
Linear Regression with Heteroskedastic Errors

Siddhant Chaudhary¹

Aditya Bhaskara¹

¹Kahlert School of Computing, University of Utah, Salt Lake City, Utah, USA

Abstract

We study the classic linear regression problem under the challenging setting of heteroskedastic noise. We study a model in which we have n sources that each observe linear measurements of an unknown d -dimensional vector β . Each source has an unknown and distinct variance for the error, and the goal is to estimate the parameter vector β , under the assumption that there exist *sufficiently many* sources with small error (specifically, m sources have variance at most 1). Our results show that β can be estimated to sub-constant error, as long as the number of small-error sources, i.e., m , is large enough. We prove two main results. First, we show that even with just one observation per source, under minimal assumptions on the linear measurements, β can be estimated to sub-constant error when $m \geq n^{1-\frac{1}{4d}}$. Second, we show that if we have access to *two* observations per source, under similar assumptions on the linear measurements, β can be estimated to sub-constant error when $m \geq n^{5/6}$.

Our results are related to the recent line of work on mean estimation with heteroskedastic variances, and more specifically, the *subset of signals* model used in this literature. To the best of our knowledge, our results provide the first sub-constant recovery guarantees for regression in this model.

1 INTRODUCTION

Linear regression is one of the most fundamental estimation problems in statistics and machine learning. In the traditional setup, we have an unknown vector $\beta \in \mathbb{R}^d$ that we wish to estimate, and we are given observations of the form (x_i, y_i) , where $x_i \in \mathbb{R}^d$ define “linear measurements”, and $y_i = \langle x_i, \beta \rangle + \eta_i$ are the outcomes of the measurement. In

the classic setting, the errors or noise values η_i are modeled as Gaussian noise, drawn from $\mathcal{N}(0, \sigma^2)$, independently for all i . This setting, where the noise in all observations shares a uniform, bounded variance, is known as the homoskedastic case, and standard estimators like Ordinary Least Squares (OLS) are well-known to be statistically efficient.

However, the uniformity assumption frequently breaks down in modern, large-scale applications—such as collecting data from diverse sensor networks or aggregating user ratings in crowdsourcing systems. In these settings, we receive linear measurements from different sources, and each source may be associated with a different variance (different noise level for different sensors, or different error levels for users; e.g., see Ghosh et al. [2011], Chierichetti et al. [2014]). More formally, suppose we have n distinct sources of measurements, and for each source, we obtain (one or more) linear measurements as before. The crucial difference is that for source i , we have an *unknown* variance σ_i^2 , that could be arbitrary and different for each i . This setting is known as the *heteroskedastic* setting of regression [White, 1980, Politis and Poulis, 2014].

The simplest case is one in which for every source, we have exactly one linear measurement. In this case, we have n observations, of the form (x_i, y_i) as before, where $y_i = \langle x_i, \beta \rangle + \eta_i$ and $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$. In the case when σ_i ’s are known, one can show that Weighted Least Squares (WLS) gives the optimal estimator for β , by down-weighting high-variance observations [White, 1980]. The problem becomes significantly more challenging, however, when these variances are unknown. This will be the focus of this paper.

Regression with heteroskedastic errors is a generalization of the problem of heteroskedastic *mean estimation*, which has been studied in several recent works. Here we are given samples from n Gaussians that have a common mean μ , but unknown and (possibly) different variances. Classic approaches, going back to the works of Chernoff [1964], rely on using “dense regions” of observations to prune out samples of high variance and estimate μ . Building on the work

of Chierichetti et al. [2014], Ghosh et al. [2011], Devroye et al. [2023], Liang and Yuan [2020], Pensia et al. [2019], Compton and Valiant [2024] developed algorithms with nearly optimal guarantees. The results were generalized to high dimensional Gaussians by Diakonikolas et al. [2025]. One of the contributions of this line of work is the so-called *subset of signals* model. Here, one assumes that m of the n Gaussians have variance ≤ 1 . Under this assumption (of course, without knowing the identity of these m Gaussians), one seeks to estimate μ to an error that is sub-constant. If one were to know the identity of the low-variance Gaussians, it is trivial to achieve an estimation error of $1/\sqrt{m}$, and the goal is to approach this bound for smaller m . In one dimension, Compton and Valiant [2024] showed how as long as $m \gg n^{1/4}$, it is possible to achieve a $o(1)$ (or sub-constant) error, that drops as m grows.

1.1 OUR RESULTS

The main question we ask is, when can we obtain sub-constant recovery error in the subset of signals model for more general problems, specifically for linear regression? The works of Yuan and Liang [2020] and Pensia et al. [2019] consider this question, but they cannot obtain sub-constant error guarantees, even for $m = n/2$. We prove that with exactly one observation per source, it is possible to obtain sub-constant error recovery in non-trivial regimes. Our main result is then proving that just two observations per source enable us to get a significantly stronger guarantee.

We now describe our model more formally, along with the results. All our results apply to the case when the dimension d is small and n, m grow. Specifically, the runtimes of the algorithms will be $O(n^{2d})$. Note that the case of mean estimation corresponds to $d = 1$.

Problem Setup. Suppose $\beta \in \mathbb{R}^d$ is the vector we wish to estimate. We have n sources of observations (e.g., sensors making measurements or users providing ratings). For each $i \in [n]$, there is a distribution $\mathcal{D}_i$ over $\mathbb{R}^d$ and a noise level σ_i . We obtain t observations for each source, where $t \in \mathbb{N}$ is pre-specified.¹ Each observation has the form $(x_i^{(j)}, y_i^{(j)})$,

$$y_i^{(j)} := \langle x_i^{(j)}, \beta \rangle + \eta_i^{(j)} \quad (1)$$

for $j = 1, 2, \dots, t$ and $x_i^{(j)} \sim \mathcal{D}_i, \eta_i^{(j)} \sim \mathcal{N}(0, \sigma_i^2)$. We assume that $x_i^{(j)}$ and $\eta_i^{(j)}$ are independent across i, j . We will also need to assume that the distributions $\mathcal{D}_i$ are scaled correctly and sufficiently “spread out”. These will be formalized in Section 2.

We consider the analog of the subset-of-signals model for regression: for m of the n sources, we have $\sigma_i \leq 1$. Once again, if we knew the identity of these sources, for any $t \geq 1$,

¹For example, in the crowdsourcing application, this corresponds to each user providing ratings for two items.

we can estimate the vector β up to an ℓ_2 error of $\sqrt{\frac{d}{mt}}$. If we do not know the identity of the sources, to the best of our knowledge, none of the known algorithms (see Pensia et al. [2019], Yuan and Liang [2020]) can achieve a sub-constant error.

Our first result is to prove that even with $t = 1$, we can achieve the following guarantee:

Theorem 1.1. *For regression with heteroskedastic errors in the subset-of-signals model and $t = 1$ observation per source, when $m \geq n^{1-\frac{1}{4d}}$, we can obtain an estimate $\hat{\beta}$ that satisfies $\|\hat{\beta} - \beta\| \leq O\left(d^{\frac{3}{2}} \frac{n^{2d-\frac{1}{2}}}{m^{2d}}\right)$. Further, the estimation algorithm runs in time $\text{poly}(n^d)$.*

While the requirement on m degrades as d grows, the theorem still covers the practically interesting setting where $m = \Omega(n)$ (i.e., an unknown but constant fraction of the sources have small error). This result also illustrates the following natural approach for regression with heteroskedastic noise: consider the case $t = 1$. Now consider *any* d indices $i_1, i_2, \dots, i_d$. Assuming the vectors $x_i^{(1)}$ are all “spread out”, we can express each vector of the standard basis, say e_r for some $r \in [d]$, as

$$e_r = \sum_{j \in [d]} \alpha_j x_{i_j}^{(1)},$$

where the coefficients satisfy $\sum_j \alpha_j^2 \leq O(1)$ (if the vectors form a “nearly orthogonal” basis). Now from (1), if we take the combination $\sum_j \alpha_j y_{i_j}^{(1)}$, this would equal $\langle e_r, \beta \rangle + \sum_j \alpha_j \eta_{i_j}^{(1)}$. The first term is precisely the r th entry of β . The second term is a sum of Gaussian errors, and so it is also distributed as a Gaussian. If *all* the indices $\{i_1, \dots, i_d\}$ satisfy $\sigma_{i_j} \leq 1$, then the second term would have a small variance. This suggests the following meta-algorithm for estimating the r th coordinate of β : (i) for each d -tuple of indices (denoted τ), form a linear combination as above (call it $u_{\tau,r}$), and (ii) feed all the values $u_{\tau,r}$ to a heteroskedastic mean estimation algorithm to estimate $\langle e_r, \beta \rangle$.

The main catch with this approach is that the errors are *dependent*. Specifically, if τ and τ' share an index i , they involve the same η_i term. Meanwhile, all the known algorithms for heteroskedastic mean estimation require independence of the noise. We observe that one of the algorithms, specifically that of Liang and Yuan [2020], works also with noise terms that are potentially dependent, as long as they satisfy a sufficiently good “cancellation of sums” property. We isolate this condition (see Section 3.1) and show that the idea above can be used to prove Theorem 1.1.

Our second (and main) result shows that even with $t = 2$ samples per source, we can achieve a much better dependence between m and n . We show the following:

Theorem 1.2. *For regression with heteroskedastic errors in the subset-of-signals model and $t = 2$ observations per source, when $m \geq n^{5/6}$, we can obtain an estimate $\hat{\beta}$ that satisfies $\|\hat{\beta} - \beta\| \leq \tilde{O}\left(d^3 \left(\frac{n^5}{m^6}\right)^{1/3}\right)$. The estimation algorithm once again runs in time $\text{poly}(n^d)$.*

This algorithm is based on *modal estimators*, an idea for heteroskedastic mean estimation that goes back to the classic work of Chernoff [1964]. The idea is that in mean estimation with independent errors, the center of the highest “density” region is a good estimate of the mean. But once again, the usual analysis heavily relies on independence across samples. We thus perform a two-step algorithm. First, we prune away some of the indices, and ensure that the number of high-variance indices is small. For the remaining (un-pruned) indices, we consider the modal estimator on a carefully constructed set of values, and argue that it leads to an estimate of β with sub-constant error, assuming $m \geq n^{5/6}$. Both the steps involve novel ideas. For pruning, we use an idea similar to the one above for Theorem 1.1: we “guess” d indices $i_1, i_2, \dots, i_d$ such that $\sigma_{i_j} \leq 1$. We show how the estimate of $\langle e_r, \beta \rangle$ as before can be used to predict (up to some error) whether a *new index* i' has $\sigma_{i'} < \log n$ or not. Using this “test”, we can prune away several large $\sigma_{i'}$ values, ending up with only a subset of the indices.

Then we come to the second step of the algorithm, where we restrict to the pruned set of indices/sources; we use the second observations from the sources, and analyze the modal estimator on appropriately constructed estimates for $\langle e_r, \beta \rangle$. The main technical result is to prove that it yields a sub-constant error in estimating β , with high probability.

1.2 RELATED WORK

Linear regression with heteroskedastic errors has been extensively studied empirically, over several decades [White, 1980, Engle, 1982, Atkinson et al., 2016, Amado et al., 2023]. 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], Yuan and Liang [2020], 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. [2025] extends these results to high dimensions using new ideas.

Classical heteroskedastic regression can be solved by using the standard weighted ordinary least squares (WLS) regressor. However, WLS, wherein the samples are weighted according to the inverse of their variances $\left(\frac{1}{\sigma_i}\right)$, assumes knowledge of the variances (or attempts to estimate them). In contrast, this paper considers the setting of heteroskedasticity where the learner has no knowledge of the variances, i.e, the variances are *unknown* and *arbitrary*. To the best of our knowledge, this problem has been considered only in Pensia et al. [2019] and Yuan and Liang [2020]. The algorithm of Pensia et al. [2019] (the *modal interval regression estimator*) achieves the error bound $\|\hat{\beta} - \beta\| \leq \frac{c' n \sigma_{(\frac{cd \log n}{\lambda_{\min}})}}{\lambda_{\min}}$, where c', c are constants and the covariates x_i are assumed to be normally distributed with $x_i \sim N(\mu', \Sigma'_i)$ and $\lambda_{\min} := \min_i \lambda_{\min}(\Sigma'_i)$. Crucially, this error bound *increases* with the number of samples n . In contrast, our algorithms make weaker assumptions on the distributions of x_i ’s, and achieve a sub-constant error bound. Yuan and Liang [2020] study a regression algorithm based on *iterative trimming*. They make similar assumptions (as to ours) for the distributions of the covariates, and their algorithm obtains the error bound $\|\hat{\beta} - \beta\| \leq cc_1 \sigma_{(\lceil \alpha n \rceil)}$, where c, c_1 and α are constants. Again, their error bound does not vanish with an increasing number of data points, which is not ideal for this problem. Other outlier models in regression have also been studied in recent ML literature (see Diakonikolas et al. [2018] and references therein), e.g., including the Huber contamination model, where an adversary is allowed to arbitrarily corrupt an ϵ -fraction of the input observations. However, these settings are different from the one we have considered, and to the best of our knowledge, only Pensia et al. [2019] and Yuan and Liang [2020] qualify as directly comparable baselines. One of our algorithms also shows resemblance to the trimmed-means approach used in mean estimation for heavy-tailed distributions [Oliveira and Resende, 2025, Lugosi and Mendelson, 2019]

2 PRELIMINARIES AND ASSUMPTIONS

Notation. The notation $[n]$ is used to denote a set of n items, i.e $[n] := \{1, 2, \dots, n\}$. For any subset $T \subseteq [n]$ and any k such that $1 \leq k \leq n$, we use $\binom{T}{k}$ to denote the set of all k -sized subsets of $[n]$ with elements from T , i.e $\binom{T}{k} := \{(i_1, \dots, i_k) : i_j \in T \forall j \in [k], i_1 < i_2 < \dots < i_k\}$. Throughout the manuscript, we will use the terms *k-subsets* and *k-tuples* interchangeably. The notation $a \lesssim b$ (equivalently $b \gtrsim a$) is used to denote the fact that $a \leq Cb$

for a positive constant $C > 0$. For a matrix X , the notation $\sigma_d(X)$ is used to denote the d th largest singular value of X . For any set $\{s_1, s_2, \dots, s_p\}$, the i th smallest element is denoted by $s_{(i)}$ for any $1 \leq i \leq p$. For a vector x , $\|x\|$ denotes its ℓ_2 norm. The notation $\tilde{O}(\cdot)$ hides log factors in n and d .

Problem Setup and Assumptions. As described in the Introduction, we have n sources of observations, each with a distribution $\mathcal{D}_i$ over $\mathbb{R}^d$. Observations $\{(x_i^{(j)}, y_i^{(j)})\}_{i \in [n], j \in [t]}$ are produced as in (1). We assume that $x_i^{(j)}$ and $\eta_i^{(j)}$ are chosen independently for different i, j .

The assumptions on $\mathcal{D}_i$ are now as follows:

1. (Scaling) $\forall i$ and all x_i in the support of $\mathcal{D}_i$, $\|x_i\| \leq 1$.
2. (Well-Spread) Suppose we fix j and consider the vectors $x_i^{(j)} \sim \mathcal{D}_i$, for $i \in [n]$. For any $2d$ -tuple $(i_1, i_2, \dots, i_{2d})$, we assume that $\{x_{i_\ell}^{(j)}\}_{\ell \in [2d]}$ satisfy the c -well-conditioned property with probability at least 0.9, over the sampling of $x_i^{(j)}$, where c is some positive constant.

