On efficient robust regression with subquadratic samples

Deeksha Adil DEADIL@ETHZ.CH
ETH Zurich

Jarosław Błasiok JAROSLAW.BLASIOK@UNIBOCCONI.IT
Bocconi University

Hongjie Chen HONGJIE.CHEN@INF.ETHZ.CH
ETH Zurich

Deepak Narayanan Sridharan DSRIDHARAN@INF.ETHZ.CH
ETH Zurich

Editors: Steve Hanneke and Tor Lattimore

We revisit the problem of robust linear regression under Gaussian covariates with an unknown covariance matrix of condition number κ. For this fundamental problem, significant gaps remain in our understanding of the trade-offs among sample complexity, condition number, runtime, and prediction error for efficient algorithms. Our first result is a near-linear-time algorithm that uses $\widetilde{O}(d/\varepsilon^4)$ samples, where d is the dimension and ε is the corruption rate, and achieves prediction error $O(\sqrt{\varepsilon\kappa})$ under the condition $\varepsilon\kappa \lesssim 1$, improving over all prior works. We complement this result with a Statistical Query (SQ) lower bound showing that efficient SQ algorithms achieving error $o(\sqrt{\varepsilon\kappa})$ when $\varepsilon\kappa \lesssim 1$ require queries that take $\Omega(d^2)$ samples to simulate. Finally, we prove a low-degree polynomial lower bound that gives fine-grained evidence that, without assumptions such as $\varepsilon\kappa \lesssim 1$, efficient algorithms may require $\tilde{\Omega}\big(\min\{d\varepsilon^2\kappa^2,\ \varepsilon^2 d^2\}\big)$ samples to significantly outperform the trivial estimator that always guesses 0.

Keywords: linear regression, robust regression, information-computation tradeoffs, lower bounds

Contents

1. Introduction

Linear regression is one of the most fundamental tools in statistics, optimization, and machine learning; however, its standard formulations are extremely sensitive to corrupted observations Rousseeuw and Leroy (2003). Even a small fraction of outliers—whether in the labels or covariates—can arbitrarily bias the least squares estimator. This has motivated a long line of work in robust statistics Huber (1964); Hampel (1974); Rousseeuw (1984) on designing statistically efficient estimators that can effectively handle outliers.

We study robust linear regression under *strong* contamination. The strong contamination model imposes no structural or distributional assumptions on the corruptions; it only assumes that at most a small fraction of the dataset is corrupted (potentially adversarially). This makes it a flexible framework and subsumes several other commonly studied data corruption models, such as Huber contamination (see Diakonikolas and Kane (2023) for a detailed discussion). We now give the formal definition.

Definition 1 (Robust linear regression) *Given a regression vector $\beta \in \mathbb{R}^d$, let $D_{X,Y}$ be the joint distribution on $\mathbb{R}^d \times \mathbb{R}$ defined by the linear model*

$$Y = \langle X, \beta \rangle + \eta,$$

where $X \sim D_X$ is a mean-zero distribution with covariance Σ, and η denotes observation noise independent from X. We say $\{(X_i', Y_i')\}_{i=1}^n$ are ε-corrupted samples from $D_{X,Y}$ if $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} D_{X,Y}$ and $(X_i', Y_i') = (X_i, Y_i)$ for at least $(1-\varepsilon)n$ indices i. Given ε-corrupted samples, the goal is to output an estimate $\widehat{\beta}$ with small error $\|\widehat{\beta} - \beta\|_\Sigma$.[1]

While robust regression has been studied for several decades under a variety of settings (see e.g. Rousseeuw and Leroy (2003)), polynomial-time algorithms robust to strong contamination were developed only recently, beginning with the works of Klivans et al. (2018); Diakonikolas et al. (2019a,c); Prasad et al. (2020). Since the initial papers, a growing body of work has proposed polynomial-time algorithms under various distributional assumptions[2] Cherapanamjeri et al. (2020); Jambulapati et al. (2021); Bakshi and Prasad (2021); Zhu et al. (2022); Pensia et al. (2025). Despite this progress, even in the basic setting of Gaussian covariates with unknown covariance, the optimal trade-off between sample complexity, robustness, and prediction error—whether it can be achieved by polynomial-time algorithms—remains unknown.

Fast algorithm for Gaussian covariates. In this paper, we focus on this basic setting when X is Gaussian with unknown covariance Σ. Under this setting, Gao (2020) showed that given $O(d/\varepsilon^2)$ samples, one can in exponential time find an estimate achieving the information-theoretically optimal error $O(\sigma\varepsilon)$. The work by Diakonikolas et al. (2019c) gave a polynomial-time algorithm that uses $\widetilde{O}(d^2/\varepsilon^2)$ samples and achieves a near-optimal error of $O(\sigma\varepsilon\log(1/\varepsilon))$. Subsequently, one line of work has focused on designing *fast* robust regression algorithms under broader distributional assumptions that encompass Gaussians, such as bounded fourth-moment distributions Diakonikolas et al. (2019a); Cherapanamjeri et al. (2020); Jambulapati et al. (2021).

When specialized to Gaussian covariates, Cherapanamjeri et al. (2020) designed an algorithm that uses $n = \widetilde{O}(d/\varepsilon)$ samples, runs in time $\widetilde{O}(nd\kappa/\varepsilon^6)$, and achieves an error $O(\sigma\sqrt{\varepsilon\kappa})$, where κ is a condition number of the covariance matrix Σ. Note that with n samples in $\mathbb{R}^d$, the size of the input is nd, so in the well-conditioned case, this algorithm runs in near-linear time in its input. However, their algorithm requires the additional assumption $\varepsilon\kappa^2 \lesssim 1$ to succeed. Under the same assumption, Jambulapati et al. (2021) improved the runtime to $\widetilde{O}(nd\sqrt{\kappa})$ while maintaining the sample complexity and error guarantee. In the same work, Jambulapati et al. (2021) gave another algorithm with runtime $\widetilde{O}(nd\sqrt{\kappa}/\varepsilon)$, while achieving a smaller error $O(\sigma\sqrt{\varepsilon\kappa})$, under the *milder* assumption $\varepsilon\kappa \lesssim 1$, at the expense of increasing the sample complexity to $\widetilde{O}(d^2/\varepsilon^3 + d/\varepsilon^4)$. A summary of the above discussion is in Table 1.

Therefore, it remains widely open to determine what the optimal trade-off is among the condition number κ, corruption rate ε, sample complexity, and runtime. In particular, Jambulapati et al. (2021) asked the following question:

Does achieving error $O(\sigma\sqrt{\varepsilon\kappa})$ under the milder assumption of $\varepsilon\kappa \lesssim 1$ necessarily come at the cost of increased sample complexity for fast algorithms?

Our first result answers this question for Gaussian covariates with unknown covariance. We show that there is an algorithm with runtime $\widetilde{O}(nd\sqrt{\kappa}/\varepsilon)$ that achieves the sharper error bound of $O(\sigma\sqrt{\varepsilon\kappa})$ under the milder condition $\varepsilon\kappa \lesssim 1$, and requires only $\widetilde{O}(d/\varepsilon^4)$ samples.

1. This error is equivalent to prediction error under the covariate distribution, as $\mathbb{E}_{X\sim D_X} |\langle X, \beta\rangle - \langle X, \widehat{\beta}\rangle|^2 = \|\beta - \widehat{\beta}\|_\Sigma^2$, where $\|v\|_\Sigma^2 := v^T\Sigma v$.
2. Some form of distributional assumption on the covariates that is stronger than bounded covariance is information-theoretically necessary Bakshi and Prasad (2021).

Theorem 2 (Fast, robust linear regression for Gaussians) *Consider the setting in Definition 1 where $D_X = \mathcal{N}(0, \Sigma)$ with $\mu \cdot I_d \preceq \Sigma \preceq L \cdot I_d$, and the noise distribution has mean zero, variance σ^2, and fourth moment $O(\sigma^4)$. Let $\kappa := L/\mu$. Then there is an algorithm that, for any $\varepsilon \lesssim 1/\kappa$, given*

$$n = O\left(\frac{d \log \frac{d}{\varepsilon}}{\varepsilon^4} + \frac{d \log \frac{d}{\varepsilon^4}}{\varepsilon^2}\right)$$

ε-corrupted samples from D_{XY}, runs in time $\widetilde{O}(nd\sqrt{\kappa}/\varepsilon)$, and returns an estimate $\widehat{\beta}$ satisfying $\|\widehat{\beta} - \beta\|_\Sigma \leqslant O\big(\sigma\sqrt{\kappa\varepsilon}\big)$ with probability at least $9/10$.

We provide an overview of the proof of our result in Section 3.1, and present the formal proof in Appendix B.

