then apply an ϵ-net argument over the ball $\mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$ to prove the lemma. Observe that

$$\left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i \widetilde{z}_i(\beta, \delta)\right\|_{\Sigma_n} = \left\|\Sigma_n^{-\frac{1}{2}}\frac{1}{n}\sum_{i=1}^n x_i \widetilde{z}_i(\beta, \delta)\right\|_2$$

Let us define

$$v(\beta, \delta) := \Sigma_n^{-\frac{1}{2}}\frac{1}{n}\sum_{i=1}^n x_i \widetilde{z}_i(\beta, \delta).$$

Now, from the definition of $z_i(\beta, \delta)$ and $\mathbb{E}z_i(\beta, \delta)$, both these quantities are in the interval $[x_i^\top(\beta - \beta^*) - \delta, x_i^\top(\beta - \beta^*) + \delta]$. In particular, we see that $|\widetilde{z}_i(\beta, \delta)| \leq 2\delta$. Thus, if we define

$$a_i := \frac{1}{n}\Sigma_n^{-\frac{1}{2}} x_i \tilde{z}_i(\beta, \delta),$$

then $\{a_i\}_{i=1}^n$ are independent, mean-zero vectors, and satisfy

$$\|a_i\|_2 \leq \frac{2\delta}{n}\|\Sigma_n^{-\frac{1}{2}} x_i\|_2.$$

We now apply a Hilbert-space (vector-valued) Hoeffding inequality (e.g., Pinelis and Sakhanenko [1986]) to the sum of these vectors: for all $s > 0$,

$$\Pr(\|v(\beta, \delta)\|_2 \geq s) \leq 2\exp\left(-\frac{c'' s^2}{\sum_{i=1}^n \|a_i\|_2^2}\right)$$

for an absolute constant $c'' > 0$. Next, let us evaluate the denominator:

$$\sum_{i=1}^n \|a_i\|_2^2 \leq \frac{4\delta^2}{n^2}\sum_{i=1}^n \|\Sigma_n^{-\frac{1}{2}} x_i\|_2^2.$$

Using the fact that $\|u\|_2^2 = \mathrm{tr}(uu^\top)$ for any vector u, and that $\Sigma_n^{-\frac{1}{2}}$ is symmetric, we can write

$$\sum_{i=1}^n \left\|\Sigma_n^{-\frac{1}{2}} x_i\right\|_2^2 = \mathrm{tr}\left(\Sigma_n^{-\frac{1}{2}}\left(\sum_{i=1}^n x_i x_i^\top\right)\Sigma_n^{-\frac{1}{2}}\right) = \mathrm{tr}(nI_d) = nd \tag{7}$$

Thus,

$$\Pr(\|v(\beta, \delta)\|_2 \geq s) \leq 2\exp\left(-\frac{cn^2 s^2}{\delta^2 nd}\right) = 2\exp\left(-\frac{cns^2}{\delta^2 d}\right).$$

for an absolute constant $c = c''/4$. We now choose

$$s = C_3\delta\sqrt{\frac{d\ln n}{n}}, \qquad C_3 := \sqrt{10/c}.$$

This ensures that, for each fixed β, the failure probability is at most $2\exp(-10\,d\ln n)$.

Thus, we have proven a high-probability norm bound for a fixed $\beta \in \mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$. We will now do an ϵ-net argument over $\mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$. Let us choose $\epsilon = \frac{\delta_e}{4dn^2}$. Let $\mathcal{N}_u$ be an ϵ-net for the ball $\mathcal{B}(0, \delta_e)$. In that case, the set $\mathcal{N}_{\text{param}} := \{\beta^* + \Sigma_n^{-\frac{1}{2}} u : u \in \mathcal{N}_u\}$ is an ϵ-net for the ball $\mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$ in the $\|\cdot\|_{\Sigma_n}$ norm. In particular, for any $\beta \in \mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$, there is a $\beta' \in \mathcal{N}_{\text{param}}$ with $\|\beta - \beta'\|_{\Sigma_n} \le \epsilon$. In that case, we have the following set of inequalities:

$$\begin{aligned}
\|v(\beta,\delta)\|_2 &\le \|v(\beta',\delta)\|_2 + \|v(\beta,\delta) - v(\beta',\delta)\|_2 \\
&\le \|v(\beta',\delta)\|_2 + \left\|\frac{1}{n}\sum_{i=1}^n \Sigma_n^{-\frac{1}{2}} x_i[\widetilde{z}_i(\beta,\delta) - \widetilde{z}_i(\beta',\delta)]\right\|_2 \\
&\le \|v(\beta',\delta)\|_2 + \frac{1}{n}\sum_{i=1}^n \left\|\Sigma_n^{-\frac{1}{2}} x_i\right\|_2 \cdot |\widetilde{z}_i(\beta,\delta) - \widetilde{z}_i(\beta',\delta)| \\
&\overset{(a)}{\le} \|v(\beta',\delta)\|_2 + \frac{4}{n}\sum_{i=1}^n \left\|\Sigma_n^{-\frac{1}{2}} x_i\right\|_2 \left|x_i^\top(\beta - \beta')\right|
\end{aligned}$$

where in (a) above, we have used Lemma C.1. Now, note that we can write

$$\begin{aligned}
\left|x_i^\top(\beta - \beta')\right| &= \left|(\Sigma_n^{-\frac{1}{2}} x_i)^\top \Sigma_n^{\frac{1}{2}}(\beta - \beta')\right| \\
&\le \left\|\Sigma_n^{-\frac{1}{2}} x_i\right\|_2 \|\beta - \beta'\|_{\Sigma_n}
\end{aligned}$$

where the above follows from Cauchy-Schwarz inequality, and the fact that $\left\|\Sigma_n^{\frac{1}{2}} u\right\|_2 = \|u\|_{\Sigma_n}$. So, combining the above with the previous inequality, we get

$$\begin{aligned}
\|v(\beta,\delta)\|_2 &\le \|v(\beta',\delta)\|_2 + \frac{4}{n}\sum_{i=1}^n \left\|\Sigma_n^{-\frac{1}{2}} x_i\right\|_2^2 \|\beta - \beta'\|_{\Sigma_n} \\
&\overset{(a)}{=} \|v(\beta',\delta)\|_2 + \frac{4}{n} \cdot nd\,\|\beta - \beta'\|_{\Sigma_n} \\
&\overset{(b)}{\le} \|v(\beta',\delta)\|_2 + \frac{\delta_e}{n^2}
\end{aligned}$$

where above in (a) we used (7) and in (b) we used our choice of ϵ. This finishes our ϵ-net argument over the ball $\mathcal{B}_{\Sigma_n}(\beta^*, \delta_e)$. Finally, from standard covering number bounds, we know that $|\mathcal{N}| \le \left(1 + \frac{2\delta_e}{\epsilon}\right)^d < (9dn^2)^d = \exp(d\ln 9dn^2)$. Thus, using the choice of s above, we can take a union bound over $\mathcal{N}$ to conclude that

$$\Pr[\exists \beta' \in \mathcal{N} \text{ such that } \|v(\beta',\delta)\|_2 \ge s] \le 2\exp\left(d\ln(9dn^2) - 10\,d\ln n\right) \le 2\exp(-6\,d\ln n) \le n^{-3},$$

where we used $d \le n$ (as $\Sigma_n \succ 0$) and $n \ge 9$, so that $\ln(9dn^2) \le 4\ln n$.

and the proof of the lemma is complete.

□

Next, we focus on bounding the second term of (6) and prove the following lemma:

Lemma C.3. *Let $\delta > 0$ be fixed, and let $e := \beta - \beta^*$. Then*

$$|\mathbb{E}[z_i(\beta,\delta)]| \le \left|x_i^\top e\right|(1 - \gamma_i)$$

where $\gamma_i := \frac{1}{5}\frac{\delta}{\max\left(\left|x_i^\top e\right|,\delta\right)}\frac{\delta}{\max(\sigma_i,\delta)}$. Moreover, we can write $\mathbb{E}[z_i(\beta,\delta)] = b_i x_i^\top e$ where $0 \le b_i \le 1 - \gamma_i$. In particular, $\mathbb{E}[z_i(\beta,\delta)]$ and $x_i^\top e$ share the same sign.