We consider two notions of vectors being well conditioned:

Definition 2.1. (c -Well-conditioned property) A collection of vectors $\{x_1, x_2, \dots, x_m\} \subset \mathbb{R}^d$ is said to be c -well-conditioned for some parameter $c > 0$ if the matrix X obtained by taking $\{x_i\}_{i \in [m]}$ as the columns satisfies $\sigma_d(X) \geq c$.

Note that for a collection of vectors drawn uniformly at random over the sphere in $\mathbb{R}^d$, we expect X to be full rank with high probability, even for $m = d$. If we select $m = 2d$ vectors, the well-known results of Rudelson and Vershynin [2009] imply that $\sigma_d(X) \geq \Omega(1)$, with probability $1 - \exp(-\Omega(d))$. The following lemma is then easy:

Lemma 2.2. Suppose $\{x_1, \dots, x_m\}$ satisfy the c -well-conditioned property from Definition 2.1. Then for any unit vector $v \in \mathbb{R}^d$, we can express

$$v = \sum_j \alpha_j x_j, \text{ with } \sum_j \alpha_j^2 \leq 1/c^2.$$

For our main result (Theorem 1.2) we require a stronger notion, which also holds w.h.p. for random vectors:

Definition 2.3. (Leave-one-out c -well-conditioned property). A collection of vectors $\{x_1, x_2, \dots, x_m\} \subset \mathbb{R}^d$ is said to satisfy the leave-one-out c -well-conditioned property for a parameter $c > 0$ if for all $(m-1)$ -sized subsets T of $[m]$, the matrix X_T obtained by taking $\{x_i\}_{i \in T}$ as the columns satisfies $\sigma_d(X_T) \geq c$.

This stronger well-conditioned property allows us to obtain a version of Lemma 2.2 with the additional property that $|\alpha_j|$ are also *lower bounded*, provided we use $2d$ vectors:

Theorem 2.4. Let $X = [x_1, \dots, x_{2d}] \in \mathbb{R}^{d \times 2d}$ be a matrix whose columns are $x_i \in \mathbb{R}^d$. Suppose that the collection $\{x_1, \dots, x_{2d}\}$ satisfies the leave-one-out c -well-conditioned property. Additionally, assume that for every index $i \in \{1, \dots, 2d\}$, the submatrix X_{-i} (formed by removing the i -th column) has full rank.

Then for any $u \in \mathbb{R}^d$, there exists a “well-spread” solution $\alpha \in \mathbb{R}^{2d}$ to $X\alpha = u$. Specifically, there exists α such that:

1. $X\alpha = u$
2. $\frac{1}{4} \cdot \|u\|^2 \leq \|\alpha\|^2 \leq 8 \left(1 + \frac{2}{c^2}\right) \cdot \|u\|^2$.
3. For all $i \in \{1, \dots, 2d\}$:

$$|\alpha_i| \geq \frac{\|u\|}{8\sqrt{2}d^{1.5}}.$$

Furthermore, the solution can be found in (expected) polynomial time in d .

The proof of the above can be found in Appendix B. Note that the two notions above apply to a given $2d$ -tuple. Assumption (2) above implies that the *expected* number of tuples that satisfy the well conditioned property is $0.9 \binom{n}{2d}$. We can convert this into a high-probability bound as well:

Theorem 2.5. Define $\mathcal{T} = \binom{[n]}{2d}$, and for a tuple $\tau \in \mathcal{T}$, let X_τ denote the indicator for whether τ is well-conditioned. Suppose $\mathbb{E}[X_\tau] \geq 0.9$ for all τ . Then (for large enough n)

$$\Pr \left[\sum_{\tau} X_\tau < 0.8 \binom{n}{2d} \right] < \frac{1}{n^4}.$$

The proof is deferred to Appendix A. It follows from applying Janson’s concentration inequality [Janson, 2004] for sums of correlated random variables.

For any subset $T \subseteq [n]$, $k \in [n]$ and $j \in [t]$, we define $\binom{T}{k}_{\text{good}}^{(j)}$ as the set of all tuples $(i_1, \dots, i_k) \in \binom{T}{k}$ such that $\{x_{i_1}^{(j)}, \dots, x_{i_k}^{(j)}\}$ form a well-conditioned basis. We also use the short-hand $\mathcal{T}_{\text{good}} := \binom{[n]}{2d}_{\text{good}}^{(j)}$, where $j = 1$ or $j = 2$ is understood from the context. From Theorem 2.5, we see that with high probability $|\mathcal{T}_{\text{good}}| \gtrsim \binom{n}{2d}$.

3 ITERATIVE TRUNCATION

We first discuss the setting where we have only $t = 1$ observations per user. We thus abuse notation slightly and write $x_i := x_i^{(1)}$, $y_i := y_i^{(1)}$ and $\eta_i := \eta_i^{(1)}$. In this case, we adapt the *iterative truncation* algorithm [Liang and Yuan, 2020], by leveraging additional linear algebraic structure.

As outlined in the introduction, Algorithm 1 first collects all $2d$ -tuples $\tau \in \mathcal{T}_{\text{good}}$. Using any such tuple, we can form an

estimate of the vector β . Let us focus on the r th coordinate. We can represent the unit vector e_r as

$$e_r = \sum_{j=1}^{2d} \alpha_{j,r}(\tau) x_{i_j} \quad (2)$$

Note that, as mentioned in Theorem 2.4, these coefficients can be computed using a randomized algorithm (see Appendix B for more details). Throughout the manuscript, wherever these coefficients are computed, implicit use of this randomized algorithm will be assumed. Given this representation, consider $u_{\tau,r}$ defined by

$$u_{\tau,r} := \sum_{j=1}^{2d} \alpha_{j,r}(\tau) y_{i_j}$$

Re-writing each y_{i_j} in the above equation as in (1) and using (2), it follows that

$$u_{\tau,r} = \beta_r + \sum_{j=1}^{2d} \alpha_{j,r}(\tau) \eta_{i_j}. \quad (3)$$

In particular, we see that $u_{\tau,r} \sim \mathcal{N}(\beta_r, \sigma_{\tau,r}^2)$ where $\sigma_{\tau,r}^2 := \sum_{j=1}^{2d} \alpha_{j,r}^2(\tau) \sigma_{i_j}^2$. To summarize, these linear representations allow us to obtain $|\mathcal{T}_{\text{good}}| \gtrsim \binom{n}{2d}$ normally distributed samples about β_r . If there are m indices with $\sigma_i \leq 1$, we can expect $\gtrsim \binom{m}{2d}$ of the $u_{\tau,r}$ to have low variance. Thus it is natural to now feed these values $u_{\tau,r}$ to a mean estimation procedure. The catch is that even though each $u_{\tau,r}$ is distributed as a Gaussian, the $u_{\tau,r}$ values are not independent! (E.g., if τ, τ' share an index i_j , the “error” in both terms depends on η_{i_j} .) Because of this limitation, most existing algorithms for mean estimation in one-dimension [Pensia et al., 2019, Yuan and Liang, 2020, Devroye et al., 2023, Compton and Valiant, 2024] cannot be applied to our setting right out of the box.

3.1 MEAN ESTIMATION OF DEPENDENT DATA USING ITERATIVE TRUNCATION

Our result here will be to show that the iterative truncation algorithm of Liang and Yuan [2020] still applies to our dependent regime, albeit with weaker guarantees. To recall, their algorithm truncates samples to a specified interval, and computes the mean of all truncated samples as the current estimate of the true mean. To that end, for any interval $[a, b]$ and $u \in \mathbb{R}$, define

$$\phi(u; [a, b]) := \begin{cases} a, & \text{if } u < a \\ u, & \text{if } a \leq u \leq b \\ b, & \text{if } u > b \end{cases} \quad (4)$$

We defer the pseudocode of their algorithm to the appendix (see Algorithm 4). In the rest of this section, we show how to extend their work to the case of *dependent* input samples.

To achieve this, we utilize machinery from Janson [2004]. Suppose $u_1, u_2, \dots, u_N$ are (possibly dependent) input samples such that $u_i \sim \mathcal{N}(\mu^*, \sigma_i^2)$ for all $i \in [N]$, where $\mu^* \in \mathbb{R}$ is the unknown mean to be estimated. We say that these inputs satisfy the *weak cancellation property*, parametrized by Γ , if for every $\epsilon > 0$, with probability at least $1 - \epsilon$, it holds that

$$\left| \sum_{i \in [N]} (u_i - \mathbb{E}[u_i]) \right| \leq O\left(\sqrt{N\Gamma \ln(1/\epsilon)}\right).$$

For *independent* random variables, we have $\Gamma = 1$ by the Chernoff bound (when $\sigma_i \leq 1$). The notion of using Γ to quantify “independent randomness” has been used in several settings. A famous example from number theory concerns the Möbius function (e.g., see Odlyzko and Riele [1985]). In the context of random walks, the so-called Hurst exponent is similar in spirit to the definition above (see Mandelbrot and Van Ness [1968]).

As it turns out, we can show that the guarantees of [Liang and Yuan, 2020] hold (up to Γ losses) as long as the inputs have the weak cancellation property, extended to projections onto intervals. We formalize this in Theorem 3.1 below:

Theorem 3.1. *Suppose $u_1, u_2, \dots, u_N$ are drawn from a joint distribution. Assume that the marginals satisfy $u_i \sim \mathcal{N}(\mu^*, \Sigma_i^2)$, and suppose that $\Sigma_{(M)} \leq C$ for a parameter $1 \leq M \leq N$ and a constant $C > 0$. Furthermore, suppose that for any fixed interval $\Delta = [\mu - \delta, \mu + \delta]$, the projections $w_i := \phi(u_i; \Delta)$ satisfy the weak cancellation property for some $\Gamma \geq 1$: i.e, with probability at least $1 - \epsilon$,*

$$\left| \sum_{i \in [N]} (w_i - \mathbb{E}[w_i]) \right| \leq C' \left(\delta \sqrt{N\Gamma \ln(1/\epsilon)} \right), \quad (5)$$

for any $\epsilon > 0$, where $C' > 0$ is another constant. Then with probability at least $\geq 1 - \frac{4}{N}$, whenever M is large enough such that $M > 4\sqrt{N\Gamma \ln N}$, the iterative truncation algorithm (Algorithm 4) on inputs $u_1, u_2, \dots, u_N$ outputs an estimate $\hat{\mu}$ that satisfies

$$|\hat{\mu} - \mu^*| \leq O\left(\frac{\sqrt{\Gamma N \ln N}}{M}\right) \quad (6)$$

We defer the proof of this theorem to Appendix C.

3.2 REGRESSION WITH ITERATIVE TRUNCATION

Having adapted Liang and Yuan [2020]’s algorithm to the case of dependent input samples, we now state and prove the estimation guarantee of our algorithm.

Algorithm 1: Iterative Truncation Based Regression

Input: Input-output pairs (y_i, x_i) for $i \in [n]$ satisfying
 (1) (with superscripts dropped because $t = 1$),
 initialization β_0 and parameters B, m such that
 $B \geq 2 \|\beta - \beta_0\|, \sigma_{(m)} \leq 1$.

```

1 for each  $r \in [d]$  do
  // Estimate each dimension
2   for each  $\tau \in \mathcal{T}_{\text{good}}$  do
3     Represent  $e_r = \sum_{j=1}^{2d} \alpha_{j,r}(\tau) x_{ij}$ ;
4     Compute  $u_{\tau,r} := \sum_{j=1}^{2d} \alpha_{j,r}(\tau) y_{ij}$ ;
5   end
6   Estimate  $\hat{\beta}_r \leftarrow$ 
     IterTrun  $\left( \{u_{\tau,r}\}, (\beta_0)_r, B, M := |\mathcal{T}'_{\text{good}}|, C \right)$ ,
     where  $C := 8 \left(1 + \frac{2}{c^2}\right)$  and the constant  $c$  is from
     assumption 2;
7 end
Output:  $\hat{\beta} := (\hat{\beta}_1, \dots, \hat{\beta}_d)$ .
```

Theorem 3.2. With probability at least $1 - \frac{4d}{N}$, Algorithm 1 returns an estimate $\hat{\beta}$ which satisfies

$$\|\hat{\beta} - \beta\| \leq \tilde{O} \left(d^{\frac{3}{2}} \frac{n^{2d-\frac{1}{2}}}{m^{2d}} \right) \quad (7)$$

for $m \geq n^{1-\frac{1}{4d}}$ where above, $N := |\mathcal{T}_{\text{good}}|$.

We now sketch the proof of the error bound of Theorem 3.2. Throughout, we put $N := |\mathcal{T}_{\text{good}}|$. Additionally, let $\mathcal{T}'_{\text{good}}$ be the subset of $\mathcal{T}_{\text{good}}$ containing all tuples $(i_1, \dots, i_{2d}) \in \mathcal{T}_{\text{good}}$ such that $\sigma_{i_j}^2 \leq 1$ for all $j \in [2d]$ (i.e all indices are small-variance). Now, put $M := |\mathcal{T}'_{\text{good}}|$. By an argument similar to that of Theorem 2.5, we see that with high probability, $\binom{n}{2d} \geq N \gtrsim \binom{n}{2d}$ and that $\binom{m}{2d} \geq M \gtrsim \binom{m}{2d}$.

Fix some $r \in [d]$; we will first give a bound on $|\hat{\beta}_r - \beta_r|$, i.e we'll analyze lines 2-6 of Algorithm 1. Let $\tau = (i_1, \dots, i_{2d}) \in \mathcal{T}'_{\text{good}}$ be any tuple. In that case, observe that

$$\begin{aligned}
\sigma_{\tau,r}^2 &= \sum_{j=1}^{2d} \alpha_{j,r}^2(\tau) \sigma_{i_j}^2 \\
&\leq \sum_{j=1}^{2d} \alpha_{j,r}^2(\tau) \quad (1 \leq i_j \leq m \ \forall j \in [2d]) \\
&\leq 8 \left(1 + \frac{2}{c^2}\right) \quad (\text{Well-conditioned basis})
\end{aligned}$$

where in the last step above we're using the upper bound from property 2 of Theorem 2.4. So, from the above, we see

that there are at least M tuples τ with $\sigma_{\tau,r}^2 \leq 8 \left(1 + \frac{2}{c^2}\right)$. Also, note that the total number of tuples is N . So, we feed inputs $u_{\tau,r}$ as computed in equation (3) over all $\tau \in \mathcal{T}_{\text{good}}$ to the iterative truncation algorithm (Algorithm 4) and invoke Theorem 3.1. Note that in order to invoke Theorem 3.1, we need to show that the inputs $u_{\tau,r}$ satisfy the weak cancellation property as in (5).

Lemma 3.3. The inputs $u_{\tau,r}$ defined in (3) satisfy the weak cancellation property of (5) for $\Gamma := \binom{n}{2d-1}$.

We prove Lemma 3.3 in Appendix D, where we use a simple argument using Janson's bound. Hence, by Theorem 3.1, we see that with probability at least $1 - \frac{4}{N}$, the error bound of line 6 of Algorithm 1 is $|\hat{\beta}_r - \beta_r| \leq O \left(\frac{\sqrt{\Gamma N \ln N}}{M} \right)$.