Table 1: Selected algorithms for robust regression with Gaussian covariates and Gaussian noise

Paper	Covariance	Samples	Running Time	Error	Assumptions
Diakonikolas et al. (2019c)	Identity	$\widetilde{O}(d/\varepsilon^2)$	$\mathrm{poly}(d)$	$\sigma\varepsilon \log 1/\varepsilon$	—
Diakonikolas et al. (2019c)	Unknown	$\widetilde{O}(d^2/\varepsilon^2)$	$\mathrm{poly}(d)$	$\sigma\varepsilon \log 1/\varepsilon$	—
Anderson et al. (2025)	Unknown	$\widetilde{O}(d^2/\varepsilon^2)$	$\mathrm{poly}(d)$	$\sigma\varepsilon \log 1/\varepsilon$	—
Klivans et al. (2018)	Unknown	$\widetilde{O}(d^2/\varepsilon^2)$	$\mathrm{poly}(d)$	$\sigma\sqrt{\varepsilon}$	—
Bakshi and Prasad (2021)	Unknown	$\widetilde{O}(d^2/\varepsilon^2)$	$\mathrm{poly}(d)$	$\sigma\varepsilon^{3/4}$	—
Cherapanamjeri et al. (2020)	Unknown	$\widetilde{O}(d/\varepsilon)$	$\widetilde{O}(nd\kappa/\varepsilon^6)$	$\sigma\sqrt{\varepsilon}\kappa$	$\varepsilon\kappa^2 \lesssim 1$
Jambulapati et al. (2021)	Unknown	$\widetilde{O}(d/\varepsilon)$	$\widetilde{O}(nd\sqrt{\kappa})$	$\sigma\sqrt{\varepsilon}\kappa$	$\varepsilon\kappa^2 \lesssim 1$
Jambulapati et al. (2021)	Unknown	$\widetilde{O}(d^2/\varepsilon^3 + d/\varepsilon^4)$	$\widetilde{O}(nd\sqrt{\kappa}/\varepsilon)$	$\sigma\sqrt{\varepsilon\kappa}$	$\varepsilon\kappa \lesssim 1$
This work	Unknown	$\widetilde{O}(d/\varepsilon^4)$	$\widetilde{O}(nd\sqrt{\kappa}/\varepsilon)$	$\sigma\sqrt{\varepsilon\kappa}$	$\varepsilon\kappa \lesssim 1$

Impact of ill-conditioning on robust linear regression. Comparing the results in Table 1, we observe two common limitations of all linear-time algorithms that use $\widetilde{O}(d)$ samples:

1. For fixed noise level σ and corruption rate ε, the estimation error grows with the condition number κ.

2. The algorithms work only if the corruption rate is smaller than a threshold that depends on the condition number — otherwise, no meaningful guarantees are provided.

These two limitations are not inherent to the problem; there exist polynomial-time algorithms that use $\Omega(d^2)$ samples and avoid both of these drawbacks, e.g., Diakonikolas et al. (2019c); Anderson et al. (2025). Indeed, whenever an efficient robust covariance estimation algorithm is applicable with $n \gtrsim d^2$ samples, there is no need for the error rate or tolerated fraction of corruptions to degrade with the condition number. One way to see this is that we can first use $\Theta(d^2)$ samples to robustly estimate the covariance matrix Σ, and use it to reduce to a well-conditioned instance ($\kappa = O(1)$). We can then run a robust linear-regression algorithm with $\kappa = O(1)$ (see Appendix F).

We would like to point out an apparent parallel between the error term appearing in the algorithms in Table 1 and the state-of-the-art results for the related problem of *robust covariance-aware mean*

estimation. For the latter problem, when the distribution of the underlying random variable is ill-conditioned ($\kappa = \omega(1)$), the error term grows proportionally to $\sqrt{\varepsilon\kappa}$ when $n = \widetilde{O}(d)$ but an efficient algorithm is nevertheless applicable, even when ε is constant and κ superconstant (see Appendix G). *A priori* it is not unreasonable to hope for a similar behavior in robust linear regression.

Error rate. A natural question to ask is whether fast algorithms can achieve error $o(\sigma\sqrt{\varepsilon\kappa})$ when $\varepsilon\kappa \lesssim 1$ while using $\widetilde{O}(d)$ samples. We provide a negative answer to this question in the Statistical Query (SQ) model. At a high level, our lower bound shows that any efficient SQ algorithm achieving error $o(\sigma\sqrt{\varepsilon\kappa})$ must make either exponentially many queries, or at least one query which requires roughly $\Omega(d^2)$ samples to simulate (up to ε, κ factors).

Theorem 3 (SQ Lower Bound (informal, see Theorem 23)) *Let $\varepsilon\kappa \lesssim 1$. For robust linear regression with Gaussian covariates of unknown covariance Σ with condition number κ and unknown label noise variance $\sigma^2 \leqslant 1$, no efficient Statistical Query algorithm can achieve error $o(\sqrt{\varepsilon\kappa})$ on all instances. In particular, any such algorithm must either make at least $2^{d^{\Omega(1)}}$ queries, or make at least one query that requires $\Omega\big(d^2/\big(\sqrt{\kappa}e^{O(1/\varepsilon)}\big)\big)$ samples to be simulated.*

Our result can be viewed as a generalization of the SQ lower bound of Diakonikolas et al. (2019c) to the setting in which the condition number κ may be arbitrary. In particular, Diakonikolas et al. (2019c) show that when $\kappa = O(1)$, no efficient SQ algorithm can achieve prediction error $o(\sqrt{\varepsilon})$ unless it uses queries of accuracy that require $\Omega(d^2)$ samples to simulate. When $\kappa = O(1)$, our lower bound recovers their result as a special case, in both the achievable error and query tolerance. Similar to Diakonikolas et al. (2019c), our lower bound yields meaningful quantitative guarantees for $\varepsilon = \Omega(1/\log d)$, and mildly growing values of κ. We present an overview of our proof in Section 3.2 and complete formal proofs in Appendix C.

We next turn to a more fine-grained study of the trade-off between sample complexity and error guarantees. Building on the question posed by Jambulapati et al. (2021), we study our problem as the sample size varies from $n = \widetilde{O}(d)$ to $n = \widetilde{O}(d^2)$ (See Section 3.3 and Appendix D).

Corruption rate and condition number. As noted earlier, when $\widetilde{O}(d^2)$ samples are available, there exist efficient algorithms that achieve near-optimal prediction error $\widetilde{O}(\varepsilon)$ whenever ε is at most a sufficiently small constant, *regardless* of κ Diakonikolas et al. (2019c); Anderson et al. (2025). In contrast, known fast algorithms that use $\widetilde{O}(d)$ samples—including our result Theorem 2—impose much stronger restrictions on the corruption rate ε. Specifically, they all require ε to be smaller than a quantity that depends on κ, ranging from $\varepsilon \lesssim 1/\kappa^2$ to $\varepsilon \lesssim 1/\kappa$ for the stronger results.

At a high level, the κ-dependence comes from the fact that these methods avoid robustly preconditioning the covariates. Robust gradient descent methods, such as Cherapanamjeri et al. (2020), use robust mean estimation at the current iterate β_t on the corrupted gradient samples $(\langle X_i, \beta_t\rangle - y_i)X_i$, to estimate the population gradient $\Sigma(\beta_t - \beta)$. The robust mean estimation guarantee gives an additive error of size roughly $\sqrt{\varepsilon}(\|\beta_t - \beta\|_2)$ in Euclidean norm. However, the descent step needs this error to be small compared to the true gradient $\Sigma(\beta_t - \beta)$. After normalizing $\|\Sigma\|_2 \leqslant 1$, we have

$$\|\beta_t - \beta\|_2 \leqslant \|\Sigma^{-1}\Sigma(\beta_t - \beta)\| \leqslant \kappa\|\Sigma(\beta_t - \beta)\|_2.$$

Thus the relative error in the robust gradient estimate can be as large as $\sqrt{\varepsilon}\kappa$. For robust gradient descent to converge, it has to absorb the bias and can thus guarantee convergence under a condition of the form $\sqrt{\varepsilon}\kappa \lesssim 1$, equivalently $\varepsilon \lesssim 1/\kappa^2$. The natural geometry for robust regression is the

Mahalanobis norm induced by Σ. Inspired by the identifiability proof of Bakshi and Prasad (2021), Jambulapati et al. (2021) do their analysis directly in the Mahalanobis norm, and as a result are able to achieve a stronger condition of $\varepsilon \lesssim 1/\kappa$.

Several existing algorithms that avoid joint assumptions on ε and κ, either implicitly or explicitly, reduce the problem to a well-conditioned setting, i.e., to an instance with condition number $O(1)$, via a robust preconditioning step. This preconditioner is obtained via robust covariance estimation in the relative spectral norm (see, e.g., Kothari et al. (2018)). When $n = O(d^{2-\Omega(1)})$, Diakonikolas et al. (2017) showed that for $\varepsilon = \Omega(1/\log d)$, robust covariance estimation in this norm is computationally hard in the SQ model.