Proof. The proof closely follows the proof of Lemma 4 of Liang and Yuan [2020]. First define $e := \beta - \beta^*$. Then, note that

$$z_i(\beta, \delta) = x_i^\top e + \phi(\epsilon_i - x_i^\top e; [-\delta, \delta])$$

Then again, note that $\phi(\epsilon - x_i^\top e; [-\delta, \delta]) = \phi(\epsilon; [x_i^\top e - \delta, x_i^\top e + \delta]) - x_i^\top e$, and hence we see that

$$z_i(\beta, \delta) = \phi(\epsilon_i; [x_i^\top e - \delta, x_i^\top e + \delta])$$

At this point, we can apply Lemma 4 of Liang and Yuan [2020] to this z_i to get

$$|\mathbb{E}[z_i(\beta, \delta)]| \leq \left|x_i^\top e\right| \left(1 - \frac{1}{5} \frac{\delta}{\max(\left|x_i^\top e\right|, \delta)} \frac{\delta}{\max(\sigma_i, \delta)}\right)$$

proving our claim. We now focus on the second part of the claim, i.e proving that $\mathbb{E}[z_i(\beta, \delta)]$ and $x_i^\top e$ share the same sign. First, observe that

$$\mathbb{E}[z_i(\beta, \delta)] = \mathbb{E}\phi(\epsilon_i; [x_i^\top e - \delta, x_i^\top e + \delta])$$

Let $f_i(x)$ be the density of $\epsilon_i \sim N(0, \sigma_i^2)$. Then, we can write

$$\begin{aligned}
\mathbb{E}[z_i(\beta, \delta)] &= \int_{-\infty}^{\infty} \phi(t; [x_i^\top e - \delta, x_i^\top e + \delta]) f_i(t) dt \\
&\overset{(a)}{=} \int_{-\infty}^{0} \phi(t; [x_i^\top e - \delta, x_i^\top e + \delta]) f_i(t) dt + \int_{0}^{\infty} \phi(t; [x_i^\top e - \delta, x_i^\top e + \delta]) f_i(t) dt \\
&\overset{(b)}{=} \int_{0}^{\infty} \phi(-t; [x_i^\top e - \delta, x_i^\top e + \delta]) f_i(t) dt + \int_{0}^{\infty} \phi(t; [x_i^\top e - \delta, x_i^\top e + \delta]) f_i(t) dt \\
&= \int_{0}^{\infty} \left[\phi(-t; [x_i^\top e - \delta, x_i^\top e + \delta]) + \phi(t; [x_i^\top e - \delta, x_i^\top e + \delta])\right] f_i(t) dt
\end{aligned}$$

where in step (a) above we have simply split the integral into two intervals, and in step (b) we substituted $t \mapsto -t$ in the first integral and used the symmetry of the density function $f_i(t)$. Now, noting the fact that $\phi(x; [A, B]) = -\phi(-x; [-B, -A])$, we see that $\phi(-t; [x_i^\top e - \delta, x_i^\top e + \delta]) = -\phi(t; [-x_i^\top e - \delta, -x_i^\top e + \delta])$. Hence, the integral becomes

$$\mathbb{E}[z_i(\beta, \delta)] = \int_{0}^{\infty} \left[\phi(t; [x_i^\top e - \delta, x_i^\top e + \delta]) - \phi(t; [-x_i^\top e - \delta, -x_i^\top e + \delta])\right] f_i(t) dt$$

At this point, we deal with three cases:

1. In the first case, assume that $x_i^\top e > 0$. In this case, the interval $[x_i^\top e - \delta, x_i^\top e + \delta]$ is strictly shifted to the right compared to the interval $[-x_i^\top e - \delta, -x_i^\top e + \delta]$. Hence, the integrand is non-negative for all $t \geq 0$ (and strictly positive for all sufficiently large t), and as a result the integral is positive.
2. The case $x_i^\top e < 0$ is identical to the case above, and the integral in that case will be negative.
3. If $x_i^\top e = 0$, the integral is also 0.

Hence, $\mathbb{E}[z_i(\beta, \delta)]$ indeed shares the same sign as $x_i^\top e$. In particular, we can write $\mathbb{E}[z_i(\beta, \delta)] = b_i x_i^\top e$ for some $b_i \geq 0$. Moreover, from the first part of the lemma, we clearly have that $b_i \leq (1 - \gamma_i)$, completing the proof of the lemma. □

We can now combine the above two lemmas to get our final contractive bound on the error, similar to the mean estimation setting. Here, we will make use of the assumption on Σ_S in Theorem 4.1. Recall that $\Sigma_S := \frac{1}{m} \sum_{i \in S} x_i x_i^\top$ (i.e the sample covariance matrix of the "good" indices). We have assumed that $\Sigma_S \succeq c_1 \Sigma_n$ for some absolute constant $c_1 \in (0, 1]$, i.e., sample covariance matrix of the good indices "captures" the overall covariance matrix up to a constant. We have also assumed a fourth-moment condition on the covariates (see the statement of Theorem 4.1).