All that is left to be done is to now simplify the above bound. To that end, note that $\binom{n}{2d-1} \leq \frac{n^{2d-1}}{(2d-1)!}$. Moreover, since $N \leq \binom{n}{2d}$, we see that $N \leq \frac{n^{2d}}{(2d)!}$ and that $\ln N \leq \ln(n^{2d}) = 2d \ln n$. So, we see that

$$\begin{aligned}
\sqrt{\binom{n}{2d-1} N \ln N} &\leq \sqrt{\frac{n^{2d-1}}{(2d-1)!} \cdot \frac{n^{2d}}{(2d)!} \cdot 2d \ln n} \\
&= \sqrt{\frac{2d \cdot n^{4d-1} \ln n}{(2d-1)!(2d)!}}
\end{aligned}$$

Also, note that $M \gtrsim \binom{m}{2d}$. So we have $M \gtrsim \frac{(m-2d)^{2d}}{(2d)!} = \frac{m^{2d}}{(2d)!} \left(1 - \frac{2d}{m}\right)^{2d}$. We can choose n large enough so that $\left(1 - \frac{2d}{m}\right)^{2d} \geq \frac{1}{2}$; in particular, we see that $M \gtrsim \frac{m^{2d}}{2(2d)!}$. Using these bounds, we see that

$$\begin{aligned}
e_T^{(K)} &\lesssim \sqrt{\frac{2d \cdot n^{4d-1} \ln n}{(2d-1)!(2d)!}} \cdot \frac{2(2d)!}{m^{2d}} \\
&= 2 \sqrt{\frac{2d(2d)!}{(2d-1)!}} \cdot \frac{\sqrt{n^{4d-1} \ln n}}{m^{2d}} \\
&= 4d \frac{\sqrt{n^{4d-1} \ln n}}{m^{2d}} \quad (8)
\end{aligned}$$

Finally, note that the above bound is a bound on $|\hat{\beta}_r - \beta_r|$. Aggregating this bound over all $r \in [d]$, the proof of the theorem is complete.

4 HYBRID ALGORITHM

We now discuss the setting where we have $t = 2$ observations per source. For this setting, we design a hybrid algorithm that involves i) a *screening step* where large variance indices are pruned and ii) an estimation step where we employ the well-known *modal estimator* [Chernoff, 1964] to estimate each coordinate of the unknown weight vector

β . Many previous works in the mean-estimation setting use such a hybrid procedure to obtain sub-constant error rates [Chierichetti et al., 2014, Pensia et al., 2019, Devroye et al., 2023]. Since we have 2 observations for each source, we use the observations corresponding to $t = 1$ in the screening step (Algorithm 2), and the observations corresponding to $t = 2$ in the estimation step (Algorithm 3).

4.1 PRUNING HIGH-VARIANCE INDICES

The key idea behind the pruning step is that if the algorithm can “guess” $2d$ indices i with $\sigma_i \leq 1$, then it can create a good way to test or screen a *new index* i' and check if $\sigma_{i'} \leq O(\log n)$ or not. The idea is as follows. Suppose we know a set of $2d$ indices τ all having small σ values. Assuming $\{x_i\}_{i \in \tau}$ are well-conditioned, we can express $e_1 = \sum_{i \in \tau} \alpha_i x_i$. Likewise, for any index i' , we can write $e_1 - x_{i'} = \sum_{i \in \tau} \gamma_i x_i$. Now the key observation is that if we look at the corresponding combination of the y values and take the difference $\sum_{i \in \tau} \gamma_i y_i - \sum_{i \in \tau} \alpha_i y_i$, then this quantity turns out to be $\eta_{i'} \pm O(\log n)$ with high probability. Thus if an $\eta_{i'} \gg \log n$, we can expect to detect it w.h.p.!

Algorithm 2: Generate Inlier Candidates

Input: Input-output pairs $(y_i^{(1)}, x_i^{(1)})$ for $i \in [n]$ satisfying (1) such that $\sigma_{(m)} \leq 1$.

```

1  $T \leftarrow \{\}$ ;
2 for all  $2d$ -tuples  $\tau = (i_1, \dots, i_{2d}) \in \mathcal{T}_{\text{good}}$  do
3   Represent  $e_1 = \sum_{j=1}^{2d} \alpha_j(\tau) x_{i_j}^{(1)}$ ;
4    $u_\tau \leftarrow \sum_{j=1}^{2d} \alpha_j(\tau) y_{i_j}^{(1)}$ ;
5    $I \leftarrow [u_\tau - 6 \log n, u_\tau + 6 \log n]$ ;
6    $T_{\text{candidate}} \leftarrow \{\}$ ; // Empty set
7   for  $k \in [n] \setminus \tau$  do
8     Represent  $e_1 - x_k^{(1)} \leftarrow \sum_{j=1}^{2d} \gamma_j(\tau) x_{i_j}^{(1)}$ ;
9     if  $y_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) y_{i_j}^{(1)} \in I$  then
10       $T_{\text{candidate}} \leftarrow T_{\text{candidate}} \cup \{k\}$ ;
11    end
12  end
13  if  $|T| < |T_{\text{candidate}}|$  then
14     $T \leftarrow T_{\text{candidate}}$ ;
15  end
16 end
Output:  $T$ 

```

The main steps in the analysis will be to prove that for some choice of τ , all the low-variance indices survive, and not many of the high-variance indices do. Our first lemma will deal with the first portion of the argument, i.e., the candidate set T of indices returned by Algorithm 2 is at least $0.9m$. For a proof, we refer the reader to Appendix E.

Lemma 4.1 (Good indices survive). *Let m be such that $m - 2d \geq \frac{9}{10}m$. With probability at least $1 - n \exp\left(-\frac{9 \log^2 n}{2}\right)$ –*

$n \exp\left(-\frac{9 \log^2 n}{10C_0}\right)$, there is a tuple $\tau \in \mathcal{T}_{\text{good}}$ for which Algorithm 2 returns a set T such that $|T| \geq \frac{9}{10}m$. Specifically, if $\tau = (i_1, \dots, i_{2d}) \in \mathcal{T}_{\text{good}}$ is such that $\sigma_{i_j}^2 \leq 1$ for all $j \in [2d]$, then with high probability, all indices $i \in [n] \setminus \tau$ with $\sigma_i^2 \leq 1$ pass the condition of line 9 of Algorithm 2 for such a τ .

Next, we formalize and prove the fact that the total number of indices $k \in [n]$ in the returned set with high σ_k is small. At a high-level, we do this by dividing the range of the variances σ_k into intervals with endpoints being powers of 2, and then bounding the sizes of each of the intervals. For the proof, we refer the reader to Appendix F.

Lemma 4.2 (Variance level counts). *Let T be the set output by Algorithm 2, and let $L = \frac{1}{4} \log n$. For any integer $\ell \in [0, L]$, let n_ℓ be the number of indices $k \in T$ with $\sigma_k \in \text{Range}_\ell$, where*

$$\text{Range}_\ell := \begin{cases} [0, 2^\ell] & , \text{ if } \ell = 0 \\ [2^{\ell-1}, 2^\ell] & , \text{ if } 0 < \ell < L \\ [2^{\ell-1}, \infty) & , \text{ if } \ell = L \end{cases} \quad (9)$$

Also define $\sigma_\ell := 2^\ell$ for $\ell \in [0, L]$. Then, with probability at least $1 - \frac{n^{2d} \log n}{4} \exp\left(-K_0 n^{\frac{1}{2}} \log^2 n\right)$, we have that $n_\ell \leq K_1 \frac{n \log n}{\sigma_\ell}$ for all integers $\ell \in [0, L]$ where $K_0 = \frac{46656}{\pi}$ and $K_1 = \frac{240}{\sqrt{2\pi}}$.

4.2 RUNNING THE MODAL ESTIMATOR ON THE PRUNED SET

We now discuss stage two of the hybrid algorithm. In this stage, we generate inputs $u_{\tau,r}$ as in (3) using observations corresponding to $t = 2$ with a slight catch: in our $2d$ -tuples, we only use indices from the set T returned by the screening step (Algorithm 2); see lines 2-4 of Algorithm 3. We then feed these inputs to the *modal estimator* [Chernoff, 1964]. See Algorithm 3 for the pseudocode. Throughout our discussion in this section, we additionally assume that

$$\sigma_i^2 \geq \frac{1}{2} \quad (10)$$

for all $i \in [n]$. To achieve this, we add a constant Gaussian variance of 0.5 to *every* observation. Because of this, the new “small” variance bound now increases from 1 to 1.5 (i.e small variances are still bounded by a constant), but the analysis and the error bounds still remain the same.

Theorem 4.3. *With probability at least $\frac{9}{10}$, Algorithm 3 run with $w = 1$ returns an estimate $\hat{\beta}$ such that*

$$\|\hat{\beta} - \beta\| \leq \tilde{O}\left(d^3 \left(\frac{n^5}{m^6}\right)^{1/3}\right) \quad (11)$$

for all $m \geq n^{5/6}$.

Algorithm 3: Modal Estimator for Regression

Input: Input-output pairs $(y_i^{(2)}, x_i^{(2)})$ for $i \in [n]$ satisfying (1) such that $\sigma_{(m)} \leq 1$, testing interval width w .

```

// Prune high variance indices
1  $T \leftarrow$  Pruned set of indices generated using Algorithm 2;
// Estimate each dimension
2  $\mathcal{T}' := \binom{T}{2d} \cap \mathcal{T}_{\text{good}}$ ;
3 for each dimension  $r \in [d]$  do
4   for each 2d-tuple  $\tau \in \mathcal{T}'$  do
5     Represent  $e_r = \sum_{j=1}^{2d} \alpha_{j,r}(\tau) x_{i_j}^{(2)}$ ;
6      $u_{\tau,r} \leftarrow \sum_{j=1}^{2d} \alpha_{j,r}(\tau) y_{i_j}^{(2)}$ ;
7   end
// Compute the empirical mode
8  $\hat{\beta}_r \leftarrow \operatorname{argmax}_{u \in \mathbb{R}} \sum_{\tau \in \mathcal{T}'} \mathbb{1}_{u_{\tau,r} \in [u-w, u+w]}$ ;
9 end

```

Output: $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_d)$

We give a sketch of how the error bound of Theorem 4.3 is obtained. As in line 2 of Algorithm 3, define $\mathcal{T}' := \binom{T}{2d} \cap \mathcal{T}_{\text{good}}$ to be the set of all well-conditioned 2d-tuples with indices in T (recall that T is the set of indices returned by the screening step). By an argument similar that of Theorem 2.5, we see that $\binom{T}{2d} \geq |\mathcal{T}'| \gtrsim \binom{T}{2d}$. We fix the dimension $r \in [d]$, and give an error bound on $|\hat{\beta}_r - \beta_r|$.

For any $u \in \mathbb{R}$, let $K(u) := \sum_{\tau \in \mathcal{T}'} \mathbb{1}_{u_{\tau,r} \in [u-w, u+w]}$. In simple words, $K(u)$ is the number of inputs $u_{\tau,r}$ that are at a distance of at most w from u . Since the modal estimator (line 8 of Algorithm 3) is translation invariant, we also assume without loss of generality that $\beta_r = 0$.

Now, let $\Delta > 0$ be a small perturbation amount which we will later determine. We will show that, with high probability, for any $u' \in \mathbb{R}$ with $|u'| > \Delta$, we have that $K(0) - K(u') > 0$; in other words, with high probability, line 8 of Algorithm 3 will return an estimate $\hat{\beta}_r$ such that $|\beta_r| \leq \Delta$. Note that this will imply a high-probability error bound of the form $|\hat{\beta}_r - \beta_r| \leq \Delta$. In fact, we will obtain a Δ that is sub-constant.

We now formalize the above intuition. For any $u' \in \mathbb{R}$, we define $\text{Diff}(u') := K(0) - K(u')$, $\mu(u') := \mathbb{E}[\text{Diff}(u')]$ and $\Sigma(u') := \text{Var}(\text{Diff}(u'))$. Observe that the random variable $K(u')$ takes at most $2|\mathcal{T}'|$ distinct values; this is because the count $K(u')$ is a step function and only changes its value when the boundary of the sliding window $[u' - w, u' + w]$ hits or exits a data point $u_{\tau,r}$, i.e. when $u' = u_{\tau,r} \pm w$. Since there are at most $|\mathcal{T}'|$ input tuples, it follows that $K(u')$ can take at most $2|\mathcal{T}'|$ distinct values. This in turn implies that the quantity $\text{Diff}(u')$ can take at most $2|\mathcal{T}'|$ distinct values; this observation will allow us to now prove a uniform bound

over all $|u'| > \Delta$. First, we prove a lower bound on the signals $\mu(\cdot)$ via the following lemma:

Lemma 4.4 (Lower bounding the signal). *Let $\mu(\cdot)$ be as defined above. Then, for $w = 1$, we have*

$$\mu(\Delta) \geq \mu := C_1 \sum_{\tau \in \mathcal{T}'} \frac{\Delta^2}{\sigma_{\tau,r}^3}$$

where $\sigma_{\tau,r}^2 := \sum_{j=1}^{2d} \alpha_{j,r}(\tau)^2 \sigma_{i_j}^2$ for any $\tau = (i_1, \dots, i_{2d}) \in \mathcal{T}'$ and C_1 is a positive constant. In fact, the same proof shows that if $|u'| \geq \Delta$, then

$$\mu(u') \geq \left(\frac{u'}{\Delta}\right)^2 \cdot \mu \quad (12)$$

For a proof of the above, we refer the reader to Appendix G. The proof is based on simple properties of Gaussian density functions. Next, we upper bound the noise terms in the following lemma:

Lemma 4.5. *Let $\Sigma(\cdot)$ be as defined above. Then, for $w = 1$, we have*

$$\Sigma(\Delta) \leq \Sigma := C_2 \cdot \chi \cdot \sum_{\tau \in \mathcal{T}'} \frac{\Delta}{\sigma_{\tau,r}}$$

where C_2 is a positive constant, χ is the fractional chromatic number of the dependency graph defined by the tuples $\tau \in \mathcal{T}'^2$, and $\sigma_{\tau,r}$ is defined as in Lemma 4.4. In fact, the same proof shows that for any $|u'| \geq \Delta$,

$$\Sigma(u') \leq \left|\frac{u'}{\Delta}\right| \cdot \Sigma \quad (13)$$

For a proof of the above, we refer the reader to Appendix H. We now prove the uniform bound for all u' with $|u'| \geq \Delta$ in the following:

Lemma 4.6. *Let μ , $\mu(\cdot)$, Σ , $\Sigma(\cdot)$, χ be as defined in Lemmas 4.4 and 4.5. Then, with probability at least $1 - 2n^{2d} \exp\left(-S_0 \frac{\mu^2}{\Sigma + 4\chi \cdot \mu}\right)$, for any $u' \in \mathbb{R}$ with $|u'| > \Delta$, we have $\text{Diff}(u') > 0$ for some positive constant S_0 . In particular, with probability at least $1 - 2n^{2d} \exp\left(-S_0 \frac{\mu^2}{\Sigma + 4\chi \cdot \mu}\right)$, line 8 of Algorithm 3 outputs $\hat{\beta}_r$ with $|\hat{\beta}_r - \beta_r| \leq \Delta$.*

The proof of Lemma 4.6 can be found in Appendix I and is based on a simple concentration argument using Janson's inequality [Janson, 2004]. Now, we want to choose a sub-constant Δ for which the probability $1 - 2n^{2d} \exp\left(-S_0 \frac{\mu^2}{\Sigma + 4\chi \cdot \mu}\right) \geq 1 - \frac{1}{10d}$. In other words, we want to solve the inequality

$$n^{2d} \exp\left(-S_0 \frac{\mu^2}{\Sigma + 4\chi \cdot \mu}\right) \leq \frac{1}{10d} \quad (14)$$

²In our setting, we will have $\chi \leq \binom{|T|}{2d} \cdot \frac{2d}{|T|}$, since we're taking 2d-tuples over the set T . See **Example 4.1** of Janson [2004].