More recently, Diakonikolas et al. (2025b) gave an efficient robust covariance estimation algorithm using $\widetilde{O}\big(\varepsilon^2 d^{2+2\delta} + d^{1+\delta}\big)$ samples for any constant $\delta > 0$, which can be significantly below d^2 when ε is small (e.g. $\varepsilon \approx d^{-0.3}$). They further complemented this result with a low-degree lower bound, suggesting that $\tilde{\Omega}(\varepsilon^2 d^2)$ samples might be necessary for efficient covariance estimation in the relative spectral norm.

To summarize, existing efficient algorithms that achieve bounded prediction error without joint assumptions on ε and κ appear to require at least $\tilde{\Omega}(\varepsilon^2 d^2)$ samples. In contrast, existing efficient algorithms that use $\widetilde{O}(d)$ samples can only achieve bounded error when $\varepsilon\kappa \lesssim 1$. However, neither existing algorithmic results nor current computational lower bounds shed light on whether the joint dependence on ε and κ is inherent for efficient algorithms operating in the regime $n = o(\varepsilon^2 d^2)$. Indeed, without such an assumption, we are not aware of *any* efficient algorithm for Gaussian robust regression that provably outperforms the trivial estimator $\widehat{\beta} = 0$. This motivates the following question:

Is the condition $\varepsilon\kappa \lesssim 1$ merely an artifact of existing algorithmic techniques, or is it inherent to efficient algorithms in the low-sample regime $n = o(\varepsilon^2 d^2)$?

In this paper, we provide the first formal evidence that such a joint assumption may be necessary for efficient algorithms in the regime $n = o(\varepsilon^2 d^2)$. Specifically, we establish lower bounds against low-degree polynomial tests, showing that without a joint assumption on ε and κ, efficient algorithms for robust regression may be unable to outperform the trivial estimator significantly unless they use $\tilde{\Omega}(\min\{d\varepsilon^2\kappa^2, \varepsilon^2 d^2\})$ samples. We show this by a reduction from the following hypothesis testing problem with input $\{(X_i, Y_i)\}_{i=1}^n$ (see Problem 38 for a formal statement). The problem requires distinguishing between the following two distributions, the null H_0: $\{(X_i, Y_i)\}_{i=1}^n$ sampled i.i.d. from $\mathcal{N}(0, I_d) \times \mathcal{N}(0, \sigma_y^2)$, and the alternate H_1: For v sampled uniformly from the unit sphere, $\{(X_i, Y_i)\}_{i=1}^n$ sampled i.i.d. from $(1-\varepsilon)D_v(X, Y) + \varepsilon E_v(X, Y)$ where $E_v(X, Y)$ is arbitrary, and $D_v(X, Y)$ is the following linear model – $X \sim \mathcal{N}(0, \Sigma_v)$, $\eta \sim \mathcal{N}(0, \sigma^2)$, and $Y = \langle X, \beta_v \rangle + \eta$ where β_v, Σ_v depend on v. We remark that under the null hypothesis H_0, there is no linear relationship between X and Y, and the alternative hypothesis H_1 is an instance of robust linear regression under the weaker Huber contamination model.

Theorem 4 (Informal, see Theorem 41) *Let $\varepsilon \gg \frac{1}{\sqrt{d}}$, $\kappa \geqslant 1$ and $\varepsilon\kappa \geqslant C$ for some constant $C > 0$ sufficiently large. Then, there exists a choice of $\beta_v, \Sigma_v, E_v(X, Y), \sigma_y^2$ such that the null and alternative above are indistinguishable for polynomials of degree $\mathrm{poly}(\log n)$ for*

$$n \ll \frac{1}{\mathrm{poly}(\log d)} \min\big(d\varepsilon^2\kappa^2, \varepsilon^2 d^2\big).$$

On the other hand, we show in Corollary 62 that if the true parameter has signal strength $\|\Sigma^{1/2}\beta\| = \alpha$ for $\alpha = \Omega(1)$, an efficient regression algorithm achieving prediction error 0.1α solves the testing problem. This provides evidence that when $n \ll \min\big(d\varepsilon^2\kappa^2, \varepsilon^2 d^2\big)$, outperforming the trivial estimator by more than constant factors might be computationally hard whenever $\varepsilon\kappa \gg 1$.

A concrete setting of parameters for which this bound is easily interpretable is to focus on the case where ε, δ are small constants (for example 0.01), and κ is moderately growing with the dimension, e.g., $\kappa = d^{\delta}$. In this case, one could *a priori* hope for an algorithm using $\widetilde{O}(d)$ samples, and returning an estimate with error $\sqrt{\varepsilon\kappa} \approx d^{\delta/2}$. Our lower bound suggests that, unless one uses $\widetilde{\Omega}(d^{1+2\delta})$ samples, this problem might be computationally hard. Our computations in Appendix D illustrate that the hardness is driven both by ill-conditioning and the underlying linear dependence of X and Y.

We remark that when $\varepsilon\kappa = \omega(1)$ and $\kappa \leqslant \sqrt{d}$, our lower bound does not rule out the existence of a low-degree polynomial algorithm that uses $o(\varepsilon^2 d^2)$ samples and achieves small error. However, we are not aware of efficient algorithms that use $n = \widetilde{O}(d\varepsilon^2\kappa^2)$ samples and achieve error even of the order of $\sqrt{\varepsilon\kappa}$. We believe that it is an interesting open problem whether there exists an efficient algorithm matching our lower bounds in the regime where $\kappa \leqslant \sqrt{d}$. We describe our hard instance in detail in Section 3.3. Our lower bound also has consequences for *differentially private* regression. In the regime $n = O(d^{2-\Omega(1)})$, it suggests an information-computation gap for efficient private algorithms, consistent with the best-known efficient algorithms in this regime Brown et al. (2024). We refer to Appendix E for details.

2. Related Work

Robust regression. Beyond the algorithms described earlier, robust regression with unknown covariance has been studied under a variety of additional settings. Oliveira et al. (2022) studied it in this setting with noise that could depend on the covariates and obtained results that are closely comparable to those of Cherapanamjeri et al. (2020). Depersin (2020) studied the case of known covariance and possibly dependent noise, and designed a spectral algorithm with a sub-Gaussian error rate and $O(\sqrt{\varepsilon})$ error under bounded fourth moment assumptions. Pensia et al. (2025) studied the problem in the well-conditioned setting. An important line of work based on the Sum-of-Squares hierarchy studies robust regression under *certifiable* moment-bounded distributions Klivans et al. (2018); Bakshi and Prasad (2021); Zhu et al. (2022), achieving information-theoretically optimal error guarantees under broader distributional assumptions at the cost of increased sample complexity and runtime. Another widely studied direction considers milder corruption models in which the covariates remain uncorrupted while the responses are corrupted. The responses may be corrupted adaptively or non-adaptively, and a non-exhaustive list of papers that study this model includes Bhatia et al. (2015, 2017); Suggala et al. (2019); d'Orsi et al. (2021); Chen et al. (2022). In some of these settings, consistent estimation is information-theoretically possible even when the fraction of corrupted responses approaches one, since the covariates remain uncorrupted and thus preserve sufficient structure, which is in sharp contrast to the strong contamination model where any procedure has to incur error scaling with ε. There is also work studying robust regression with both covariate and response corruption under weaker models, such as Huber's contamination; see, e.g., Diakonikolas et al. (2023a).

Lower bounds for statistical problems. Standard approaches to proving computational lower bounds for average-case statistical problems establish hardness against restricted classes of algo-

rithms, such as Statistical Query algorithms Kearns (1998); Feldman et al. (2017), low-degree polynomials Hopkins and Steurer (2017); Hopkins et al. (2017); Hopkins (2018), which capture a large class of spectral methods, or the Sum-of-Squares hierarchy Barak et al. (2019). A complementary line of work, which has received recent attention, proves reduction-based hardness, similar to classical computational complexity theory, via reductions from problems believed to be computationally hard; see e.g. Brennan and Bresler (2020); Bruna et al. (2021).

Algorithmic robust statistics. Since the breakthrough works of Lai et al. (2016); Diakonikolas et al. (2019b) that designed efficient algorithms to robustly estimate the mean and covariance of a Gaussian, there has been a plethora of work that has designed efficient estimators for a large number of related problems such as moment estimation, clustering and regression Diakonikolas et al. (2018); Kothari et al. (2018); Hopkins and Li (2018); Klivans et al. (2018). Earlier works also studied efficient learning under outliers Klivans et al. (2009). We refer the reader to Diakonikolas and Kane (2023) for a comprehensive overview of the developments.