Lemma C.4. *Let $\delta := \delta^{(k)}$ and suppose $\delta \geq \left\|e_t^{(k)}\right\|_{\Sigma_n}$. Then, with probability at least $1 - \frac{1}{n^3}$, we have*

$$\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n} \leq C_2\,\delta\sqrt{\frac{d\ln n}{n}} + \left\|e_t^{(k)}\right\|_{\Sigma_n}\left(1 - \frac{\kappa m}{n}\right)$$

where $C_2 := C_3 + 1$ is an absolute constant, $\kappa := \frac{c_1}{40}\min\left(1, \sqrt{\frac{c_1}{2C}}\right)$. Particularly, we have $0 < \left(1 - \frac{\kappa m}{n}\right) < 1$.

Proof. Recall from equation (6) that we have

$$\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n} \leq \left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i\widetilde{z}_i(\beta_t^{(k)}, \delta^{(k)})\right\|_{\Sigma_n} + \left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i\mathbb{E}z_i(\beta_t^{(k)}, \delta^{(k)})\right\|_{\Sigma_n}$$

We have already bounded the first term in Lemma C.2, which we apply with the deterministic radius $\delta_e := \delta$ (valid since $\left\|e_t^{(k)}\right\|_{\Sigma_n} \leq \delta$); we have that with probability at least $1 - \frac{1}{n^3}$:

$$\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n} \leq C_3\delta\sqrt{\frac{d\ln n}{n}} + \frac{\delta}{n^2} + \left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^n x_i\mathbb{E}z_i(\beta_t^{(k)}, \delta^{(k)})\right\|_{\Sigma_n}$$

Since $\frac{1}{n^2} \leq \sqrt{\frac{d\ln n}{n}}$, the first two terms are at most $C_2\delta\sqrt{\frac{d\ln n}{n}}$ with $C_2 = C_3 + 1$. Next, we invoke Lemma C.3 to bound the last term of the RHS above. Particularly, from Lemma C.3, we know that

$$\frac{1}{n}\sum_{i=1}^n x_i\mathbb{E}z_i(\beta_t^{(k)}, \delta^{(k)}) = \frac{1}{n}\sum_{i=1}^n b_i x_i x_i^\top e_t^{(k)}$$

where $0 \leq b_i \leq 1 - \gamma_i$ for each i, with γ_i as defined in the statement of Lemma C.3. Now, by the definition of γ_i, since $\sigma_i \leq 1$ for $i \in S$, and since $\delta^{(k)} = B/2^k \geq B/2^K \geq \frac{1}{2}$ for all stages k (as $K = \lceil \log_2 B \rceil < \log_2 B + 1$), we have $\frac{\delta}{\max(\sigma_i, \delta)} \geq \min(1, \delta) \geq \frac{1}{2}$, and therefore

$$\gamma_i \geq \frac{1}{10}\min\left(1, \frac{\delta}{\left|x_i^\top e_t^{(k)}\right|}\right)$$

for $i \in S$. Now, suppose $L := \sqrt{\frac{2C}{c_1}}$. Consider those indices $i \in S$ that additionally satisfy $\left|x_i^\top e_t^{(k)}\right| \leq L\delta^{(k)}$. By our fourth moment condition on the covariates, we know that

$$\begin{aligned}
\frac{1}{m}\sum_{i\in S}\left\langle x_i, e_t^{(k)}\right\rangle^4 &\leq C\left(\frac{1}{n}\sum_{i\in[n]}\left\langle x_i, e_t^{(k)}\right\rangle^2\right)^2 \\
&= C\left((e_t^{(k)})^\top\left(\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right)e_t^{(k)}\right)^2 \\
&= C\left((e_t^{(k)})^\top\Sigma_n e_t^{(k)}\right)^2 \\
&= C\left\|e_t^{(k)}\right\|_{\Sigma_n}^4.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}\frac{1}{m}\sum_{i\in S,\left|x_i^\top e_t^{(k)}\right|>L\delta^{(k)}}\left\langle x_i, e_t^{(k)}\right\rangle^2 &\le \frac{1}{L^2(\delta^{(k)})^2}\cdot\frac{1}{m}\sum_{i\in S}\left\langle x_i, e_t^{(k)}\right\rangle^4\\ &\le \frac{C\left\|e_t^{(k)}\right\|_{\Sigma_n}^4}{L^2(\delta^{(k)})^2}\\ &\le \frac{C\left\|e_t^{(k)}\right\|_{\Sigma_n}^2}{L^2}.\end{aligned}$$