Upon simplification, the above reduces to the inequality:

$$\mu^2 \geq \frac{(\Sigma + 4\chi\mu) \cdot (2d \ln n + \ln 10d)}{S_0}$$

Note that for the above, dropping constants, it is enough to find a sub-constant Δ which gives

$$\mu^2 \gtrsim O(d \ln n) \cdot \Sigma, \text{ and } \mu^2 \gtrsim O(d \ln n) \cdot \chi\mu \quad (15)$$

We prove in Appendix J that the choice $\Delta = \tilde{O}\left(d^2 \left(\frac{n^5}{m^6}\right)^{\frac{1}{3}}\right)$ solves both of the above inequalities. In particular, we see that with probability at least $1 - \frac{1}{10d}$, the error satisfies $|\hat{\beta}_r - \beta| \leq \Delta = \tilde{O}\left(d^2 \left(\frac{n^5}{m^6}\right)^{\frac{1}{3}}\right)$. Taking a union bound over all dimensions $r \in [d]$, we see that with probability at least $1 - \frac{1}{10}$, the error satisfies $\|\hat{\beta} - \beta\| \leq \tilde{O}\left(d^3 \left(\frac{n^5}{m^6}\right)^{\frac{1}{3}}\right)$. This completes our analysis.

5 CONCLUSION AND OPEN PROBLEMS

We studied the linear regression problem in the setting where the observations have heteroskedastic noise, dependent on the source of the observations. We showed how to achieve sub-constant estimation error for the unknown vector $\beta \in \mathbb{R}^d$. Our algorithms use a combination of linear algebraic ideas and probabilistic reasoning. The main drawback of our analysis is the dependence on d : our algorithms are based on building estimators using d -tuples of the observations, and thus take time $\text{poly}(n^d)$. The big open problem from our work is to address the case of larger d . It is also interesting to understand the precise threshold of when sub-constant recovery is possible in the subset-of-signals model, as a function of t . For example, with $t = 1$, can one achieve a bound like $m = n^{5/6}$? Furthermore, the error bounds in this work are *coarse*; particularly, it will be interesting if the analysis or the algorithms can be modified to obtain error bounds that scale as a function of the vector $\sigma := (\sigma_1, \dots, \sigma_n)$ of variances. Finally, it will be an interesting open problem to see if sub-constant error guarantees can also be achieved with fewer than $n^{5/6}$ small variance inputs, and to investigate corresponding lower bounds.

References

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>.

T. W. Anderson. Some inequalities for symmetric convex sets with applications. *The Annals of Statistics*, 24(2):

753–762, 1996. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/2242672>.

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>.

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](https://doi.org/10.1016/S0167-7152(99)00072-3). URL <https://www.sciencedirect.com/science/article/pii/S0167715299000723>.

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

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.

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>.

Luc Devroye, Silvio Lattanzi, Gábor Lugosi, and Nikita Zhivotovskiy. On mean estimation for heteroscedastic random variables. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 59(1):1 – 20, 2023. doi: 10.1214/21-AIHP1239. URL <https://doi.org/10.1214/21-AIHP1239>.

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

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression, 2018. URL <https://arxiv.org/abs/1806.00040>.

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, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718162. URL <https://doi.org/10.1145/3717823.3718162>.
- 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](https://onlinelibrary.wiley.com/doi/abs/0012-9682(198207)50:4<987:ACHWEO>2.0.CO;2-3).
- 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>.
- Wassily Hoeffding. Probability inequalities for sum of bounded random variables. 1963. URL <https://api.semanticscholar.org/CorpusID:121341745>.
- Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24, 2004. URL <https://api.semanticscholar.org/CorpusID:14808600>.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL <https://doi.org/10.1214/aos/1015957395>.
- 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>.
- Gabor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions—a survey, 2019. URL <https://arxiv.org/abs/1906.04280>.
- Benoit B. Mandelbrot and John W. Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968. doi: 10.1137/1010093.
- A.M. Odlyzko and H.J.J. te Riele. Disproof of the mertens conjecture. *Journal für die reine und angewandte Mathematik*, 357:138–160, 1985. URL <http://eudml.org/doc/152712>.
- Roberto I. Oliveira and Lucas Resende. Trimmed sample means for robust uniform mean estimation and regression. *The Annals of Statistics*, 53(5), October 2025. ISSN 0090-5364. doi: 10.1214/25-aos2536. URL <http://dx.doi.org/10.1214/25-AOS2536>.
- 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>.
- Ankit Pensia, Varun Jog, and Po-Ling Loh. Estimating location parameters in entangled single-sample distributions, 2019. URL <https://arxiv.org/abs/1907.03087>.
- 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.
- 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>.
- 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](https://onlinelibrary.wiley.com/doi/abs/0012-9682(198005)48:4<817:AHCMEA>2.0.CO;2-K).
- 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>.

Appendix

Siddhant Chaudhary¹

Aditya Bhaskara¹

¹Kahlert School of Computing, University of Utah, Salt Lake City, Utah, USA

A PROOF OF THEOREM 2.5

We note (e.g., Janson [2004], Example 4.1) that the fractional chromatic number of the set of all $2d$ -tuples of $[n]$ is at most $\chi := \binom{n}{2d-1}$. Using this, we have that for any deviation λ ,

$$\Pr \left[\sum_{\tau} (X_{\tau} - 0.9) > \lambda \right] \leq \exp \left(-\frac{\lambda^2}{\chi \cdot \binom{n}{2d}} \right).$$

Setting $\lambda = 0.1 \binom{n}{2d}$, the conclusion follows.

B PROOF OF THEOREM 2.4

The main challenge in the proof is to ensure that $\|\alpha\|^2$ is lower-bounded, and also to ensure that *each coordinate* of the representation is sufficiently large.

Since the entire theorem is scale-invariant, we can assume that $\|u\| = 1$. Note that the set of all solutions to $X\alpha = u$ is the affine subspace:

$$\mathcal{S} = \{\alpha^* + z \mid z \in \ker(X)\}$$

where $\alpha^* = X^\dagger u$ is the minimum ℓ_2 -norm solution, and $\ker(X)$ is the null space of X . The idea is to perturb α^* along carefully chosen directions, and argue that we end up with a solution that satisfies all the desired properties with at least a constant probability.

For each i , define $\alpha^{(i)}$ to be the vector defined as follows: write $x_i = \sum_{j \in [2d] \setminus \{i\}} \gamma_j x_j$, and take $\alpha^{(i)}$ to be the vector that is 1 at index i and $-\gamma_j$ for all other indices. Because of the least singular value bound on X_{-i} , this is firstly possible, and secondly, we can ensure that $\sum_{j \neq i} \gamma_j^2 \leq \frac{1}{c^2}$. By definition $\alpha^{(i)} \in \ker(X)$.

We now construct our desired solution as

$$\alpha = \alpha^* + \sum_{j \in [2d]} g_j \alpha^{(j)},$$

where the g_j are independent Gaussians from $\mathcal{N}(0, \rho^2)$, and $\rho^2 = 1/2d$. We will first prove part (3) of the Theorem, that all coordinates are sufficiently large. For this, we note that α_i is distributed as a Gaussian with sufficiently large variance. Specifically, the variance of α_i is

$$\rho^2 + \sum_{j \neq i} (\alpha_i^{(j)})^2 \rho^2 \geq \rho^2.$$

Thus, by standard anti-concentration properties of a Gaussian, we have that for any parameter ε , the probability that $|\alpha_i| < \varepsilon \rho$ is at most $\frac{2\varepsilon}{\sqrt{2\pi}} < \varepsilon$. We set $\varepsilon = 1/8d$, so that we can take a union bound over all the $2d$ indices and conclude that $|\alpha_i| > \frac{1}{8d} \rho$ for all i with probability at least $3/4$.

Next, we need to upper- as well as lower- bound the norm of α . We do this separately. Note that:

$$\mathbb{E}[\|\alpha\|^2] = \|\alpha^*\|^2 + \mathbb{E} \left[\left\| \sum_{j \in [2d]} g_j \alpha^{(j)} \right\|^2 \right].$$

Since $\sigma_d(X) \geq \sigma_d(X_{-i}) \geq c$, we have that $\|\alpha^*\|^2 \leq \frac{1}{c^2}$. Thus, let us focus on the second term.

The second term can be bounded by observing that it can be written as $\|Ag\|^2$, where A is the $2d \times 2d$ matrix whose i th row has entries $\alpha_i^{(j)}$ (and 1 on the diagonal). Thus $\mathbb{E}[\|Ag\|^2] = \mathbb{E}[g^T A^T A g] = \|A\|_F^2 \leq 2d(1 + \frac{1}{c^2})\rho^2$, by our earlier discussion. From the choice of $\rho^2 = 1/2d$, this is exactly $(1 + \frac{1}{c^2})$.

Thus by a simple Markov bound, we have that $\|\alpha\|^2 \leq 8(1 + \frac{2}{c^2})$ with probability at least $7/8$. Finally, let us lower bound $\|\alpha\|^2$. While a weaker lower bound follows from the bounds on the coefficients, it is $\text{poly}(d)$ off, which will be problematic later. We thus do a more careful analysis, for which we move to the eigenspace of A . Suppose we write $A = \sum_{i=1}^{2d} \lambda_i u_i v_i^T$, where $\{u_i\}, \{v_i\}$ are the left and right singular vectors respectively. Suppose $\alpha^* = \sum_{i \in [2d]} \hat{\alpha}_i^* v_i$. We will also write $g = \sum_{i \in [2d]} \hat{g}_i v_i$. Then we have

$$\|\alpha^* + Ag\|^2 = \|\hat{\alpha} + U\hat{g}\|^2 = \sum_{i \in [2d]} \lambda_i^2 (\hat{\alpha}_i^* - \hat{g}_i)^2.$$

Since the Gaussian distribution is rotationally invariant, we have that the $\hat{g}_i$ are also $\sim \mathcal{N}(0, \rho^2)$, and are independent. Now, we are interested in the probability of the form $\Pr \left[\sum_{i \in [2d]} \lambda_i^2 (\hat{\alpha}_i^* - \hat{g}_i)^2 \leq \tau \right]$, for an appropriate τ . By definition, it is the Gaussian volume of the convex body

$$K(\hat{\alpha}) := \{z \in \mathbb{R}^{2d} : \sum_{i \in [2d]} \lambda_i^2 (\hat{\alpha}_i^* - z_i)^2 \leq \tau\}.$$

This is a symmetric convex body centered around $\hat{\alpha}$. Now by Anderson's theorem on Gaussian volume Anderson [1996], the Gaussian volume is smaller than the volume of $K(0)$ (i.e., the body centered around the origin).

Thus, for any τ ,

$$\Pr \left[\sum_{i \in [2d]} \lambda_i^2 (\hat{\alpha}_i^* - \hat{g}_i)^2 \leq \tau \right] \leq \Pr \left[\sum_{i \in [2d]} \lambda_i^2 \hat{g}_i^2 \leq \tau \right]. \quad (16)$$

We will also move to unit variance Gaussians $h_i = \hat{g}_i/\rho$, and focus on the probability:

$$\Pr \left[\sum_{i \in [2d]} \lambda_i^2 h_i^2 \leq \tau/\rho^2 \right].$$

We have $\mathbb{E}[\sum_{i \in [2d]} \lambda_i^2 h_i^2] = \sum_i \lambda_i^2 = \|A\|_F^2$. We have already observed that $2d \leq \|A\|_F^2 \leq 2d(1 + \frac{1}{c^2})$. Define $Y_i = h_i^2$, so that each Y_i is distributed as a unit chi-squared random variable. Also denote $L = \sum_i \lambda_i^2$. A well-known inequality (see, e.g., Lemma 1 of Laurent and Massart [2000]) is that for all constants $x > 0$,

$$\Pr \left[\sum_i \lambda_i^2 Y_i < L - 2x\sqrt{L} \right] < \exp(-x). \quad (17)$$

Setting $x = 1/2$, since $d \geq 1$ is an integer, the function $L - \sqrt{L}$ is increasing for $L \geq 2d$, and thus we have $\Pr[\sum_i \lambda_i^2 Y_i < 2d - \sqrt{2d}] < \exp(-0.5) \approx 0.61$. Simplifying some more, $\Pr[\sum_i \lambda_i^2 Y_i < 0.5d] < 0.61$. Since $\rho^2 = \frac{1}{2d}$, we have shown a bound in the case of $\tau = 1/4$ in (16).

Now, we observe that all the failure probabilities in the steps above add up to $0.61 + 1/4 + 1/8 < 1$. Thus, with probability $\Omega(1)$, properties (1), (2), (3) in the theorem all hold. This implies that there exists such an α , also, our proof also shows the existence of a randomized algorithm that succeeds with probability $\Omega(1)$ in finding such an α . Repeating this procedure $O(\log n)$ times to ensure success with high probability, the theorem follows.

C PROOF OF THEOREM 3.1

Algorithm 4: IterTrun($\{u_i\}_{i=1}^N, \mu_0, B, M, C$)

Input: $u_1, u_2, \dots, u_N \in \mathbb{R}$, not necessarily independent, with $u_i \sim N(\mu^*, \Sigma_i^2)$ (marginal), parameters B, μ_0, M, C such that $\Sigma_{(M)} \leq C$ for some constant $C > 0$ and $B \geq 2|\mu_0 - \mu^*|$.

```

1  $\delta^{(0)} \leftarrow CB$ ;
2  $\mu_0^{(0)} \leftarrow \mu_0$ ; // Initial estimate
3  $K \leftarrow \lfloor \log_2 B \rfloor$ ;
4  $T \leftarrow \lceil \frac{64N \ln N}{M} \rceil$ ;
5 for  $k = 0, \dots, K$  do
    // One stage
6   for  $t = 0, \dots, T$  do
    // One step within a stage
    // Truncate all points and average them
7      $\Delta_t^{(k)} = [\mu_t^{(k)} - \delta^{(k)}, \mu_t^{(k)} + \delta^{(k)}]$ ;
8      $\mu_{t+1}^{(k)} = \frac{1}{N} \sum_i \phi(u_i; \Delta_t^{(k)})$ ;
9   end
    // Initialize estimate for next stage
10   $\mu_0^{(k+1)} = \mu_{T+1}^{(k)}, \delta^{(k+1)} = \frac{\delta^{(k)}}{2}$ ;
11 end
Output:  $\hat{\mu} \leftarrow \mu_{T+1}^{(K)}$ .
```

We first state and prove some key results which we will use in our proof. We begin by re-using a lemma from Liang and Yuan [2020]:

Lemma C.1. *Let $\Delta := [\mu - \delta, \mu + \delta]$ be any interval. For any i , define $z_i(\mu, \delta) := \phi(u_i; \Delta) - \mu^*$. Also define $\delta_e(\mu) := |\mu - \mu^*|$. Then we have*

$$|\mathbb{E} z_i(\mu, \delta)| \leq \delta_e(\mu) \left(1 - \frac{1}{5} \frac{\delta}{\max(\delta_e(\mu), \delta)} \cdot \frac{\delta}{\max(\Sigma_i, \delta)} \right)$$

Since the lemma deals with a single index i , the fact that u_i 's are drawn from a joint distribution is not relevant. Thus the proof of Liang and Yuan [2020] applies. The next lemma is also based on an ϵ -net argument, applied to variables that satisfy the weak cancellation property.