3. Technical Overview

In this section, we provide a detailed overview of our proof techniques. We begin by discussing our sample complexity result (Theorem 2), and then turn to our SQ lower bound (Theorem 3) and low-degree lower bound (Theorem 4).

3.1. Improved Sample Complexity for Gaussians

In this section, we outline the main ideas and techniques used to prove Theorem 2. Our algorithm is the same as that of Jambulapati et al. (2021), and our result follows by improving their analysis for the case of Gaussian distributions.

They propose an alternating algorithm that alternates between updating the regression parameter (ERM step), and removing outliers (filtering step). Specifically, the algorithm of Jambulapati et al. (2021) is able to achieve error $O(\sigma\sqrt{\varepsilon\kappa})$, improving on Cherapanamjeri et al. (2020)'s error of $O(\sigma\sqrt{\varepsilon}\kappa)$ under the milder condition $\varepsilon\kappa \lesssim 1$.

To achieve this smaller error, their algorithm requires stronger regularity conditions for the filtering step to succeed, necessitating d^2 samples. The following statement is the only bottleneck in their approach, which leads to $\Omega(d^2)$ sample complexity.

Lemma 5 *(Jambulapati et al., 2021, Lemma 19) Let $\varepsilon > 0$ be sufficiently small. Let $X_1, \dots, X_n$ be n samples from a 2-to-4 hypercontractive distribution $\mathcal{D}$ with parameter C and second moment Σ with $\mu I_d \preceq \Sigma \preceq L I_d$. Then, there exist universal constants $c, C_{est} > 0$ so that if*

$$n \geqslant c\left(\frac{d\log d}{\varepsilon^4} + \frac{d^2\log(d/\varepsilon)}{\varepsilon^3}\right),$$

then with probability at least 0.99, for every $u \in \mathbb{R}^d$, there exists a $G_u \subseteq [n]$ satisfying $|G_u| \geqslant (1-\varepsilon^2)n$, and

$$\left\| \frac{1}{|G_u|} \sum_{i \in G_u} \langle X_i, u\rangle^2 X_i X_i^\top \right\|_{\text{op}} \leqslant C_{est} L \|u\|_\Sigma^2 \,. \tag{1}$$

In this work, we prove a similar statement under a weaker sample complexity requirement, $n \gtrsim d \log d/\varepsilon^4$, when $\mathcal{D}$ is a Gaussian distribution (Theorem 9). To highlight our core ideas, we sketch the proof in the simpler case where $\Sigma = I$.

We proceed via the standard argument of applying a union bound over a δ-net. Our aim is to show that for a fixed u in the δ-net with $\|u\| = 1$, inequality (1) fails with probability $\exp(-\tilde{\Omega}(d))$. For such a fixed u, we choose the set G_u to roughly consist of all samples X_i for which $\|X_i\|^2 \lesssim O(d\varepsilon^{-2})$, and $\langle X_i, u\rangle^2 \leqslant O(1/\varepsilon)$. With this choice of G_u we first show that with high probability over the selection of X, for all u simultaneously, it holds that $|G_u| \geqslant (1-\varepsilon^2)n$ (Lemmas 10 and 11).

In order to prove the spectral norm bound (1), we again apply a net argument coupled with a union bound, i.e., for every fixed u, v (with unit norm), we are required to show that

$$\frac{1}{|G_u|}\sum_{i\in G_u}\langle X_i, u\rangle^2\langle X_i, v\rangle^2 \lesssim 1, \tag{2}$$

fails with probability at most $\exp(-\tilde{\Omega}(n))$. When $n \gtrsim d$, this allows us to take a union bound over all $\exp(\tilde{O}(d))$ points in the appropriate net over v's.

Since for every u, the set $|G_u|$ is large with high probability, Eq. (2) can be upper bounded as

$$\frac{1}{|G_u|}\sum_{i\in G_u}\langle X_i, u\rangle^2\langle X_i, v\rangle^2 \lesssim \frac{1}{n}\sum_{i\in[n]}\mathbb{1}[i\in G_u]\langle X_i, u\rangle^2\langle X_i, v\rangle^2. \tag{3}$$

We formally prove how to upper bound the right-hand side of the above inequality in the general covariance case in Lemma 12, which captures the core of our proof. Here, we proceed by decomposing $v = \alpha u + v'$, where $v' \perp u$. In general, we note that v' may depend on v, which adds to the complexity. By such a decomposition, we have essentially reduced the task of upper-bounding the right-hand side of (3) to the following two cases: either $v = u$ or $v \perp u$.

If we denote $Y_i := \mathbb{1}[i \in G_u]\langle X_i, u\rangle$, and $Z_i := \langle X_i, v'\rangle$, we need to show

$$\frac{1}{n}\sum Y_i^4 \lesssim 1, \quad \text{and} \quad \frac{1}{n}\sum Y_i^2 Z_i^2 \lesssim 1,$$

with probability $1 - \exp(-\tilde{\Omega}(n))$. In the above inequalities, Y_i and Z_i are independent[3], Z_i are standard Gaussians, and Y_i are 1-subgaussian and bounded by $1/\sqrt{\varepsilon}$ (by our choice of set G_u). Both of these statements now follow from standard Bernstein-type concentration inequalities, thus concluding the result. We refer the reader to Appendix B for a complete proof.

At a high level, our approach differs from that of Jambulapati et al. (2021) as follows. For a fixed u, they apply a matrix Chernoff bound to obtain Eq. (1), but as a result, the failure probability they obtain is only of the order $\exp(-\Omega(n/d))$. In order for this probability to be upper bounded by $\exp(-\tilde{\Omega}(d))$ (which is necessary to handle union bound over the net), they need to pick $n \gtrsim d^2$.

3.2. An Improved Statistical Query Lower Bound

In this section, we outline the main ideas and techniques used to prove Theorem 3. Our lower bound builds on the SQ lower bound for Gaussian robust regression established by Diakonikolas et al. (2019c), who show that any SQ algorithm achieving error $o(\sqrt{\varepsilon})$ must either make exponentially many queries or use tolerances so small that each query requires $n = \Omega(d^2)$ samples to simulate.

3. Independence holds *after* dropping the norm truncation indicator in Y_i and only keeping the u-projection truncation.

One of the most well-studied hard problems for the SQ model (and a basis of many hardness reductions in this area) is Non-Gaussian Component Analysis (NGCA). Informally, the goal in the NGCA task is to distinguish between the standard Gaussian and a distribution that is a known distribution A in some unknown hidden direction and the standard Gaussian in the orthogonal complement.

Diakonikolas et al. (2017) showed that for every distribution A that matches the first p moments of the Gaussian distribution, and does not have too large χ^2 divergence from that standard Gaussian, the related NGCA instance is hard for SQ algorithms (where the hardness depends on the number of moments matched). Often, after carefully selecting the distribution A to not only match the first few moments of a Gaussian, but also satisfy additional problem-specific properties, this NGCA problem can be used as a basis for reduction to show SQ hardness for other problems of interest.

In this work, following Diakonikolas et al. (2019c), we *do not* show the SQ hardness of the robust regression via a black-box reduction from NGCA. Instead, we construct a null distribution over $\mathbb{R}^d \times \mathbb{R}$, as well as a family of alternative distributions over $\mathbb{R}^d \times \mathbb{R}$ (one for each direction $v \in \mathcal{S}^{d-1}$), such that, if one had access to an accurate robust regression algorithm, one could easily distinguish the alternative from the null distribution. The proof strategy for indistinguishability in the SQ model is as follows.

- We carefully pick a (univariate) distribution R for the response variable y. The null distribution is just $\mathcal{N}(0, I) \times R$.
- For a given hidden direction v, and a value of response variable $y \in \mathbb{R}$, we prepare a conditional distribution of covariate X in the alternative case, as $\mathbb{P}[X|Y = y]$. This distribution resembles the one seen in the NGCA problem: on the hyperplane orthogonal to the hidden direction, the distribution of X is just the standard Gaussian. In the hidden direction v, it is a mixture between a Gaussian distribution and another distribution prepared such that the first three moments of the mixture match those of the standard Gaussian. The weights of the mixture depend on the value of y. The alternative distribution is then obtained by first sampling y from the same distribution R, and then X from the conditional distribution discussed.
- Integrating over y, we show that the distribution prepared this way is close in the TV-distance to a distribution $(X, \langle \beta, X\rangle + \eta)$ for Gaussian X, and β — a rescaling of v. As such, the prepared instance is a valid instance of a robust regression problem, and the hidden direction v can be recovered, if we have a good estimate of β.
- Since the conditional distribution of $\mathbb{P}[X|Y = y]$ has the same structure as in NGCA, proving the relevant correlation inequalities needed to establish SQ lower bound for distinguishing the null and alternative distribution can be reduced (by integrating over y) to the same inequalities already known for the NGCA instance[4].