On the other hand, the covariance domination condition gives

$$\frac{1}{m}\sum_{i\in S}\left\langle x_i, e_t^{(k)}\right\rangle^2 = (e_t^{(k)})^\top\Sigma_S e_t^{(k)} \ge c_1\left\|e_t^{(k)}\right\|_{\Sigma_n}^2.$$

Subtracting, we get

$$\frac{1}{m}\sum_{i\in S,\left|x_i^\top e_t^{(k)}\right|\le L\delta}\left\langle x_i, e_t^{(k)}\right\rangle^2 \ge \left(c_1-\frac{C}{L^2}\right)\left\|e_t^{(k)}\right\|_{\Sigma_n}^2.$$

Plugging in our definition of L, this is just

$$\frac{1}{m}\sum_{i\in S,\left|x_i^\top e_t^{(k)}\right|\le L\delta}\left\langle x_i, e_t^{(k)}\right\rangle^2 \ge \frac{c_1}{2}\left\|e_t^{(k)}\right\|_{\Sigma_n}^2.$$

Combining this with the expression for γ_i that we obtained above, we get

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\gamma_i\left\langle x_i, e_t^{(k)}\right\rangle^2 &\ge \frac{1}{n}\sum_{i\in S,\left|x_i^\top e_t^{(k)}\right|\le L\delta}\gamma_i\left\langle x_i, e_t^{(k)}\right\rangle^2\\ &\ge \frac{1}{10}\min\left(1,\frac{1}{L}\right)\frac{1}{n}\sum_{i\in S,\left|x_i^\top e_t^{(k)}\right|\le L\delta}\left\langle x_i, e_t^{(k)}\right\rangle^2\\ &\ge \frac{c_1}{20}\min\left(1,\frac{1}{L}\right)\frac{m}{n}\left\|e_t^{(k)}\right\|_{\Sigma_n}^2.\end{aligned}$$

Using our definition of L above, we get

$$\frac{1}{n}\sum_{i\in[n]}\gamma_i\left\langle x_i, e_t^{(k)}\right\rangle^2 \ge \frac{c_1}{20}\min\left(1,\sqrt{\frac{c_1}{2C}}\right)\frac{m}{n}\left\|e_t^{(k)}\right\|_{\Sigma_n}^2 = \frac{\alpha m}{n}\left\|e_t^{(k)}\right\|_{\Sigma_n}^2$$

where, we used the definition $\alpha := \frac{c_1}{20}\min\left(1,\sqrt{\frac{c_1}{2C}}\right)$.

Now let us define $A := \frac{1}{n}\sum_{i\in[n]} b_i x_i x_i^\top$. Since $0 \le b_i \le 1$, we see that $A \preceq \Sigma_n$. Now, let $M := \Sigma_n^{-1/2} A \Sigma_n^{-1/2}$ and let $v := \Sigma_n^{1/2} e_t^{(k)}$. So, it follows that $0 \preceq M \preceq I$, and that $M^2 \preceq M$. Therefore, we have

$$\begin{aligned}\left\|\Sigma_n^{-1}Ae_t^{(k)}\right\|_{\Sigma_n}^2 &= \|Mv\|_2^2\\ &= v^\top M^2 v\\ &\le v^\top M v\\ &= (e_t^{(k)})^\top A e_t^{(k)}\\ &= \frac{1}{n}\sum_{i\in[n]} b_i\left\langle x_i, e_t^{(k)}\right\rangle^2\\ &\le \left(1-\frac{\alpha m}{n}\right)\left\|e_t^{(k)}\right\|_{\Sigma_n}^2.\end{aligned}$$