Lemma C.2. *Let $\delta > 0$ be fixed, and $\{u_i\}_{i=1}^N$ be drawn from a joint distribution as in Theorem 3.1, satisfying the weak-cancellation property of (5) for any μ, δ , with parameter Γ . For any i , define $z_i(\mu, \delta) := \phi(u_i; [\mu - \delta, \mu + \delta]) - \mu^*$, and $\bar{z}_i(\mu, \delta) := z_i(\mu, \delta) - \mathbb{E} z_i(\mu, \delta)$. Let $\delta_e > 0$ be any parameter. Then, with probability at least $1 - \frac{4}{N}$, for every μ satisfying $|\mu - \mu^*| \leq \delta_e$, we have*

$$\left| \sum_i \bar{z}_i(\mu, \delta) \right| \lesssim \delta \sqrt{N\Gamma \ln N} + \frac{\delta_e}{N}. \quad (18)$$

The lemma basically states that we can go from having the weak-cancellation property for any given μ, δ , to one that holds for every μ (that satisfies some conditions). The proof is the same as in Liang and Yuan [2020], based on setting ϵ appropriately in the definition of weak cancellation, and taking a union bound over a net for μ .

Proof. As mentioned above, we use an ϵ -net argument. First, observe that $z_i(\mu, \delta)$ is 1-Lipschitz as a function of μ , for each $i \in [N]$. Set $\epsilon_0 = \frac{\delta_e}{N^2}$. Let $\mathcal{X}$ be an ϵ_0 -net over the interval $[\mu^* - \delta_e, \mu^* + \delta_e]$. Clearly, we will have $|\mathcal{X}| \leq 2N^2$. Now, fix some $\mu \in \mathcal{X}$. Then, we have

$$\mu - \delta - \mu^* \leq z_i(\mu, \delta) \leq \mu + \delta - \mu^*$$

where the above is true by the definition of $z_i(\mu, \delta)$. Taking expectation on all sides, we see that

$$\mu - \delta - \mu^* \leq \mathbb{E}z_i(\mu, \delta) \leq \mu + \delta - \mu^*$$

From the above two inequalities, we immediately see that

$$-2\delta \leq \bar{z}_i(\mu, \delta) \leq 2\delta \quad (19)$$

for all $i \in [N]$. So, we see that the random variables $\bar{z}_i(\mu, \delta)$ (over varying $i \in [N]$) are all bounded. However, note that the random variables $\bar{z}_i$ are not independent, and unlike Liang and Yuan [2020], we cannot apply standard concentration bounds like that of Hoeffding [1963]. To alleviate this, we now use the weak cancellation property. Particularly, for our fixed choice of μ, δ , by our assumption, we know that

$$\left| \sum_{i \in [N]} (\phi(u_i; [\mu - \delta, \mu + \delta]) - \mathbb{E}[\phi(u_i; [\mu - \delta, \mu + \delta])]) \right| \leq O(\delta \sqrt{N\Gamma \ln(1/\epsilon)})$$

with probability at least $1 - \epsilon$ for all $\epsilon > 0$. By adding and subtracting μ^* in each term on the LHS above, this implies that for all $\epsilon > 0$, with probability at least $1 - \epsilon$, we have

$$\left| \sum_{i \in [N]} (z_i(\mu, \delta) - \mathbb{E}z_i(\mu, \delta)) \right| \leq C'(\delta \sqrt{N\Gamma \ln(1/\epsilon)})$$

which can further be re-written as

$$\left| \sum_{i \in [N]} (\bar{z}_i(\mu, \delta)) \right| \leq C'(\delta \sqrt{N\Gamma \ln(1/\epsilon)})$$

for all $\epsilon > 0$ with probability at least $1 - \epsilon$. We pick $\epsilon = \frac{2}{N^3}$. So, with probability at least $1 - \frac{2}{N^3}$, we have that

$$\left| \sum_{i \in [N]} (\bar{z}_i(\mu, \delta)) \right| \leq C' \left(\delta \sqrt{N\Gamma \ln \left(\frac{N^3}{2} \right)} \right)$$

Taking a union bound over all the $\mu \in \mathcal{X}$ (note that there are at most $2N^2$ of them), we see that with probability at least $1 - \frac{4}{N}$, we have

$$\left| \sum_{i \in [N]} (\bar{z}_i(\mu, \delta)) \right| \leq C' \left(\delta \sqrt{N\Gamma \ln \left(\frac{N^3}{2} \right)} \right) \lesssim \delta \sqrt{N\Gamma \ln N}$$

Now, for any $\mu' \in [\mu^* - \delta_e, \mu^* + \delta_e]$ such that $\mu' \notin \mathcal{X}$, we know that there is a $\mu \in \mathcal{X}$ that satisfies $|\mu - \mu'| \leq \epsilon_0$. In that case, we have

$$\begin{aligned} \left| \sum_{i \in [N]} \bar{z}_i(\mu', \delta) \right| &\leq \left| \sum_{i \in [N]} \bar{z}_i(\mu, \delta) \right| + \left| \sum_{i \in [N]} \bar{z}_i(\mu', \delta) - \bar{z}_i(\mu, \delta) \right| \\ &\leq \left| \sum_{i \in [N]} \bar{z}_i(\mu, \delta) \right| + 2\epsilon_0 N \\ &\lesssim \delta \sqrt{N\Gamma \ln N} + 2\epsilon_0 N \end{aligned}$$

where in the second last step above, we used the 1-Lipschitz property of $z_i(\mu, \delta)$. Now, our choice was $\epsilon_0 = \frac{\delta_e}{N^2}$, and this completes the proof of the lemma. $\square$

We now return back to the proof of Theorem 3.1. We focus on one stage of Algorithm 4 (i.e., one iteration of the loop in line 7 of Algorithm 4). First, we define the following quantities:

$$e_t^{(k)} := |\mu_t^{(k)} - \mu^*| \quad (20)$$

$$z_i := \phi(u_i; \Delta_t^{(k)}) - \mu^* \quad (21)$$

$$\bar{z}_i := z_i - \mathbb{E}z_i \quad (22)$$

In words, $e_t^{(k)}$ denotes the error (from the true mean) of the estimate in step t of stage k . So, by the definition of the next iterate in the current stage (line 8 of Algorithm 4),

$$\mu_{t+1}^{(k)} - \mu^* = \frac{\sum_i (\phi(u_i; \Delta_t^{(k)})) - \mu^*}{N} = \frac{\sum_i z_i}{N}$$

So, in order to bound $|\mu_{t+1}^{(k)} - \mu^*|$, we would need to bound $|\sum_i z_i|$. Also, from (22), we see that

$$N e_{t+1}^{(k)} = \left| \sum_i z_i \right| \leq \left| \sum_i \bar{z}_i \right| + \sum_i |\mathbb{E}z_i|. \quad (23)$$

Lemmas C.1 and C.2 allow us to do an inductive analysis of the error $e_{t+1}^{(k)}$ in terms of $e_t^{(k)}$. So we prove the following lemma:

Lemma C.3. *If $\delta \geq e_t^{(k)}$, then with probability at least $1 - \frac{4}{N}$, we have*

$$e_{t+1}^{(k)} \lesssim 2\delta \sqrt{\frac{\Gamma \ln N}{N}} + e_t^{(k)} \left(1 - \frac{\delta}{5H_\delta^\Sigma} \right) \quad (24)$$

where H_δ^Σ is the harmonic mean of the quantities $\{\max(\Sigma_i, \delta)\}$, i.e., $H_\delta^\Sigma := \frac{N}{\sum_i \frac{1}{\max(\Sigma_i, \delta)}}$.

Proof. Using Lemma C.2 and C.1 and applying it to (23), we set $\delta_e = e_t^{(k)}$ and obtain that with probability at least $1 - \frac{4}{N}$,

$$\begin{aligned} N e_{t+1}^{(k)} &= \left| \sum_i z_i \right| \\ &\leq \left| \sum_i \bar{z}_i \right| + \sum_i |\mathbb{E}z_i| \\ &\lesssim \delta \sqrt{N \Gamma \ln N} + \frac{e_t^{(k)}}{N} + e_t^{(k)} \left(N - \frac{\delta}{5} \sum_i \frac{1}{\max(\Sigma_i, \delta)} \right) \end{aligned}$$

Dividing both sides of the above inequality by N , we get

$$e_{t+1}^{(k)} \lesssim \delta \sqrt{\frac{\Gamma \ln N}{N}} + \frac{e_t^{(k)}}{N^2} + e_t^{(k)} \left(1 - \frac{\delta}{5H_\delta^\Sigma} \right).$$

Finally, note that the second term is $\frac{e_t^{(k)}}{N^2}$, which is smaller than the first term for large enough N . So, bounding the second term by the first, the claim follows. $\square$

Finally, following Liang and Yuan [2020], we can complete the proof of Theorem 3.1. We will prove by induction that for all $k \leq K$,

$$e_0^{(k)} = |\mu_0^{(k)} - \mu^*| \leq \frac{\delta^{(k)}}{2}.$$

Note that the claim holds by construction for stage $k = 0$. Suppose the claim is true for stage k , where $k < K$. We need to show it for stage $k + 1$. Now, in stage k , our assumption implies that $\delta^{(k)} \geq 2e_t^{(k)}$ for $t = 0$.

We first show that *within* phase k , this condition holds for later time steps t as well. Suppose it is true for some step $t < T$, and we will show that it is true for $t + 1$. For any δ , define $m(\delta) := |\{i \in [N] : \Sigma_i \leq \delta\}|$. Then, observe that for any δ we have

$$H_\delta^\Sigma = \frac{N}{\sum_i \frac{1}{\max(\Sigma_i, \delta)}} \leq \frac{N}{m(\delta) \frac{1}{\delta}} = \frac{N\delta}{m(\delta)}.$$

Then by Lemma C.3 (we take $\delta = \delta^{(k)}$ in the lemma), we see that

$$\begin{aligned} e_{t+1}^{(k)} &\lesssim 2\delta^{(k)} \sqrt{\frac{\Gamma \ln N}{N}} + e_t^{(k)} \left(1 - \frac{m(\delta^{(k)})}{5N}\right) \\ &= e_t^{(k)} + \left(2\delta^{(k)} \sqrt{\frac{\Gamma \ln N}{N}} - e_t^{(k)} \frac{m(\delta^{(k)})}{5N}\right). \end{aligned}$$

Now, we have two cases. First, suppose that the second term on the RHS above is negative. In that case, we have $e_{t+1}^{(k)} \leq e_t^{(k)} \leq \frac{\delta^{(k)}}{2}$. On the other hand, if the second term is positive, then we must have that $e_t^{(k)} \leq \frac{10\delta^{(k)} \sqrt{N\Gamma \ln N}}{m(\delta^{(k)})}$. In that case, we will have that

$$\begin{aligned} e_{t+1}^{(k)} &\leq e_t^{(k)} + 2\delta^{(k)} \sqrt{\frac{\Gamma \ln N}{N}} \\ &\lesssim \delta^{(k)} \frac{\sqrt{N\Gamma \ln N}}{m(\delta^{(k)})} + \delta^{(k)} \sqrt{\frac{\Gamma \ln N}{N}} \\ &\lesssim \delta^{(k)} \frac{\sqrt{N\Gamma \ln N}}{m(\delta^{(k)})} \\ &\leq \delta^{(k)} \frac{\sqrt{N\Gamma \ln N}}{M} \\ &\lesssim \frac{\delta^{(k)}}{4}, \end{aligned}$$

where in the second last step we have used the fact that $m(\delta^{(k)}) \geq M$ by our assumption that $\Sigma_{(M)} \leq C$, since $\delta^{(k)} \geq C$ for all stages k (note that this holds because we halve $\delta^{(k)}$ after each stage and we have $\log B$ phases and that $\delta^{(0)} = CB$). Also, in the last step, we have used our assumption on M , namely that M is large enough such that $\sqrt{N\Gamma \ln N} \leq \frac{M}{4}$.

Hence, in either case, we can guarantee that $e_{t+1}^{(k)} \leq \frac{\delta^{(k)}}{2}$ for all $t \leq T$. So, we apply Lemma C.3 for all $t \leq T$ with the choice of $\delta = \delta^{(k)}$, and we get that

$$\begin{aligned} e_T^{(k)} &\lesssim \delta^{(k)} \sqrt{\frac{\Gamma \ln N}{N}} \sum_{i=0}^{T-1} \left(1 - \frac{m(\delta^{(k)})}{5N}\right)^i + \left(1 - \frac{m(\delta^{(k)})}{5N}\right)^T e_0^{(k)} \\ &= \delta^{(k)} \sqrt{\frac{\Gamma \ln N}{N}} \cdot \frac{5N}{m(\delta^{(k)})} \left(1 - \left(1 - \frac{m(\delta^{(k)})}{5N}\right)^T\right) + \left(1 - \frac{m(\delta^{(k)})}{5N}\right)^T e_0^{(k)} \\ &\lesssim \delta^{(k)} \frac{\sqrt{N\Gamma \ln N}}{m(\delta^{(k)})} \\ &\lesssim \delta^{(k)} \frac{\sqrt{N\Gamma \ln N}}{M} \\ &\lesssim \frac{\delta^{(k)}}{4} \end{aligned}$$

where above in the second step, we've simply expanded the geometric sum with common ratio $1 - \frac{m(\delta^{(k)})}{5N}$; in the third step we've used the fact that the term $\left(1 - \frac{m(\delta^{(k)})}{5N}\right)^T$ is small by our choice of T (line 4 of Algorithm 4), since

$$\left(1 - \frac{m(\delta^{(k)})}{5N}\right)^T \leq e^{-T \cdot \frac{m(\delta^{(k)})}{5N}} \approx e^{-64 \ln N \cdot \frac{m(\delta^{(k)})}{5M}} \leq e^{-\frac{64 \ln N}{5}} = o(1)$$

; in the fourth step, we've used the fact that $m(\delta^{(k)}) \geq M$ and in the final step we've used our assumption on M . Hence we get that $e_T^{(k)} \leq \frac{\delta^{(k)}}{4}$. In particular, this implies that $\delta^{(k+1)} = \frac{\delta^{(k)}}{2} \geq 2e_T^{(k)} = 2\left|\mu_0^{(k+1)} - \mu^*\right|$, which is exactly what we wanted to show by induction.

So, we have just shown that $\delta^{(k)} \geq 2\left|\mu_0^{(k)} - \mu^*\right|$ for all $k \leq K$. Now, we know that $C \geq \delta^{(K)}$ at the end of stage K , and hence we see that

$$\begin{aligned} e_T^{(K)} &\leq \delta^{(K)} \frac{\sqrt{\Gamma N \ln N}}{m(\delta^{(K)})} \\ &\leq C \frac{\sqrt{\Gamma N \ln N}}{M} \end{aligned}$$

completing the proof of the theorem.