The main technical difficulty is then crafting the conditional distribution $\mathbb{P}[X|Y = y]$. Consider first the joint distribution of (X', Y) in the uncorrupted regression case, i.e. when $X \sim \mathcal{N}(0, \Sigma_v)$, and $Y = \langle X, \beta\rangle + \eta$, where $\beta = \alpha\sqrt{\varepsilon}v$, $\Sigma_v = I_d - \gamma vv^T$ and σ^2 is chosen such that $\sigma_y^2 = 1$. In this

4. In Appendix C we prove hardness for the search version of the problem. NGCA hardness for the testing version is also well established (see e.g., (Diakonikolas and Kane, 2023, Proposition 8.14)) and has been used in regression settings Diakonikolas et al. (2025a). We use the testing version here for ease of exposition.

case, the conditional distribution is

$$[X|Y=y] \sim \mathcal{N}\big(\alpha'\sqrt{\varepsilon}yv, I_d - \gamma' vv^T\big),$$

where α' and γ' depend on α, γ. The conditional distribution of $[X|Y=y]$ which we attempt to create, should be a mixture of the Gaussian above with another "corruption" distribution (different from a standard Gaussian only in the direction parallel to v), such that the first three moments of the mixture match that of standard Gaussian, and the weight of the corruption distribution in the mixture is small for typical y.

In the work of Diakonikolas et al. (2019c), it was enough for them to pick α and γ some fixed constants smaller than 1, and as such the distribution of X (after projecting onto the non-trivial direction v) was simply $\mathcal{N}(c'\sqrt{\varepsilon}y, 2/3)$, whereas in our setting it is

$$\mathcal{N}(\underbrace{c\sqrt{\varepsilon}y}_{=:\mu_s}, \underbrace{1/\kappa - \bar{c}^2\varepsilon}_{=:\sigma_s^2}),$$

for some constants c and $\bar{c}$.

Since κ can be arbitrarily large, the variance σ_s^2 of the conditional distribution along v can be arbitrarily small, as opposed to just $2/3$ as in the case of instance constructed in Diakonikolas et al. (2019c). This qualitative difference necessitates a new moment-matching construction. We next provide an overview of this construction; full details appear in Section C.3. To complete the analysis, we additionally show that the resulting corrupted distribution can be realized as a Huber contamination of the joint distribution (X, Y) in Section C.4, and derive the required χ^2-divergence bounds in Section C.5. We omit these details here and focus on the moment-matching construction.

An important observation by Diakonikolas et al. (2019c) is that since $y \in \mathbb{R}$, and as a result μ_s can take any value, the corruption rate $\varepsilon_{\mu_s} \in (0, 1)$ must depend on μ_s. We split our construction into three regimes as opposed to the four regimes in Diakonikolas et al. (2019c):

(i) $|\mu_s| \leqslant O(\sqrt{\varepsilon})$: We adapt a recent construction of Diakonikolas et al. (2025b) (See Lemma 27) to our setting. Their construction applies to the regime of small $|\mu_s|$, and uses a mixture of four unit-variance Gaussians in the noise component, with component means of magnitude at most $O(1/\sqrt{\varepsilon})$.

(ii) $\Omega(\sqrt{\varepsilon}) \leqslant |\mu_s| \leqslant O(1)$: This is the most delicate regime to handle. In Diakonikolas et al. (2019c), for constant values of μ_s, the corruption rate ε_{μ_s} is small. If we adopt a similar relationship between μ_s and ε_{μ_s}, then the contribution of the signal component – whose variance is σ_s^2 and mixing weight is $(1 - \varepsilon_{\mu_s})$– to the second moment can become arbitrarily close to zero. At the same time, this component carries the largest mixing weight. A natural way to compensate is to increase the variance of the noise by introducing Gaussian components with variance $O(1/\varepsilon_{\mu_s})$. However, this approach is incompatible with controlling the χ^2-divergence with respect to the standard Gaussian, which is finite only when the variance remains bounded by a constant (see Fact 34). As a result, our construction requires introducing additional dependencies between μ_s and σ_s^2.

(iii) $|\mu_s| \geqslant \Omega(1)$: One can utilize a natural approach of considering a mixture of three Gaussians: one component with mean 0 and a large constant variance, and two symmetric Gaussian components with equal weights $(1 - \varepsilon_{\mu_s})$ and parameters (μ_s, σ_s^2) and $(-\mu_s, \sigma_s^2)$. While this

already suffices for our results, we provide a refined construction that maintains the variances of the noise components a constant instead of letting them be arbitrarily close to zero.

This concludes our construction generalizing Diakonikolas et al. (2019c) from the constant-condition-number setting $\kappa = O(1)$ to the general regime $\kappa = \Omega(1)$.

3.3. A New Information-Computation Trade-Off

We finally outline the ideas used in the proof of Theorem 4. We first present a natural reduction from NGCA to robust linear regression that isolates the source of hardness. This reduction, however, is insufficient for our main result and motivates the stronger construction that follows.

A first reduction. We begin with the following testing problem. Let $\mathcal{Q} \coloneqq \mathcal{N}(0, I_d)$ be the null hypothesis. To define the alternative $\mathcal{P}$, draw $v \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ and define $\mathcal{P} \coloneqq (1-\varepsilon)\mathcal{N}(0, I_d - vv^T) + \varepsilon E_v$, for a specific E_v. Given samples from either $\mathcal{P}$ or $\mathcal{Q}$, the goal is to efficiently distinguish between them. The distribution E_v can be chosen so that $\mathcal{P}$ matches the first three moments of $\mathcal{Q}$, for instance, E_v may be taken to have variance $1/\varepsilon$ in the direction v and the standard Gaussian in the orthogonal complement, yielding a special instance of the NGCA problem. For this instance, there is evidence against low-degree polynomials that indicates distinguishing requires $\tilde{\Omega}(\varepsilon^2 d^2)$ samples (Mao and Wein, 2025, Theorem 4.5). Our first step is to reduce this testing problem to robust linear regression. While related direct reductions from spiked models (these are essentially distributions of the type $\mathcal{N}(0, I + \theta vv^T)$) to regression are known, they have been intensely studied in the sparse setting Bresler et al. (2018); Brennan and Bresler (2020); Buhai et al. (2024); Kelner et al. (2024). We reduce the above NGCA instance to robust linear regression as follows. Let $Z \sim \mathcal{N}(0, I + \theta vv^T)$ for $\theta = -1$, and write $Z = (Y, X)$ where $Y = Z_1$ and $X = Z_{-1}$ (i.e., the vector obtained by dropping the first coordinate). Therefore,

$$X \sim \mathcal{N}\Big(0, I_{d-1} - v_{-1}v_{-1}^\top\Big),$$

and

$$Y \mid X = x \sim \mathcal{N}\left(-\frac{1}{v_1} \cdot \langle x, v_{-1}\rangle, 0\right).$$

Consequently, the distribution of Z admits two equivalent interpretations:

1. a negative spiked model $\mathcal{N}(0, I_d - vv^T)$,
2. a linear regression model with $X \sim \mathcal{N}\big(0, I_{d-1} - (1-1/\kappa)ww^\top\big)$, $Y = \langle X, \beta\rangle$, where $w = v_{-1}/\|v_{-1}\|$, $\kappa = 1/v_1^2$, and $\beta = -\frac{1}{v_1}v_{-1}$.

Routine calculation shows that any algorithm for Gaussian robust regression achieving non-trivial prediction error can solve the above testing problem by computing the norm of its estimator. As a result, by using the connection between spiked models and robust regression, we have shown evidence suggesting that unless an efficient algorithm uses $\tilde{\Omega}(\varepsilon^2 d^2)$ samples, it must incur at least a small constant prediction error.

We pause to make several useful observations about the above construction. (i) In contrast to our SQ lower bound, this reduction does not rely on any joint constraint — such as $\varepsilon\kappa \lesssim 1$ — between the corruption rate and the condition number. (ii) The signal strength $\|\beta\|_\Sigma$ is inherently limited, as it

is coupled to the coordinates of a random vector and the constraints that moment-matching imposes. (iii) Likewise, the condition number κ is tied to the coordinates of a random vector, restricting the construction to $\kappa = \Theta(d)$ in the typical case. (iv) Finally, the construction falls short of saying much about requiring joint assumptions concerning ε and κ.

Our lower bound instance. Building on the limitations of the previous construction, we make three key changes. First, we modify the null distribution so that the signal strength in the Mahalanobis norm can arbitrarily grow with the dimension. Second, instead of restricting to spiked models, we directly corrupt the joint distribution of (X, Y) in the Huber model by mixing with a $d+1$ dimensional Gaussian. Most importantly, we design the corruption to act in a two-dimensional subspace. Lifting to two dimensions allows us to explicitly encode the linear relationship between X and Y through the covariance, bringing out the coupling between ε and κ while preserving the $\tilde{\Omega}(\varepsilon^2 d^2)$ lower bound. We now describe the resulting testing problem and hard instance.