Also, since

$$\left\|\Sigma_n^{-1}\frac{1}{n}\sum_{i=1}^{n}x_i\mathbb{E}z_i(\beta_t^{(k)},\delta^{(k)})\right\|_{\Sigma_n}=\left\|\Sigma_n^{-\frac{1}{2}}Ae_t^{(k)}\right\|_2,$$

after taking square roots, we get the overall contraction bound

$$\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n}\leq C_2\delta\sqrt{\frac{d\ln n}{n}}+\sqrt{\left(1-\frac{\alpha m}{n}\right)}\left\|e_t^{(k)}\right\|_{\Sigma_n}.$$

At this point, we're nearly done with the proof of the lemma. The contraction factor above is $\sqrt{(1-\frac{\alpha m}{n})}$. First, we show that

$$0<1-\frac{\alpha m}{n}<1.$$

Since $c_1>0$, $C>0$ and $m>0$, we have that $\frac{\alpha m}{n}>0$, and hence $1-\frac{\alpha m}{n}<1$. Next, observe that

$$\alpha=\frac{c_1}{20}\min\left(1,\sqrt{\frac{c_1}{2C}}\right)\leq\frac{c_1}{20}.$$

Now, by our assumption, we know that $c_1\in(0,1]$ and $m\leq n$, and hence

$$0<\frac{\alpha m}{n}\leq\frac{1}{20}$$

and consequently, we have

$$\frac{19}{20}\leq1-\frac{\alpha m}{n}<1.$$

Now, using the inequality $\sqrt{1-x}\leq1-\frac{x}{2}$ for $x\in[0,1]$, we see that

$$\sqrt{1-\frac{\alpha m}{n}}\leq1-\frac{\alpha m}{2n}.$$

So, our final contraction bound is

$$\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n}\leq C_2\delta\sqrt{\frac{d\ln n}{n}}+\left(1-\frac{\alpha m}{2n}\right)\left\|e_t^{(k)}\right\|_{\Sigma_n}.$$

Since $\kappa=\alpha/2$, this is exactly the claimed bound, which completes the proof.

□

Final error bound. From this point on, the proof follows the same template as in Liang and Yuan [2020]. First, we prove (by induction) that for all stages $k\leq K$, we have

$$\left\|e_0^{(k)}\right\|_{\Sigma_n}\leq\frac{\delta^{(k)}}{2}$$

Note that the claim holds for stage $k=0$ by our assumption. Suppose the claim is true for some stage k with $k<K$. We need to show it for stage $k+1$. Now, in stage k, our assumption implies that $\delta^{(k)}\geq2\left\|e_t^{(k)}\right\|_{\Sigma_n}$ for $t=0$.

We now show that *within* stage k, this condition holds for later time steps t as well. So suppose that it is true for some time step t, and we'll prove it for time step $t+1$. By Lemma C.4, we have that

$$\begin{aligned}\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n}&\leq C_2\delta^{(k)}\sqrt{\frac{d\ln n}{n}}+\left\|e_t^{(k)}\right\|_{\Sigma_n}\left(1-\frac{\kappa m}{n}\right)\\&=\left\|e_t^{(k)}\right\|_{\Sigma_n}+\left(C_2\delta^{(k)}\sqrt{\frac{d\ln n}{n}}-\left\|e_t^{(k)}\right\|_{\Sigma_n}\frac{\kappa m}{n}\right)\end{aligned}$$

Now, we have two cases. First, suppose that the second term (inside the parentheses) on the RHS above is negative. In that case, we see that $\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n}\leq\left\|e_t^{(k)}\right\|_{\Sigma_n}\leq\frac{\delta^{(k)}}{2}$. On the other hand, if the

second term is positive, then we must have that $\left\|e_t^{(k)}\right\|_{\Sigma_n} \leq \frac{C_2 n\delta^{(k)}}{\kappa m}\sqrt{\frac{d\ln n}{n}} = \frac{C_2\delta^{(k)}\sqrt{dn\ln n}}{\kappa m}$. In that case, we will further have

$$\begin{aligned}\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n} &\leq \left\|e_t^{(k)}\right\|_{\Sigma_n} + C_2\delta^{(k)}\sqrt{\frac{d\ln n}{n}} \\ &\leq \frac{C_2\delta^{(k)}\sqrt{dn\ln n}}{\kappa m} + \frac{C_2\delta^{(k)}\sqrt{dn\ln n}}{n} \\ &\leq \frac{2C_2\delta^{(k)}\sqrt{dn\ln n}}{\kappa m} \\ &\leq \frac{\delta^{(k)}}{4}.\end{aligned}$$

In the third step we used $m \leq n$ and $\kappa \leq 1$, and in the last step we used our assumption on m, namely that $m \geq \frac{8C_2\sqrt{dn\ln n}}{\kappa}$. Hence, in either case, we can guarantee that $\left\|e_{t+1}^{(k)}\right\|_{\Sigma_n} \leq \frac{\delta^{(k)}}{2}$, and hence this inequality holds for all steps $t \leq T$.