D PROOF OF LEMMA 3.3

Let μ, δ be fixed, and we consider the interval $[\mu - \delta, \mu + \delta]$. Consider the projections $\phi(u_{\tau,r}; \Delta)$. Note that by definition these projections are bounded, i.e we have

$$\mu - \delta \leq \phi(u_{\tau,r}; \Delta) \leq \mu + \delta$$

Taking expectations, we get

$$\mu - \delta \leq \mathbb{E}\phi(u_{\tau,r}; \Delta) \leq \mu + \delta$$

From the above equations, we see that

$$-2\delta \leq \phi(u_{\tau,r}) - \mathbb{E}\phi(u_{\tau,r}) \leq 2\delta$$

So, the random variables $\phi(u_{\tau,r}) - \mathbb{E}\phi(u_{\tau,r})$ are bounded. But note that these random variables over $\tau \in \mathcal{T}_{\text{good}}$ are not independent. So, we use Janson's inequality Janson [2004]; particularly, we get that for any $t > 0$, we have

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{\tau \in \mathcal{T}_{\text{good}}} \phi(u_{\tau,r}) - \mathbb{E}\phi(u_{\tau,r})\right| \geq t\right) &\leq 2 \exp\left(-\frac{2t^2}{\binom{n}{2d-1} N (4\delta)^2}\right) \\ &= 2 \exp\left(-\frac{t^2}{8 \binom{n}{2d-1} N \delta^2}\right) \end{aligned}$$

Now, let $\epsilon > 0$ be given. We pick t such that $\epsilon \leq 2 \exp\left(-\frac{t^2}{8 \binom{n}{2d-1} N \delta^2}\right)$. This leads to the choice $t \leq C\delta \sqrt{\binom{n}{2d-1} N \ln \frac{1}{\epsilon}}$, where $C > 0$ is a positive constant. In other words, we have just shown that $\mathbb{P}\left(\left|\sum_{\tau \in \mathcal{T}_{\text{good}}} \phi(u_{\tau,r}) - \mathbb{E}\phi(u_{\tau,r})\right| \geq t\right) \leq C\delta \sqrt{\binom{n}{2d-1} N \ln \frac{1}{\epsilon}}$. So, it follows that the inputs $u_{\tau,r}$ satisfy the weak cancellation property of (5) with $\Gamma = \binom{n}{2d-1}$.

E PROOF OF LEMMA 4.1

As in the statement of the lemma, let $\tau = (i_1, \dots, i_{2d}) \in \mathcal{T}_{\text{good}}$ be any well-conditioned tuple such that $\sigma_{i_j}^2 \leq 1$ for all $j \in [2d]$. We will show that, with high probability, Algorithm 2 returns a set T with $|T| \geq \frac{9}{10}m$ for this choice of τ . This will prove the lemma.

To that end, fix any index k such that $\sigma_k^2 \leq 1$ and $k \notin \tau$. Consider the event that $k \notin T_{\text{candidate}}$, i.e the event that the condition in line 9 of Algorithm 2 fails. Note that from line 8, we have

$$y_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) y_{i_j}^{(1)} = \beta_1 + \eta_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) \eta_{i_j}^{(1)}$$

So, the event $k \notin T_{\text{candidate}}$ can be written as the event

$$\beta_1 + \eta_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) \eta_{i_j}^{(1)} \notin [u_\tau - 6 \log n, u_\tau + 6 \log n] \quad (25)$$

From lines 3-4 of Algorithm 2, we also know that

$$u_\tau = \beta_1 + \sum_{j=1}^{2d} \alpha_j(\tau) \eta_{i_j}^{(1)}$$

For convenience we define $u'_\tau := \sum_{j=1}^{2d} [\alpha_j(\tau) - \gamma_j(\tau)] \eta_{i_j}^{(1)}$ and define $(\sigma'_\tau)^2 := \sum_{j=1}^{2d} [\alpha_j(\tau) - \gamma_j(\tau)]^2 \sigma_{i_j}^2$. Now, we will use the well-conditioned property of τ ; particularly, note that the well-spread coefficients $\{\alpha_j\}_{j \in [2d]}$ are used to represent a vector of norm 1 (line 3 of Algorithm 2), and the coefficients $\{\gamma_j\}_{j \in [2d]}$ are used to represent a vector of norm at most 2 (line 8 of Algorithm 2; we also use the fact that $\|x_k^{(1)}\| \leq 1$ here). So, by the bound in point 2 of Theorem 2.4, we have

$$\begin{aligned} \sum_{j=1}^{2d} [\alpha_j(\tau) - \gamma_j(\tau)]^2 &\leq \sum_{j=1}^{2d} \alpha_j(\tau)^2 + \sum_{j=1}^{2d} \gamma_j(\tau)^2 \\ &\leq 5 \cdot 8 \cdot \left(1 + \frac{2}{c^2}\right) =: C_0 \end{aligned}$$

By our assumption on τ , we also know that $\sigma_{i_j}^2 \leq 1$ for all $j \in [2d]$. In particular, we see that $u'_\tau \sim N(0, (\sigma'_\tau)^2)$ with $(\sigma'_\tau)^2 \leq 5C_0$.

Now, observe that the event in equation (25) is equivalent to the event

$$\eta_k^{(1)} \notin [u'_\tau - 6 \log n, u'_\tau + 6 \log n] \quad (26)$$

Let's call the above event E . Now, we'll condition on the event $F := |u'_\tau| \leq 3 \log n$ (note that we can condition on u'_τ since $k \notin \tau$). So, we have

$$\begin{aligned} \mathbb{P}[E] &= \mathbb{P}(F) \mathbb{P}[E \mid F] + \mathbb{P}[F^c] \mathbb{P}[E \mid F^c] \\ &\leq \mathbb{P}[E \mid F] + \mathbb{P}[F^c] \end{aligned}$$

Let's first focus on $\mathbb{P}[E \mid F]$. By the definitions of E and F , we see that $\mathbb{P}[E \mid F]$ will be maximized when $u'_\tau = \pm 3 \log n$; we go with the choice $u'_\tau = 3 \log n$. Moreover, we know that $\eta_k^{(1)} \sim N(0, \sigma_k^2)$ where $\sigma_k \leq 1$ (by our assumption on k). Using all this, we see that

$$\begin{aligned} \mathbb{P}[E \mid F] &\leq 1 - \int_{-3 \log n}^{3 \log n} f_{N(0, \sigma_k^2)}(t) dt \\ &\leq 1 - \int_{-3 \log n}^{3 \log n} f_{N(0, \sigma_k^2)}(t) dt \\ &= \mathbb{P}\left[|\eta_k^{(1)}| \geq 3 \log n\right] \end{aligned}$$

where above $f_{N(0, \sigma_k^2)}$ is the density of $\eta_k^{(1)}$. Now, we have the following straightforward calculations:

$$\begin{aligned} \mathbb{P}\left[|\eta_k^{(1)}| \geq 3 \log n\right] &= \mathbb{P}\left[\left|\frac{\eta_k^{(1)}}{\sigma_k}\right| \geq \frac{3 \log n}{\sigma_k}\right] \\ &\leq \mathbb{P}\left[\left|\frac{\eta_k^{(1)}}{\sigma_k}\right| \geq 3 \log n\right] \end{aligned}$$

where above we simply used the fact that $\sigma_k \leq 1$. Now, observe that the last term above is a tail bound for a Gaussian with mean 0 and variance 1. So, by known tail bounds, we have

$$\mathbb{P} \left[\left| \frac{\eta_k^{(1)}}{\sigma_k} \right| \geq 3 \log n \right] \leq \exp \left(-\frac{9 \log^2 n}{2} \right)$$

So, from the above, we see that $\mathbb{P}[E \mid F] \leq \exp \left(-\frac{9 \log^2 n}{2} \right)$. Next, we focus on bounding $\mathbb{P}[F^c]$; we essentially repeat the same steps as above. By the definition of F , we have that

$$\begin{aligned} \mathbb{P}[F^c] &= \mathbb{P}[|u'_\tau| \geq 3 \log n] \\ &= \mathbb{P} \left[\left| \frac{u'_\tau}{\sigma'_\tau} \right| \geq \frac{3 \log n}{\sigma'_\tau} \right] \\ &\leq \mathbb{P} \left[\left| \frac{u'_\tau}{\sigma'_\tau} \right| \geq \frac{3 \log n}{\sqrt{5C_0}} \right] \end{aligned}$$

where in the last step above, we used the fact that $\sigma'_\tau \leq \sqrt{5C_0}$. In that case, again using the Gaussian tail bounds, we get

$$\mathbb{P} \left[\left| \frac{u'_\tau}{\sigma'_\tau} \right| \geq \frac{3 \log n}{\sqrt{5C_0}} \right] \leq \exp \left(-\frac{9 \log^2 n}{10C_0} \right)$$

and hence $\mathbb{P}[F^c] \leq \exp \left(-\frac{9 \log^2 n}{10C_0} \right)$. Combining everything, we see that

$$\mathbb{P}[E] \leq \exp \left(-\frac{9 \log^2 n}{2} \right) + \exp \left(-\frac{9 \log^2 n}{10C_0} \right) \quad (27)$$

So, we have just proven an upper bound for $\mathbb{P}[k \notin T_{\text{candidate}}]$. Taking a union bound over all such indices k such that $k \notin \tau$ (note that trivially there are at most $m \leq n$ of them), we see that with probability at least $1 - (m - 2d) \exp \left(-\frac{9 \log^2 n}{2} \right) - (m - 2d) \exp \left(-\frac{9 \log^2 n}{10C_0} \right)$, the event $k \in T_{\text{candidate}}$ is true for all such k . Clearly, the number of such k is $m - 2d \geq \frac{9}{10}m$, and this proves the lemma.

F PROOF OF LEMMA 4.2

First, note that the **for** loop in line 2 of Algorithm 2 iterates over all $2d$ -tuples $\tau \in \mathcal{T}_{\text{good}}$. There are at most $\binom{n}{2d}$ such tuples.

Now, fix such a well-conditioned tuple $\tau = (i_1, \dots, i_{2d}) \in \mathcal{T}_{\text{good}}$. Also fix some integer $\ell \in [0, L]$. We will specify a tail bound for the random variable n_ℓ . To that end, let $A_\ell := \{i \in [n] : \sigma_i \in \text{Range}_\ell\}$.

From line 7 of Algorithm 2, note that we only pick k which are not in τ . So, let $k \notin \tau$ be any index such that $k \in A_\ell$. Consider the event $k \in T_{\text{candidate}}$; note that this event is true if and only if the condition in line 9 of Algorithm 2 is true. So, from lines 8-9 of Algorithm 2, we have the following:

$$\mathbb{P}[k \in T_{\text{candidate}}] = \mathbb{P} \left[y_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) y_{i_j}^{(1)} \in I \right]$$

Now, observe that $I = [u_\tau - 6 \log n, u_\tau + 6 \log n]$. Also, similar to the proof of Lemma 4.1, we see that

$$y_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) y_{i_j}^{(1)} = \beta_1 + \eta_k^{(1)} + \sum_{j=1}^{2d} \gamma_j(\tau) \eta_{i_j}^{(1)}$$

For convenience, let us define $\eta_\tau := \sum_{j=1}^{2d} \gamma_j \eta_{i_j}^{(1)}$. In this case, we condition on the values of the random variables u_τ and η_τ , i.e we consider

$$\begin{aligned} \mathbb{P}[k \in T_{\text{candidate}} \mid u_\tau, \eta_\tau] &= \mathbb{P} \left[\beta_1 + \eta_k^{(1)} + \eta_\tau \in I \mid u_\tau, \eta_\tau \right] \\ &= \mathbb{P} \left[\eta_k^{(1)} \in I' \mid u_\tau, \eta_\tau \right] \end{aligned} \quad (28)$$

where above $I' := [(u_\tau - \eta_\tau - \beta_1) - 6 \log n, (u_\tau - \eta_\tau - \beta_1) + 6 \log n]$. Now, note that $\eta_k^{(1)} \sim N(0, \sigma_k^2)$, and the interval I' has size $12 \log n$; particularly, note that $\eta_k^{(1)}$ is normally distributed about 0. So, we see that the above conditional probability can be bounded by

$$\begin{aligned} \mathbb{P}[\eta_k^{(1)} \in I' \mid u_\tau, \eta_\tau] &\stackrel{(a)}{\leq} \mathbb{P}[\eta_k^{(1)} \in [-6 \log n, 6 \log n] \mid u_\tau, \eta_\tau] \\ &\stackrel{(b)}{=} \int_{-6 \log n}^{6 \log n} f_{N(0, \sigma_k^2)}(t) dt \\ &\leq 12 \log n \cdot f_{N(0, \sigma_k^2)}(0) \\ &= \frac{12 \log n}{\sigma_k \sqrt{2\pi}} \\ &\stackrel{(c)}{\leq} \frac{24 \log n}{\sigma_\ell \sqrt{2\pi}} \end{aligned}$$

where above in (a) we used the fact that $k \notin \tau$ and that $\eta_k^{(1)}$ is centered at 0 (hence the interval of size $12 \log n$ centered at 0 has the largest probability mass); in (b), $f_{N(0, \sigma_k^2)}$ is the density function of $\eta_k^{(1)}$, and in (c) we have used the fact that $\sigma_k \in \text{Range}_\ell$ and hence $\sigma_k \geq \frac{\sigma_\ell}{2}$. So, by the tower property of conditional expectations, plugging the above bound in (28) we get that

$$\mathbb{P}[k \in T_{\text{candidate}}] \leq \frac{24 \log n}{\sigma_\ell \sqrt{2\pi}} \quad (29)$$

Next, for each $k \in A_\ell \setminus \tau$, we define the random variable $X_k = \mathbb{1}_{k \in T_{\text{candidate}}}$. Note that the random variables $(X_k)_{k \in A_\ell \setminus \tau}$ are *independent*. Also, we can write $n_\ell = \sum_{k \in A_\ell \setminus \tau} X_k$. From (29), we see that

$$\begin{aligned} \mathbb{E}[n_\ell] &= \sum_{k \in A_\ell \setminus \tau} \mathbb{E}X_k \\ &\leq |A_\ell \setminus \tau| \frac{24 \log n}{\sigma_\ell \sqrt{2\pi}} \\ &\leq \frac{24n \log n}{\sigma_\ell \sqrt{2\pi}} \end{aligned}$$

where in the last step above we have used the trivial bound $|A_\ell \setminus \tau| \leq n$. Now, we'll apply Hoeffding's inequality to n_ℓ and get the following for $x = \frac{24n \log n}{\sigma_\ell \sqrt{2\pi}}$:

$$\begin{aligned} \mathbb{P}[n_\ell \geq 10x] &\stackrel{(a)}{\leq} \mathbb{P}[n_\ell \geq \mathbb{E}[n_\ell] + 9x] \\ &\stackrel{(b)}{\leq} \exp\left(-\frac{2 \cdot 9^2 \cdot x^2}{|A_\ell \setminus \tau|}\right) \\ &\stackrel{(c)}{\leq} \exp\left(-\frac{2 \cdot 9^2 \cdot x^2}{n}\right) \end{aligned}$$

where above in (a), we're simply using the fact that $\mathbb{E}[n_\ell] \leq x$; in (b), we're applying Hoeffding's bound, and in (c) we're using the fact that $|A_\ell \setminus \tau| \leq n$. Substituting the value of x in the above bound, we get that

$$\begin{aligned} \mathbb{P}\left[n_\ell \geq 10 \cdot \frac{24n \log n}{\sigma_\ell \sqrt{2\pi}}\right] &\leq \exp\left(-\frac{2 \cdot 9^2 \cdot (24)^2 n^2 \log^2 n}{2\pi \sigma_\ell^2 n}\right) \\ &\stackrel{(a)}{=} \exp\left(-\frac{K_0 n^2 \log^2 n}{\sigma_\ell^2 n}\right) \\ &\stackrel{(b)}{\leq} \exp\left(-\frac{K_0 n^2 \log^2 n}{n^{\frac{1}{2}} \cdot n}\right) \\ &= \exp\left(-K_0 n^{\frac{1}{2}} \log^2 n\right) \end{aligned} \quad (30)$$

where above in (a), we've simply reduced all constants to the constant K_0 , and in (b) we've used the fact that $\sigma_\ell^2 \leq \sigma_L^2 = n^{\frac{1}{2}}$.