1. $\mathcal{Q}$: The joint distribution over (X, Y) where $X \sim \mathcal{N}(0, I_d)$ and $Y \sim \mathcal{N}(0, \sigma_y^2)$ are independent.

2. $\mathcal{P}$: $(1-\varepsilon) \cdot D_v(X, Y) + \varepsilon \cdot E_v(X, Y)$, where $v \sim \text{Unif}(\mathcal{S}^{d-1})$, $D_v(X, Y)$ is the distribution of (X, Y) for $Y = \langle X, \beta_v \rangle + \eta$, $X \sim \mathcal{N}(0, \Sigma_v)$, and $\eta \sim N(0, \sigma^2)$ independent of X.

For the hard instance, we set $\Sigma_v = I_d - (1 - 1/\kappa) vv^\top$, $\beta_v = \delta v$, and take $E_v(X, Y)$ to be a mean-zero Gaussian distribution in $\mathbb{R}^{d+1}$. Under this choice, the alternative distribution $\mathcal{P}$ is given by,

$$\mathcal{P}(X, Y) = (1-\varepsilon) \cdot \mathcal{N}\left(0, \begin{bmatrix} I_d - (1-1/\kappa) vv^T & \frac{\delta}{\kappa} v \\ \frac{\delta}{\kappa} v^T & \sigma_y^2 \end{bmatrix}\right) + \varepsilon \cdot \mathcal{N}(0, \Sigma_E).$$

We choose Σ_E so that the first three moments of $\mathcal{P}$ match those of the null distribution $\mathcal{Q}$. Therefore, we have Σ_E as

$$\begin{bmatrix} I_d + \frac{(1-\varepsilon)}{\varepsilon} \cdot (1 - 1/\kappa) vv^T & -\frac{(1-\varepsilon)}{\varepsilon} \frac{\delta}{\kappa} v \\ -\frac{(1-\varepsilon)}{\varepsilon} \frac{\delta}{\kappa} v^T & \sigma_y^2 \end{bmatrix}.$$

Observe that Σ_E is the identity matrix on the $(d-1)$-dimensional subspace orthogonal to both v and Y; all corruption is therefore confined to the two-dimensional subspace spanned by v and the last coordinate Y. This matrix is positive semidefinite if $\varepsilon\kappa \geqslant 1 - \varepsilon$ (Lemma 52).

We now examine the distribution $\mathcal{P}$ more closely. In the limit when $\kappa \to \infty$, the off-diagonal covariance terms vanish, and one can observe that distinguishing $\mathcal{P}$ and $\mathcal{Q}$ is equivalent to distinguishing the X part since the Y part is independent of X. Note that this is equivalent to the distinguishing problem in our first reduction as the X part in our alternative has the form $(1-\varepsilon)\mathcal{N}(0, I - vv^T) + \varepsilon E_v$. In the general case, the off-diagonal terms encode the correlation between X and Y. This correlation term plays a central role in our lower bound. After appropriate calculations, the diagonal covariance terms give rise to the $\tilde{\Omega}(\varepsilon^2 d^2)$ term, corresponding to the embedded NGCA hardness. In contrast, the off-diagonal block—responsible for the linear dependence between X and Y through β_v—yields the additional $\tilde{\Omega}(d\varepsilon^2\kappa^2)$ term. This latter contribution precisely provides evidence for the necessity of joint assumptions on ε and κ in robust regression when $n = o(\varepsilon^2 d^2)$ for algorithms that use $n = \widetilde{O}(d)$ samples. We note that this construction subsumes the earlier construction in the following sense: setting $\kappa = \Theta(d)$ in our construction yields a sample lower bound of $\tilde{\Omega}(\varepsilon^2 d^2)$.

We remark that a related construction appears in reduction-based lower bounds for robust sparse linear regression Brennan and Bresler (2020); however our setting and objective are different. To establish our results formally, we show in Theorem 41 that the advantage of degree-$O(\log n)$ tests is $1 + o(1)$ unless

$$n = \tilde{\Omega}\big(\min\{d\varepsilon^2\kappa^2, \varepsilon^2 d^2\}\big)$$

samples are used. Our analysis follows the approach of Mao and Wein (2025) by expressing the distinguishing advantage in terms of Hermite coefficients via a Fourier decomposition. Along the way, we prove a generalization of their result involving expectations of products of Hermite polynomials which may be of independent interest (see Lemma 49). The remaining arguments are combinatorial and we defer the computations to Appendix D.

4. Conclusions

We study robust linear regression under strong contamination with Gaussian covariates of unknown covariance, focusing on the tradeoffs between sample complexity, condition number, runtime and prediction error. We give a near-linear-time algorithm which requires $\widetilde{O}(d/\varepsilon^4)$ samples under mild conditions ($\varepsilon\kappa \lesssim 1$), with error $O(\sigma\sqrt{\varepsilon\kappa})$ for Gaussian covariates, improving upon prior works, and answering an open question of Jambulapati et al. (2021)—improved error guarantees for fast algorithms do not require increased sample complexity.

We complement this with a statistical query lower bound that provides evidence that when $\varepsilon\kappa \lesssim 1$, achieving error $o(\sigma\sqrt{\varepsilon\kappa})$ might be hard unless $\Omega(d^2/(\sqrt{\kappa}e^{O(1/\varepsilon)}))$ samples are used. We further provide a low-degree lower bound that gives evidence that without assumptions such as $\varepsilon\kappa \lesssim 1$, efficient algorithms may require $n = \tilde{\Omega}(\min\{d\varepsilon^2\kappa^2, \varepsilon^2 d^2\})$ samples to perform substantially better than the trivial estimator that always guesses zero. We leave open the problem of whether there is an efficient algorithm that uses $n = \tilde{O}(d\varepsilon^2\kappa^2)$ samples when $\kappa \leqslant \sqrt{d}$ and achieves non-trivial error whenever $\varepsilon\kappa \gg 1$.

Acknowledgments

We thank David Steurer for insightful discussions, and Stefan Tiegel for enlightening conversations regarding prior work. We are also thankful to the anonymous reviewers for their comments regarding presentation and pointing us to relevant prior work.

References

Prashanti Anderson, Ainesh Bakshi, Mahbod Majid, and Stefan Tiegel. Sample-optimal private regression in polynomial time. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2341–2349, 2025.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166. PMLR, 2018.

Matthew Brennan, Guy Bresler, Samuel B. Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In *Proceedings of the 34th Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 708–760. PMLR, 2021.

Guy Bresler, Sung Min Park, and Madalina Persu. Sparse pca from sparse linear regression. *Advances in Neural Information Processing Systems*, 31, 2018.

Gavin Brown, Jonathan Hayase, Samuel Hopkins, Weihao Kong, Xiyang Liu, Sewoong Oh, Juan C Perdomo, and Adam Smith. Insufficient statistics perturbation: Stable estimators for private least squares extended abstract. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 750–751. PMLR, 2024.

Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous lwe. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 694–707, 2021.

Rares-Darius Buhai, Jingqiu Ding, and Stefan Tiegel. Computational-statistical gaps for improper learning in sparse linear regression. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 752–771. PMLR, 2024.

Rares-Darius Buhai, Jun-Ting Hsieh, Aayush Jain, and Pravesh K Kothari. The quasi-polynomial low-degree conjecture is false. *arXiv preprint arXiv:2505.17360*, 2025.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 684–695. IEEE, 2022.

Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I. Jordan, Nicolas Flammarion, and Peter L. Bartlett. Optimal robust linear regression in nearly linear time, 2020. URL https://arxiv.org/abs/2007.08137.

Jules Depersin. A spectral algorithm for robust regression with subgaussian rates. *arXiv preprint arXiv:2007.06072*, 2020.

Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019a.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019b.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019c.

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Near-optimal algorithms for gaussians with huber contamination: Mean estimation and linear regression. *Advances in Neural Information Processing Systems*, 36:43384–43422, 2023a.

Ilias Diakonikolas, Daniel Kane, Lisheng Ren, and Yuxin Sun. Sq lower bounds for non-gaussian component analysis with weaker assumptions. *Advances in Neural Information Processing Systems*, 36:4199–4212, 2023b.

Ilias Diakonikolas, Chao Gao, Daniel Kane, John Lafferty, and Ankit Pensia. Information-computation tradeoffs for noiseless linear regression with oblivious contamination. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Ilias Diakonikolas, Samuel B Hopkins, Ankit Pensia, and Stefan Tiegel. Sos certificates for sparse singular values and their applications: Robust statistics, subspace distortion, and more. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1701–1709, 2025b.

Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32, 2019.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM Symposium on Theory of computing*, pages 371–380, 2009.

Tommaso d'Orsi, Gleb Novikov, and David Steurer. Consistent regression when oblivious outliers overwhelm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2): 1–37, 2017.