So, we can now apply Lemma C.4 for all steps $t < T$ with the choice of $\delta = \delta^{(k)}$, and unrolling the recursion gives

$$\begin{aligned}\left\|e_T^{(k)}\right\|_{\Sigma_n} &\leq C_2\delta^{(k)}\sqrt{\frac{d\ln n}{n}}\sum_{i=0}^{T-1}\left(1-\frac{\kappa m}{n}\right)^i + \left(1-\frac{\kappa m}{n}\right)^T\left\|e_0^{(k)}\right\|_{\Sigma_n} \\ &\leq C_2\delta^{(k)}\sqrt{\frac{d\ln n}{n}}\cdot\frac{n}{\kappa m} + \left(1-\frac{\kappa m}{n}\right)^T\frac{\delta^{(k)}}{2}.\end{aligned}$$

By our choice $T \geq \frac{\sqrt{n}\ln n}{\kappa}$, and since $m \geq \sqrt{n}$ (which follows from our assumption on m), we have

$$\left(1-\frac{\kappa m}{n}\right)^T \leq e^{-T\frac{\kappa m}{n}} \leq e^{-\ln n\cdot\frac{m}{\sqrt{n}}} \leq \frac{1}{n}.$$

Hence

$$\left\|e_T^{(k)}\right\|_{\Sigma_n} \leq \frac{C_2\delta^{(k)}\sqrt{dn\ln n}}{\kappa m} + \frac{\delta^{(k)}}{2n} \leq \frac{\delta^{(k)}}{8} + \frac{\delta^{(k)}}{8} = \frac{\delta^{(k)}}{4}, \tag{8}$$

where we used $m \geq \frac{8C_2\sqrt{dn\ln n}}{\kappa}$ and $n \geq 4$. In particular, this implies that $\delta^{(k+1)} = \frac{\delta^{(k)}}{2} \geq 2\left\|e_T^{(k)}\right\|_{\Sigma_n}$, which is what we wanted to show by induction.

Overall, we have shown that $\left\|e_0^{(k)}\right\|_{\Sigma_n} \leq \frac{\delta^{(k)}}{2}$ for all stages $k \leq K$. Applying the first inequality of (8) in the last stage, and using $\delta^{(K)} = B/2^K \leq 1$ together with $\frac{1}{n} \leq \frac{\sqrt{dn\ln n}}{m}$ (as $m \leq n$), we get

$$\left\|e_T^{(K)}\right\|_{\Sigma_n} \leq \frac{C_2\delta^{(K)}\sqrt{dn\ln n}}{\kappa m} + \frac{\delta^{(K)}}{2n} \leq \left(\frac{C_2}{\kappa}+\frac{1}{2}\right)\frac{\sqrt{dn\ln n}}{m}.$$

Since $\widehat{\beta} = \beta_T^{(K)}$, this shows that the algorithm satisfies $\left\|\widehat{\beta}-\beta^*\right\|_{\Sigma_n} \leq C'\frac{\sqrt{dn\ln n}}{m}$ with $C' := \frac{C_2}{\kappa}+\frac{1}{2}$. Finally, the ℓ_2 bound in Theorem 4.1 follows from $\|v\|_2 \leq \lambda_{\min}(\Sigma_n)^{-1/2}\|v\|_{\Sigma_n}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims are reflected in the content of the paper. The claims made are formally proven in the body of the paper and the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions needed for the theoretical results are explicitly discussed. Comparisons and weaknesses of the results are also made.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full set of assumptions is provided with full proofs, with technical details in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not contain experiments

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not provide experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Ethic guidelines have been reviewed and accepted by the authors.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is theoretical in nature and so discussion on societal impacts would be very indirect and vague.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work is theoretical in nature and the raised concerns do not apply.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No licensed artifact has been used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No assets have been introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper is theoretical in nature and it does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper is theoretical in nature and it does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used for improving the writing and for finding technical/calculation errors. They were not used to develop the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.