Now, observe that the bound in (30) holds for a fixed integer $\ell \in [0, L]$ and a fixed $2d$ -tuple τ . Noting that the number of $2d$ -tuples is trivially bounded by n^{2d} and that the number of ℓ 's is at most $\frac{1}{4} \log n$, we can now take a union bound over all $2d$ -tuples τ and all $\ell \in [0, L]$ and get that with probability at least $1 - \frac{n^{2d} \log n}{4} \exp\left(-K_0 n^{\frac{1}{2}} \log^2 n\right)$, the claim of the lemma holds.

G PROOF OF LEMMA 4.4

We note that $w = 1$ (as in the statement of the lemma). Also, as in the main text, for any $u \in \mathbb{R}$, we define $K(u)$ to be the number of well-conditioned $\tau \in \binom{T}{2d}$ with indices in T such that $u_{\tau,r} \in [u - w, u + w]$. Then, note that for any $\Delta < w$, we have

$$\mu := \mathbb{E}[K(0) - K(\Delta)] = \sum_{\tau \in \mathcal{T}'} \mathbb{P}[u_{\tau,r} \in [-w, w]] - \mathbb{P}[u_{\tau,r} \in [-w + \Delta, w + \Delta]]$$

Now fix some well-conditioned $\tau \in \binom{T}{2d}$. We define:

$$P_{\tau,r}(\Delta) := \mathbb{P}[u_{\tau,r} \in [-w + \Delta, w + \Delta]]$$

Then, note that $P_{\tau,r}(\Delta)$ can be written as an integral:

$$P_{\tau,r}(\Delta) = \int_{-w+\Delta}^{w+\Delta} f_{N(0, \sigma_{\tau,r}^2)}(t) dt$$

where $f_{N(0, \sigma_{\tau,r}^2)}$ is the density of $u_{\tau,r}$. Using Taylor's theorem on $P_{\tau,r}$ about 0, we can write:

$$P_{\tau,r}(\Delta) = P_{\tau,r}(0) + P'_{\tau,r}(0)\Delta + \frac{P''_{\tau,r}(y_{\tau,r})\Delta^2}{2}$$

for some $y_{\tau,r} \in (0, \Delta)$. Now, note that

$$P'_{\tau,r}(x) = f_{N(0, \sigma_{\tau,r}^2)}(x + w) - f_{N(0, \sigma_{\tau,r}^2)}(x - w)$$

and hence we see that $P'_{\tau,r}(0) = 0$ (by the symmetry of the density function about 0). Also, for any $\theta \in \mathbb{R}$, we have that

$$f'_{N(0, \sigma_{\tau,r}^2)}(\theta) = -\frac{\theta}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\theta^2}{2\sigma_{\tau,r}^2}}$$

Now, we have

$$P'_{\tau,r}(x) = f'_{\tau,r}(x + w) - f'_{\tau,r}(x - w)$$

So, combining everything above, we can write the following:

$$\begin{aligned} P_{\tau,r}(0) - P_{\tau,r}(\Delta) &= -\frac{\Delta^2}{2} [f'_{\tau,r}(y_{\tau,r} + w) - f'_{\tau,r}(y_{\tau,r} - w)] \\ &= -\frac{\Delta^2}{2} \left[-\frac{y_{\tau,r} + w}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y_{\tau,r} + w)^2}{2\sigma_{\tau,r}^2}} - \frac{-(y_{\tau,r} - w)}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y_{\tau,r} - w)^2}{2\sigma_{\tau,r}^2}} \right] \\ &= \frac{\Delta^2}{2} \left[\frac{y_{\tau,r} + w}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y_{\tau,r} + w)^2}{2\sigma_{\tau,r}^2}} + \frac{(w - y_{\tau,r})}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(w - y_{\tau,r})^2}{2\sigma_{\tau,r}^2}} \right] \end{aligned}$$

Since $y_{\tau,r} < \Delta < w = 1$, we see that both of the terms on the RHS above are non-negative. Hence, we can lower-bound the LHS above by any of the terms on the RHS; so we choose the first term on the RHS and get the bound:

$$P_{\tau,r}(0) - P_{\tau,r}(\Delta) \geq \frac{\Delta^2}{2} \cdot \frac{y_{\tau,r} + w}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y_{\tau,r} + w)^2}{2\sigma_{\tau,r}^2}}$$

Now, observe that $y_{\tau,r} + w \geq 1$ (since $w = 1$ and $y_{\tau,r} > 0$). Also, we see that $y_{\tau,r} + w \leq 2$, and hence $e^{-\frac{(y_{\tau,r} + w)^2}{2\sigma_{\tau,r}^2}} \geq e^{-\frac{2^2}{2\sigma_{\tau,r}^2}} = e^{-\frac{2}{\sigma_{\tau,r}^2}}$. Hence, we see that

$$P_{\tau,r}(0) - P_{\tau,r}(\Delta) \geq \frac{\Delta^2}{2} \cdot \frac{1}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{2}{\sigma_{\tau,r}^2}}$$

Next, note that $\sigma_{\tau,r}^2 = \sum_{j=1}^{2d} \alpha_{j,r}(\tau)^2 \sigma_{i_j}^2$. Recall by our assumption that $\sigma_{i_j}^2 \geq \frac{1}{2}$ for each i_j (10), and that $\sum_{j=1}^{2d} \alpha_{j,r}(\tau)^2 \geq c_0$ since τ is a well-conditioned tuple; here, we're using the lower bound in part 2 of Theorem 2.4 with $c_0 = \frac{1}{4}$. Hence, we see that $\sigma_{\tau,r}^2 \geq \frac{1}{2} \cdot c_0$, and hence $e^{\frac{-2}{\sigma_{\tau,r}^2}} \geq e^{-\frac{4}{c_0}}$. Using this we get the lower bound:

$$\begin{aligned} P_{\tau,r}(0) - P_{\tau,r}(\Delta) &\geq \frac{\Delta^2}{2} \cdot \frac{1}{\sigma_{\tau,r}^3} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{4}{c_0}} \\ &= C_1 \cdot \frac{\Delta^2}{\sigma_{\tau,r}^3} \end{aligned}$$

where $C_1 := \frac{1}{2\sqrt{2\pi}} \cdot e^{-\frac{4}{c_0}}$. Importantly, note that C_1 is a positive constant that is independent of n, m, σ_i or d . Summing over all the well-conditioned $\tau \in \binom{T}{2d}$, we get that

$$\mathbb{E}[K(0) - K(\Delta)] \geq C_1 \sum_{\tau \in \mathcal{T}'} \frac{\Delta^2}{\sigma_{\tau,r}^3}$$

H PROOF OF LEMMA 4.5

As in the statement of the lemma, we have $w = 1$. Now, we will upper bound the variance of the $K(0) - K(\Delta)$. To that end, define $X_{\tau,r} := \mathbb{1}_{u_{\tau,r} \in [-w, w]} - \mathbb{1}_{u_{\tau,r} \in [-w+\Delta, w+\Delta]}$. So, we see that

$$\text{Var}[K(0) - K(\Delta)] = \text{Var} \left[\sum_{\tau \in \mathcal{T}'} X_{\tau,r} \right] \quad (31)$$

The issue with a direct variance calculation is that the $X_{\tau,r}$ s are dependent. To that end, we resort to the machinery developed by Janson [2004] to get deviation bounds for sums of dependent random variables. Particularly, define $S := \sum_{\tau \in \mathcal{T}'} \text{Var}(X_{\tau,r})$, and let χ be the chromatic number of the dependency graph defined by the $2d$ -tuples $\tau \in \mathcal{T}'$. From Theorem K.1 (which we prove in Appendix K), we see that $\text{Var} \left[\sum_{\tau \in \mathcal{T}'} X_{\tau,r} \right] \leq \chi \cdot S$. So, the only thing we need to figure out is an upper bound on S .

Note that

$$X_{\tau,r} = \mathbb{1}_{u_{\tau,r} \in [-w, -w+\Delta]} - \mathbb{1}_{u_{\tau,r} \in [w, w+\Delta]}$$

and so we have

$$\begin{aligned} \text{Var}(X_{\tau,r}) &\leq \mathbb{E}[X_{\tau,r}^2] \\ &= \mathbb{P}[u_{\tau,r} \in [-w, -w+\Delta]] + \mathbb{P}[u_{\tau,r} \in [w, w+\Delta]] \\ &= \int_{-w}^{-w+\Delta} f_{N(0, \sigma_{\tau,r}^2)}(t) dt + \int_w^{w+\Delta} f_{N(0, \sigma_{\tau,r}^2)}(t) dt \\ &\leq \Delta f_{N(0, \sigma_{\tau,r}^2)}(-w+\Delta) + \Delta f_{N(0, \sigma_{\tau,r}^2)}(w) \\ &\leq 2\Delta f_{N(0, \sigma_{\tau,r}^2)}(0) \\ &= \frac{2\Delta}{\sigma_{\tau,r} \sqrt{2\pi}} \\ &= C_2 \frac{\Delta}{\sigma_{\tau,r}} \end{aligned}$$

where again note that C_2 is independent of n, m, σ_i or d . So, summing over all the well-conditioned $\tau \in \binom{T}{2d}$ we get that

$$\text{Var}[K(0) - K(\Delta)] \leq C_2 \cdot \chi \cdot \sum_{\tau \in \mathcal{T}'} \frac{\Delta}{\sigma_{\tau,r}} \quad (32)$$

and this completes the proof of the lemma.

I PROOF OF LEMMA 4.6

We now show how to obtain a uniform bound that works for all $|u'| > \Delta$. So first, let u' be fixed such that $|u'| > \Delta$. Then, note that $\text{Diff}(u')$ is a sum of dependent random variables that takes values in the set $\{-1, 0, 1\}$; so, we can apply Janson's inequality [Janson, 2004] to $\text{Diff}(u')$ ¹; particularly, we get that for any $t > 0$, we have

$$\mathbb{P}[\text{Diff}(u') \leq \mu(u') - t] \leq \exp \left(-\frac{8t^2}{25 \cdot \Sigma(\mu') + 25 \cdot \chi \cdot \frac{4t}{3}} \right)$$

In the above, we put $t = \left(\frac{u'}{\Delta}\right)^2 \mu$. Doing so, we get

$$\begin{aligned} \mathbb{P}[\text{Diff}(u') \leq 0] &\stackrel{(a)}{\leq} \mathbb{P} \left[\text{Diff}(u') \leq \mu(u') - \left(\frac{u'}{\Delta}\right)^2 \mu \right] \\ &\leq \exp \left(-\frac{8 \left(\frac{u'}{\Delta}\right)^4 \mu^2}{25 \cdot \Sigma(\mu') + 25 \cdot \chi \cdot \frac{4 \left(\frac{u'}{\Delta}\right)^2 \mu}{3}} \right) \end{aligned}$$

where in (a) above we simply use the bound of Lemma 4.4. Now, further using the bound in Lemma 4.5, we get that

$$\mathbb{P}[\text{Diff}(u') \leq 0] \leq \exp \left(-\frac{8 \left(\frac{u'}{\Delta}\right)^4 \cdot \mu^2}{25 \cdot \left|\frac{u'}{\Delta}\right| \cdot \Sigma + 25 \cdot \chi \cdot \frac{4 \left(\frac{u'}{\Delta}\right)^2 \mu}{3}} \right)$$

Now, noting that $|\mu'| > \Delta$, and hence $\left(\frac{u'}{\Delta}\right)^2 \geq \left|\frac{u'}{\Delta}\right|$. Using this in the denominator above, we get that

$$\begin{aligned} \mathbb{P}[\text{Diff}(u') \leq 0] &\leq \exp \left(-\frac{8 \left(\frac{u'}{\Delta}\right)^4 \cdot \mu^2}{25 \cdot \left(\frac{u'}{\Delta}\right)^2 \cdot \Sigma + 25 \cdot \chi \cdot \frac{4 \left(\frac{u'}{\Delta}\right)^2 \mu}{3}} \right) \\ &= \exp \left(-\frac{8 \left(\frac{u'}{\Delta}\right)^2 \cdot \mu^2}{25 \cdot \Sigma + 25 \cdot \chi \cdot \frac{4\mu}{3}} \right) \end{aligned}$$

Finally, again noting that $\left(\frac{u'}{\Delta}\right)^2 \geq 1$, we get the bound

$$\begin{aligned} \mathbb{P}[\text{Diff}(u') \leq 0] &\leq \exp \left(-\frac{8 \cdot \mu^2}{25 \cdot \Sigma + 25 \cdot \chi \cdot \frac{4\mu}{3}} \right) \\ &= \exp \left(-S_0 \frac{\mu^2}{\Sigma + 4\chi\mu} \right) \end{aligned}$$

where above we just absorbed constants into $S_0 > 0$. Now, as noted before, since $\text{Diff}(u')$ takes at most $2|\mathcal{T}'|$ distinct values, noting that $2|\mathcal{T}'| \leq 2n^{2d}$, we can take a union bound over the $2|\mathcal{T}'|$ distinct values and see that the claim of the lemma follows.

J SOLVING THE SIGNAL-TO-NOISE INEQUALITY FOR ACHIEVING SUB-CONSTANT Δ

We now demonstrate how to solve (15) to obtain a sub-constant Δ . Particularly, we show that for a sub-constant Δ we have $\mu^2 \gtrsim O(d \ln n) \cdot \Sigma$ and that $\mu^2 \gtrsim O(d \ln n) \cdot \chi\mu$. First, let's focus on the inequality $\mu^2 \gtrsim O(d \ln n) \cdot \Sigma$. Using Lemma 4.4 and 4.5, our signal-to-noise inequality (15) becomes the following:

¹See Theorem 2.3 of Janson [2004]; particularly, we take $b = 2$.

$$\left(\sum_{\tau \in \mathcal{T}'} \frac{\Delta^2}{\sigma_{\tau,r}^3} \right)^2 \gtrsim O(d \ln n) \cdot \chi \cdot \sum_{\tau \in \mathcal{T}'} \frac{\Delta}{\sigma_{\tau,r}} \quad (33)$$

For any tuple $\tau \in \mathcal{T}'$, define ℓ_τ to be the maximum level of any index in τ (where levels are defined as in Lemma 4.2). Additionally, for any integer $\ell \in [0, L]$ let S_ℓ denote the set of all tuples $\tau \in \mathcal{T}'$ such that $\ell_\tau = \ell$.

Now let $\tau \in \mathcal{T}'$ be such that $\tau \in S_\ell$. In that case, from the definition of $\sigma_{\tau,r}$, we have that

$$\frac{c_d}{2} \cdot \sigma_\ell \stackrel{(a)}{\leq} \sigma_{\tau,r} \stackrel{(b)}{\leq} 8 \left(1 + \frac{2}{c_d^2} \right) \sigma_\ell \quad (34)$$

Inequality (a) above follows from the fact that all coefficients satisfy $|\alpha_{j,r}(\tau)| \geq c_d$, where $c_d := \frac{1}{8\sqrt{2}d^{1.5}}$ (see condition 3 of Theorem 2.4, and that $\ell_\tau = \ell$ by assumption (and hence $\sigma_{i_j}^2 \geq \frac{\sigma_\ell^2}{2}$ for at least one $i_j \in \tau$). Inequality (b) above simply follows from the well-conditioned property of τ and the fact that $\sigma_{i_j} \leq \sigma_\ell$ for all $i_j \in \tau$ (also, see condition 3 of Theorem 2.4).