Chao Gao. Robust regression via multivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

Kristian Georgiev and Samuel Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in neural information processing systems*, 35:6829–6842, 2022.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Samuel Hopkins. *Statistical inference and the sum of squares method*. Cornell University, 2018.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. *Advances in Neural Information Processing Systems*, 34: 4475–4488, 2021.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Lasso with latents: Efficient estimation, covariate rescaling, and computational-statistical gaps. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2840–2886. PMLR, 2024.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.

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

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022.

Cheng Mao and Alexander S Wein. Optimal spectral recovery of a planted vector in a subspace. *Bernoulli*, 31(2):1114–1139, 2025.

Aaron Meurer, Christopher P Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K Moore, Sartaj Singh, et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Roberto I Oliveira, Zoraida F Rico, and Philip Thompson. A spectral least-squares-type method for heavy-tailed corrupted regression with unknown covariance\& heterogeneous noise. *arXiv preprint arXiv:2209.02856*, 2022.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association*, 120(550):1002–1013, 2025.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627, 2020.

Steven Roman. *The Umbral Calculus*. Academic Press, New York, 1984.

Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2003.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. See Theorem 6.8 (VC Inequality).

Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

Gábor Szegö. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1939.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Alexander S Wein. Computational complexity of statistics: New insights from low-degree polynomials. *arXiv preprint arXiv:2506.10748*, 2025.

Gian-Carlo Wick. The evaluation of the collision matrix. *Physical review*, 80(2):268, 1950.

Wikipedia contributors. Hermite polynomials. https://en.wikipedia.org/wiki/Hermite_polynomials, 2025. Accessed: 2026-01-07.

Inc. Wolfram Research. Mathematica, Version 14.2, 2024. Champaign, IL, 2024.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients. *Information and Inference: A Journal of the IMA*, 11(2):581–636, 2022.

Appendix A. Preliminaries

A.1. Organization

In Appendix B we present the proof of our improved analysis of Jambulapati et al. (2021) resulting in Theorem 2. In Appendix C we present the formal statement for our SQ lower bound and provide complete proofs. In Appendix D we present our low-degree lower bounds formally and provide all accompanying proofs, along with our reduction. In Appendix E we discuss the consequences of our low-degree lower bound in the context of differentially private regression. Appendix F discusses how preconditioning can get rid of the joint conditions on ε and κ and Appendix G discusses a fast algorithm for covariance-aware mean estimation. Finally Appendix H provides a link to a repository where we provide code for verifying computations that arise as part of our SQ lower bound.

A.2. Notation

We use the following convention: $\mathbb{N}$ is the set of natural numbers and $\mathbb{R}$ is the set of real numbers. $\mathbb{R}^d$ is the set of real vectors in d dimensions. For a positive integer n, $[n]$ is the set $\{1, 2, \ldots, n\}$. Unless explicitly stated, the base of the logarithm is e. Unless otherwise specified, all vector norms ($\|.\|$ or $\|.\|_2$) are the euclidean norm, and all matrix norms are the spectral norm. We use the notation $O(.), \Theta(.), \Omega(.), \lesssim, \gtrsim$ to hide absolute constants. We use $\mathbb{1}[.]$ for the indicator variable. We use $\widetilde{O}(.), \tilde{\Omega}$ to hide logarithmic factors. We denote the identity matrix in d dimensions by I_d. Let $A, B \in \mathbb{R}^{d\times d}$. Then $A \preccurlyeq B$ or $A \preceq B$ is the ordering of A and B in Löwner order. We use RHS and LHS to refer to the right and the left of an inequality respectively. We use $\mathrm{diag}(I_d, a)$ for some scalar a to denote the diagonal matrix in $\mathbb{R}^{d+1}$ that has 1 in the first d coordinates and a in the last coordinate. $f(n) \ll g(n)$ indicates that there exists a constant C such that $f(n) \leqslant g(n)/(\log n)^C$.

Appendix B. Improved Sample Complexity for Gaussian Distributions

As detailed in the technical overview, the core of our improvement is Lemma 12. Before we state and prove our improvement, we will require a few important facts that we will utilize in the rest of this section. We will first state them.

B.1. Concentration Inequalities

Fact 6 (Scalar Bernstein)

Suppose $X_1, X_2, \dots, X_n$ are drawn iid from a distribution with mean 0 on the real line and each $|X_i| \leqslant M$. Then we have for all positive t

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i > t\right] \leqslant \exp\left(-\frac{n^2t^2}{2\cdot\left(\sum_i \mathbb{E}[X_i^2] + \frac{1}{3}Mnt\right)}\right)$$

Fact 7 (Subexponential Tail Bounds Vershynin (2018)) *Let $S = \sum_{i=1}^n a_i X_i$, where each X_i is an independent chi-squared random variable with one degree of freedom, i.e., $X_i \sim \chi_1^2$, and $\{a_i\}_{i=1}^n$ are fixed weights. The key observation is that if we define the centered variable*

$$Y_i = X_i - \mathbb{E}[X_i] = X_i - 1,$$

then Y_i is a mean-zero subexponential random variable. Consequently, we can apply Bernstein's inequality to the centered sum:

$$S_c = S - \mathbb{E}[S] = \sum_{i=1}^{n} a_i(X_i - 1).$$

In particular, for any $t \geqslant 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i(X_i-1)\right| > t\right) \leqslant 2\exp\left(-c\cdot\min\left(\frac{t^2}{\sum_{i=1}^n a_i^2 V}, \frac{t}{\max_i |a_i| K}\right)\right)$$

where c is a universal positive constant, $V = Var(X_i)$ is the variance of a χ_1^2 random variable and K is a constant related to the sub-exponential norm of a centered χ_1^2 random variable.

Fact 8 *(VC Inequality)(Shalev-Shwartz and Ben-David (2014)) Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{D}$ be samples in $\mathbb{R}^d$. For each unit vector $u \in \mathbb{R}^d$, define*

$$S_{n,u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[\langle X_i, u\rangle^2 \geqslant \frac{20}{\varepsilon}\right], \qquad p_u = \mathbb{E}[S_{n,u}] = \mathbb{P}\left[\langle X, u\rangle^2 \geqslant \frac{20}{\varepsilon}\right].$$

Let

$$\mathcal{H} = \left\{ H_u = \{x \in \mathbb{R}^d : \langle x, u\rangle^2 \geqslant 20/\varepsilon\} \ : \ \|u\|_\Sigma = 1\right\}$$

be the corresponding concept class, and let $S_{\mathcal{H}}(n)$ denote its growth function. Then, for every $\alpha > 0$,

$$\mathbb{P}\left[\sup_u \left|S_{n,u} - p_u\right| > \alpha\right] \ \leqslant \ 8\, S_{\mathcal{H}}(n)\, e^{-n\alpha^2/32}.$$

We now state and prove our improved version of the result from Jambulapati et al. (2021).

Theorem 9 *Let $X_1, X_2, \dots X_n$ be n samples drawn from a d-dimensional Gaussian, $D_X = \mathcal{N}(0, \Sigma)$ where $\mu \cdot I_d \preceq \Sigma \preceq L \cdot I_d$. Let $\kappa := L/\mu$ and let $\varepsilon > 0$ be such that $\varepsilon\kappa \lesssim 1$. Then there exist universal constants $c, C_{est} > 0$ such that if*

$$n \geqslant c \cdot \left(\frac{d \log \frac{d}{\varepsilon}}{\varepsilon^4} + \frac{d \log \frac{d}{\varepsilon^4}}{\varepsilon^2} \right)$$

we have with probability at least 0.9 that for every $u \in \mathbb{R}^d$, there exists a $S_u \subseteq [n]$ satisfying $|S_u| \geqslant (1 - \varepsilon^2)n$, and

$$\left\| \frac{1}{|S_u|} \sum_{i \in S_u} \langle X_i, u \rangle^2 X_i X_i^T \right\|_{op} \leqslant C_{est} L \|u\|_\Sigma^2$$

We first prove the above result assuming Lemma 12, which we prove in Section B.3. The proof of Theorem 2 then follows similarly to Theorem 5 in Jambulapati et al. (2021), and we have included it in Section B.3.

B.2. Proof of Theorem 9

To prove our result, we first note that it suffices to prove for $\|u\|_\Sigma = 1$. We first prove that most samples have a small norm.