Using this, we can lower-bound the LHS of (33) and upper bound the RHS of (33). Doing so, we get

$$\sum_{\tau \in \mathcal{T}'} \frac{\Delta^2}{\sigma_{\tau,r}^3} \geq \sum_{\ell=0}^L |S_\ell| \cdot \frac{\Delta^2}{\sigma_\ell^3}$$

and that

$$\sum_{\tau \in \mathcal{T}'} \frac{\Delta}{\sigma_{\tau,r}} \leq \sum_{\ell=0}^L \frac{2}{c_d} |S_\ell| \cdot \frac{\Delta}{\sigma_\ell}$$

So, to solve (33), it is enough to solve the inequality:

$$\left(\sum_{\ell=0}^L |S_\ell| \cdot \frac{\Delta^2}{\sigma_\ell^3} \right)^2 \gtrsim O(d \ln n) \cdot \chi \cdot \sum_{\ell=0}^L \frac{2}{c_d} |S_\ell| \cdot \frac{\Delta}{\sigma_\ell}$$

We simplify further; to solve the previous inequality, we square each term on the LHS, drop all the other terms and solve the resulting inequality:

$$\sum_{\ell=0}^L |S_\ell|^2 \cdot \frac{\Delta^4}{\sigma_\ell^6} \gtrsim O(d \ln n) \cdot \chi \cdot \sum_{\ell=0}^L \frac{2}{c_d} |S_\ell| \cdot \frac{\Delta}{\sigma_\ell}$$

We again drop the constants and only keep c_d . Recall that that c_d is dependent on d . We also cancel Δ from both sides, and the resultant inequality becomes:

$$c_d \sum_{\ell=0}^L |S_\ell|^2 \cdot \frac{\Delta^3}{\sigma_\ell^6} \gtrsim O(d \ln n) \cdot \chi \cdot \sum_{\ell=0}^L |S_\ell| \cdot \frac{1}{\sigma_\ell} \quad (35)$$

Now, we focus on the chromatic number χ . Note that since we're summing over well-conditioned tuples τ with indices in T , the chromatic number is bounded above by $\chi \leq \frac{\binom{|T|}{2d}}{\binom{|T|}{2d}} = \frac{2d \binom{|T|}{2d}}{\binom{|T|}{2d}}$ [Janson, 2004]. Also, by our definition of S_ℓ , we know that $\sum_{\ell=0}^L |S_\ell| = |\mathcal{T}'| \gtrsim \binom{|T|}{2d}$; so, we have the bound

$$\chi \lesssim \frac{4d \sum_{\ell=0}^L |S_\ell|}{|T|}$$

Plugging this in (35), we see that it is enough to solve the inequality

$$c_d \sum_{\ell=0}^L |S_\ell|^2 \cdot \frac{\Delta^3}{\sigma_\ell^6} \gtrsim O(d \ln n) \cdot \frac{4d \sum_{\ell=0}^L |S_\ell|}{|T|} \cdot \sum_{\ell=0}^L |S_\ell| \cdot \frac{1}{\sigma_\ell} \quad (36)$$

Now, let $i := \operatorname{argmax}_\ell |S_\ell|$. For any $\ell > i$, the ℓ th term on the RHS of (36) can be bounded as

$$\frac{|S_\ell|}{\sigma_\ell} \leq \frac{|S_i|}{\sigma_i}$$

where above we've used the fact that $\sigma_\ell > \sigma_i$. So, we see that $\sum_{\ell \geq i} \frac{|S_\ell|}{\sigma_\ell} \leq \frac{L|S_i|}{\sigma_i}$. Moreover, let $\ell^* := \operatorname{argmax}_{\ell < i} \frac{|S_\ell|}{\sigma_\ell}$. In that case, we have $\sum_{\ell < i} \frac{|S_\ell|}{\sigma_\ell} \leq \frac{L|S_{\ell^*}|}{\sigma_{\ell^*}}$. Also, note that $\frac{4d \sum_{\ell=0}^L |S_\ell|}{|T|} \leq \frac{4dL|S_i|}{|T|}$. Overall, the RHS of (36) can be bounded as

$$\text{RHS of (36)} \lesssim O(d \ln n) \cdot \frac{4dL|S_i|}{|T|} \cdot L \left(\frac{|S_{\ell^*}|}{\sigma_{\ell^*}} + \frac{|S_i|}{\sigma_i} \right)$$

So, in order to solve (36), it is enough to only keep the terms corresponding to ℓ^* and i on the LHS; in other words, it is enough to solve the following:

$$c_d \Delta^3 \left(\frac{|S_{\ell^*}|^2}{\sigma_{\ell^*}^6} + \frac{|S_i|^2}{\sigma_i^6} \right) \gtrsim O(d \ln n) \cdot \frac{4dL|S_i|}{|T|} \cdot L \left(\frac{|S_{\ell^*}|}{\sigma_{\ell^*}} + \frac{|S_i|}{\sigma_i} \right) \quad (37)$$

We now consider two cases:

1. In the first case, the first term on the RHS is dominant, i.e. $\frac{|S_{\ell^*}|}{\sigma_{\ell^*}} > \frac{|S_i|}{\sigma_i}$, which can be written as $|S_{\ell^*}| > \frac{\sigma_{\ell^*}|S_i|}{\sigma_i}$. So, in this case, by only keeping the first term of the LHS in (37), we see that the LHS is

$$> c_d \Delta^3 \frac{|S_{\ell^*}|}{\sigma_{\ell^*}^6} \cdot \frac{\sigma_{\ell^*}|S_i|}{\sigma_i}$$

So, it is enough to solve the inequality

$$c_d \Delta^3 \frac{|S_{\ell^*}|}{\sigma_{\ell^*}^6} \cdot \frac{\sigma_{\ell^*}|S_i|}{\sigma_i} \gtrsim O(d \ln n) \cdot \frac{4dL|S_i|}{|T|} \cdot 2L \cdot \frac{|S_{\ell^*}|}{\sigma_{\ell^*}}$$

Cancelling terms and simplifying, this inequality reduces to

$$\Delta^3 \gtrsim O(d \ln n) \cdot \frac{8dL^2 \sigma_{\ell^*}^4 \sigma_i}{c_d |T|}$$

Now, observe that $\sigma_{\ell^*} < \sigma_i$; so, it is enough to solve

$$\Delta^3 \gtrsim O(d \ln n) \frac{8dL^2 \sigma_i^5}{c_d |T|} \quad (38)$$

By Lemma 4.2, we know that (with high probability) $n_i \leq K_1 \frac{n \log n}{\sigma_i}$, which implies that $\sigma_i \leq \frac{n \log n}{n_i}$. From our choice of i , we also have that $\binom{|T|}{2d} \lesssim 2 \sum_{\ell=0}^L |S_\ell| \leq 2L |S_i|$. But, we know that $|S_i| \leq n_i \binom{\sum_{j=0}^i n_j}{2d-1} \leq n_i \binom{|T|}{2d-1}$ (this inequality simply follows by picking one index in level i and all other $2d-1$ indices in levels at most i), and this implies that

$$\binom{|T|}{2d} \lesssim 2L \cdot n_i \binom{|T|}{2d-1}$$

which upon simplifying gives us

$$n_i \gtrsim \frac{|T| - 2d + 1}{2dL}$$

In particular, we see that

$$\sigma_i \leq K_1 \frac{n \log n}{n_i} \leq K_1 \frac{n \log n \cdot 2dL}{|T| - 2d + 1}$$

Now, we know from Lemma 4.1 that with high probability, $|T| \geq \frac{9}{10} m \geq \frac{9}{10} n^{\frac{5}{6}}$, and hence for large enough n , we have $|T| - 2d + 1 \geq 0.9 |T|$ (so we can effectively ignore the $2d-1$ term). Plugging this in (38), we again get rid of all constants, and use the facts that $L = \frac{1}{4} \log n$ and $|T| \geq \frac{9}{10} m$ to get

$$\Delta^3 \gtrsim O(d \ln n) \frac{d^6 \log^{12} n \cdot n^5}{c_d m^6}$$

and this leads to $\Delta \geq O(d \ln n)^{\frac{1}{3}} \cdot \frac{d^2 \log^4 n}{c_d^{\frac{1}{3}}} \cdot \left(\frac{n^5}{m^6} \right)^{\frac{1}{3}}$, which is sub-constant as long as $m \geq n^{5/6}$.

2. In the second case, the right term on the RHS is dominant, i.e. $\frac{|S_{\ell^*}|}{\sigma_{\ell^*}} \leq \frac{|S_i|}{\sigma_i}$. In this case, we only keep the second term on the LHS in (37), and solve the resultant inequality, which is

$$c_d \Delta^3 \frac{|S_i|^2}{\sigma_i^6} \geq O(d \ln n) \frac{4dL |S_i|}{|T|} \cdot 2L \frac{|S_i|}{\sigma_i} \quad (39)$$

Simplifying the above, we get

$$\Delta^3 \geq O(d \ln n) \frac{8dL^2 \sigma_i^5}{d_d |T|} \quad (40)$$

From this point on, we follow the same steps as in case 1, and get the same error bound.

In any case, we see that the choice $\Delta \gtrsim O(d \ln n)^{1/3} \frac{d^2 \log^4 n}{c_d^{1/3}} \cdot \left(\frac{n^5}{m^6}\right)^{1/3}$ works. Note that $c_d \propto \frac{1}{d^{1.5}}$ (for reference, see condition 3 of Theorem 2.4), and hence this bound becomes (after ignoring log factors) $\Delta \gtrsim \tilde{O}\left(d^{2+\frac{1}{3}-\frac{1}{2}} \left(\frac{n^5}{m^6}\right)^{1/3}\right) = \tilde{O}\left(d^2 \left(\frac{n^5}{m^6}\right)^{1/3}\right)$. In particular, we've just shown that a choice of sub-constant Δ solves the first part of (15).

Next, we show that the same choice of Δ also solves the second part of inequality 15. So, recall that the inequality we want to solve is

$$\mu^2 \gtrsim O(d \ln n) \cdot \chi \mu$$

which is the same as the inequality $\mu \gtrsim O(d \ln n) \chi$. Again, we use the bound in Lemma 4.4, and we see that this inequality is equivalent to solving:

$$\sum_{\tau \in \mathcal{T}'} \frac{\Delta^2}{\sigma_{\tau,r}^3} \gtrsim O(d \ln n) \cdot \chi$$

As in the above discussion (J), we use the bound $\chi \lesssim \frac{4d \sum_{\ell=0}^L |S_\ell|}{|T|}$. So, we want to solve the inequality

$$\sum_{\tau \in \mathcal{T}'} \frac{\Delta^2}{\sigma_{\tau,r}^3} \gtrsim O(d \ln n) \cdot \frac{4d \sum_{\ell=0}^L |S_\ell|}{|T|} \quad (41)$$

Again, pick $i := \operatorname{argmax}_\ell |S_\ell|$. Then, the second factor in the RHS above can be bounded above by $\frac{4d \cdot L |S_i|}{|T|}$. Also, re-writing things in terms of levels on the LHS, our resultant inequality becomes:

$$\sum_{\ell=0}^L \frac{\Delta^2 |S_\ell|}{\sigma_\ell^3} \gtrsim O(d \ln n) \cdot \frac{4d \cdot L |S_i|}{|T|}$$

We only keep the i th term on the LHS, and solve:

$$\Delta^2 \cdot \frac{|S_i|}{\sigma_i^3} \geq O(d \ln n) \cdot \frac{4d \cdot L |S_i|}{|T|}$$

At this point, we can solve the above using the same argument as in the previous discussion. In fact, in this case we get a milder condition on Δ ; particularly, in this case we get $\Delta \gtrsim \tilde{O}\left(\operatorname{poly}(d) \left(\frac{n^3}{m^4}\right)^{1/2}\right)$. This completes the proof.

K VARIANCE BOUND FOR SUMS OF DEPENDENT RANDOM VARIABLES USING JANSON'S ARGUMENT

In this section, we will prove a simple variance bound for sums of dependent random variables using ideas from Janson [2004]. Janson's inequality Janson [2004] is a concentration bound for tail probabilities of sums of dependent random variables, where the degree of dependency is controlled by the *fractional chromatic number of the dependency graph*. For the sake of completeness, we reproduce the setup from Janson [2004]. Particularly, we have the following:

1. A set $\mathcal{A}$ of indices, and random variables $\{Y_\alpha\}_{\alpha \in \mathcal{A}}$ which are possibly dependent.
2. A *proper exact fractional cover* $\{(\mathcal{A}_j, w_j)\}_j$ of $\mathcal{A}$; this is a collection of pairs $(\mathcal{A}_j, w_j)$ where $\mathcal{A}_j \subseteq \mathcal{A}$ is *independent*, $w_j \in [0, 1]$ such that $\sum_{j: \alpha \in \mathcal{A}_j} w_j = 1$ for all $\alpha \in \mathcal{A}$.
3. $\chi^*(\mathcal{A}) := \min \sum_j w_j$, where the minimum is taken over all possible proper exact fractional covers of $\mathcal{A}$. $\chi^*(\mathcal{A})$ is called the *fractional chromatic number*.

We now prove our main variance bound.

Theorem K.1. *Let $X = \sum_{\alpha \in \mathcal{A}} Y_\alpha$, where $\mathcal{A}$ and Y_α are as defined above. Let $S := \sum_{\alpha \in \mathcal{A}} \text{Var}(Y_\alpha)$. Then*

$$\text{Var}(X) \leq \chi^*(\mathcal{A}) \cdot S \quad (42)$$

Proof. We proceed as in the proof of Theorem 2.1 of Janson [2004]. By subtracting the means, we assume without loss of generality that $\mathbb{E}Y_\alpha = 0$, and hence $\mathbb{E}X = 0$. Let $\{(\mathcal{A}_j, w_j)\}$ be an exact proper fractional cover of $\mathcal{A}$ such that $\sum_j w_j = \chi^*(\mathcal{A})$. Define $X_j := \sum_{\alpha \in \mathcal{A}_j} Y_\alpha$. From Lemma 3.1 of Janson [2004], we know that

$$X = \sum_j w_j X_j$$

So, we see that

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[X^2] && (\mathbb{E}X = 0 \text{ by assumption}) \\ &= \mathbb{E} \left[\left(\sum_j w_j X_j \right)^2 \right] \\ &= \mathbb{E} \left[\left(\sum_j \sqrt{w_j} \cdot \sqrt{w_j} X_j \right)^2 \right] \\ &\leq \mathbb{E} \left[\left(\sum_j w_j \right) \left(\sum_j w_j X_j^2 \right) \right] && (\text{Cauchy-Schwarz Inequality}) \\ &= \chi^*(\mathcal{A}) \sum_j w_j \mathbb{E}X_j^2 \end{aligned}$$

where in the last step we used $\sum_j w_j = \chi^*(\mathcal{A})$ and linearity of expectation. Next, note that because X_j is 0-mean for each j , we have that $\mathbb{E}X_j = \text{Var}(X_j)$. Next, observe that each X_j is a sum of *independent* random variables, and hence $\text{Var}X_j = \sum_{\alpha \in \mathcal{A}_j} \text{Var}(Y_\alpha)$. Combining all of this, we get that

$$\begin{aligned} \text{Var}(X) &\leq \chi^*(\mathcal{A}) \sum_j w_j \sum_{\alpha \in \mathcal{A}_j} \text{Var}(Y_\alpha) \\ &= \chi^*(\mathcal{A}) \sum_{\alpha \in \mathcal{A}} \text{Var}(Y_\alpha) \sum_{j: \alpha \in \mathcal{A}_j} w_j \\ &= \chi^*(\mathcal{A}) \sum_{\alpha \in \mathcal{A}} \text{Var}(Y_\alpha) \end{aligned}$$

where in the last step above we used the fact that $\sum_{j: \alpha \in \mathcal{A}_j} w_j = 1$ for all $\alpha \in \mathcal{A}$ (exact fractional covers). This proves the theorem. $\square$