Lemma 10 *Let $X_1, \cdots, X_n$ be n samples drawn from the distribution D_X. If $n \geqslant \Omega(1/\varepsilon^2)$, then*

$$\mathbb{P}\left[\left| \left\{ i : \|X_i\|_2^2 \geqslant \frac{20Ld}{\varepsilon^2} \right\} \right| \leqslant \frac{n\varepsilon^2}{10} \right] \geqslant 0.99.$$

Proof. By Markov's inequality,

$$\mathop{\mathbb{P}}_{X \sim D_X}\left[\|X\|_2^2 \geqslant \frac{20Ld}{\varepsilon^2} \right] \leqslant \frac{\mathbb{E}_{X \sim D_X}\left[\|X\|^2\right]}{\frac{20Ld}{\varepsilon^2}} \leqslant \frac{\varepsilon^2}{20},$$

where in the last inequality, we used that $\mathbb{E}_{X \sim D_X}[\|X\|^2] \leqslant Ld$. Now, by an application of Scalar Bernstein (Fact 6) on the indicator random variable $\mathbb{1}\left[\|X_i\|_2^2 \geqslant \frac{20Ld}{\varepsilon^2}\right]$, we have with probability at least 0.999,

$$\left| \left\{ i : \|X_i\|_2^2 \geqslant \frac{20Ld}{\varepsilon^2} \right\} \right| \leqslant \frac{n\varepsilon^2}{10}$$

whenever $n = \Omega(1/\varepsilon^2)$, which concludes the proof.

□

We next show that along any arbitrary direction u, the one-dimensional projections of the samples cannot be large.

Lemma 11 *Let $X_1, \cdots, X_n$ be n samples drawn from the distribution D_X, and $u \in \mathbb{R}^d$ such that $\|u\|_\Sigma = 1$. Let H_u be defined as,*

$$H_u := \left\{ X \in \mathbb{R}^d : \langle X, u\rangle^2 \geqslant \frac{20}{\varepsilon} \right\}.$$

Then, if

$$n \geqslant \Omega\left(\frac{d \log \frac{d}{\varepsilon}}{\varepsilon^4} \right)$$

then we have

$$\mathbb{P}\left[\sup_u \frac{|i : X_i \in H_u|}{n} \leqslant \frac{\varepsilon^2}{50} \right] \geqslant 0.99.$$

Proof. For any vector u with $\|u\|_\Sigma = 1$, define H_u to be the following set:

$$H_u := \left\{ X \in \mathbb{R}^d : \langle X, u\rangle^2 \geqslant \frac{20}{\varepsilon} \right\}$$

Now for a fixed u, we have using Markov's inequality,

$$\mathop{\mathbb{P}}_{X\sim D_X}\left[\langle X, u\rangle^2 \geqslant \frac{20}{\varepsilon} \right] = \mathop{\mathbb{P}}_{X\sim D_X}\left[\langle X, u\rangle^4 \geqslant \frac{400}{\varepsilon^2} \right] \leqslant \frac{\varepsilon^2}{400} \cdot \mathop{\mathbb{E}}_{X\sim D_X}[\langle X, u\rangle^4]$$

Note that,

$$\mathop{\mathbb{E}}_{X\sim D_X}[\langle X, u\rangle^4] = \mathop{\mathbb{E}}_{Z\sim N(0,I_d)}[\langle Z, \Sigma^{1/2}u\rangle^4] = \mathop{\mathbb{E}}_{Y\sim N(0,1)}[Y^4] = 3$$

Therefore,

$$\mathop{\mathbb{P}}_{X\sim D_X}\left[\langle X, u\rangle^2 \geqslant \frac{20}{\varepsilon} \right] \leqslant \frac{3\varepsilon^2}{400} \leqslant \frac{\varepsilon^2}{100}$$

Next, we look at the fraction of samples that have a large univariate projection along the fixed direction u. Define,

$$S_{n,u} = \frac{1}{n} \sum_{i\in[n]} \mathbb{1}\left[\langle X_i, u\rangle^2 \geqslant \frac{20}{\varepsilon} \right]$$

We know that $\mathbb{E}[S_{n,u}] = \mathbb{P}\left[\langle X_i, u\rangle^2 \geqslant \frac{20}{\varepsilon}\right] \leqslant \frac{\varepsilon^2}{100}$. We will now be interested in

$$\mathbb{P}\left[\sup_u |S_{n,u} - \mathbb{E}[S_{n,u}]| > \frac{\varepsilon^2}{100} \right]$$

which by an application of the VC inequality (Fact 8) can be bounded as

$$\mathbb{P}\left[\sup_u |S_{n,u} - \mathbb{E}[S_{n,u}]| > \frac{\varepsilon^2}{100} \right] \leqslant 8 \cdot \mathcal{S}(H_u, n) \cdot e^{-\Omega(n\varepsilon^4)}$$

where $\mathcal{S}$ is the shatter coefficient or the growth function. From Sauer's Lemma (Sauer (1972)), we know that

$$\mathcal{S}(H_u, n) \leqslant (n+1)^{VC(H_u)}$$

where $\mathrm{VC}(H_u)$ is the VC-dimension of the concept class H_u. In our specific case, the set H_u is the union of two halfspaces. It is a well-known fact that union of finitely many halfspaces in $\mathbb{R}^d$ has VC dimension $O(d)$. Therefore we have that as long as

$$n = \Omega\left(\frac{d \log \frac{d}{\varepsilon}}{\varepsilon^4}\right) \geqslant \Omega\left(\frac{d \log n}{\varepsilon^4}\right)$$

with probability at least 0.999 that

$$\sup_u S_{n,u} = \sup_{H_u} \frac{|i : X_i \in H_u|}{n} \leqslant \frac{\varepsilon^2}{50},$$

concluding the proof. Note that since we required

$$\frac{n}{\log n} \gtrsim \frac{d}{\varepsilon^4}$$

it sufficed to take $n = \Omega\left(\frac{d \log \frac{d}{\varepsilon}}{\varepsilon^4}\right)$. □

The crux of our proof is the following lemma which we prove in Section B.3.

Lemma 12 *With probability at least* $1 - \exp\big(-\Omega\big(d \cdot \log\big(d/\varepsilon^4\big)\big)\big)$ *over the choice of* $\{X_i\}_{i=1}^n$*, we have for a fixed* u *with* $\|u\|_\Sigma = 1$ *that*

$$\sup_{v:\|v\|_2=1} \frac{1}{n} \sum_{i\in[n]} \langle X_i, u\rangle^2 \langle X_i, v\rangle^2 \mathbb{1}\left[\langle X_i, u\rangle^2 \leqslant \frac{20}{\varepsilon}\right] \mathbb{1}\left[\|X_i\|^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \leqslant L \cdot C''$$

for an absolute constant $C'' > 0$ *whenever* $n = \Omega\big(\frac{d}{\varepsilon^2} \cdot \log \frac{d}{\varepsilon^4}\big)$ *and* $\varepsilon\kappa \lesssim 1$.

We now prove Theorem 9.

Proof of Theorem 9.

We will use a γ-net argument to prove this theorem. Let $\mathcal{N}_\gamma$ denote a γ-net on the ellipsoid $\|u\|_\Sigma = 1$. We first claim that,

$$|\mathcal{N}_\gamma| \leqslant \left(\frac{3}{\gamma}\right)^d.$$

This follows since $\Phi : \mathcal{E} \to \mathcal{S}^{d-1}$ defined as $\Phi(u) = \Sigma^{1/2}u$ where $\mathcal{E} = \{u \in \mathbb{R}^d : \|u\|_\Sigma = 1\}$ is a bijection between the ellipsoid and the unit sphere. This implies that whenever $\|u - u'\|_\Sigma \leqslant \delta$, then correspondingly $\|\Phi(u) - \Phi(u')\|_2 \leqslant \delta$ as well since Φ is also an isometry between the metric spaces $(\mathbb{R}^d, \|.\|_\Sigma)$ and $(\mathbb{R}^d, \|.\|_2)$. Therefore, the number of elements in $\mathcal{E}$ is the same as the number of elements in a γ-net over a sphere.

We first claim that it is sufficient to prove the result only for $u \in \mathcal{N}_\gamma$ when $\gamma = \Theta(\varepsilon^5/d^2)$. To prove this we assume the result holds for all $u \in \mathcal{N}_\gamma$ and show that it then holds for all u satisfying $\|u\|_\Sigma = 1$. To this end, we first define the set for all $u \in \mathbb{R}^d$, $\|u\|_\Sigma = 1$,

$$G_u := \left\{ i : X_i \notin H_u \text{ and } \|X_i\|_2^2 \leqslant \frac{20Ld}{\varepsilon^2} \right\}.$$

Now, assume that, for every $v \in \mathcal{N}_\gamma$, $S_v = G_v \subseteq [n]$ satisfying $|S_v| \geqslant (1-\varepsilon^2)n$, and

$$\left\| \frac{1}{|S_v|} \sum_{i \in S_v} \langle X_i, v\rangle^2 X_i X_i^T \right\|_{\text{op}} \leqslant C_{est} L.$$

Let $u \notin \mathcal{N}_\gamma$, $\|u\|_\Sigma = 1$, and let $u' \in \mathcal{N}_\gamma$ denote the point on the net closest to u, and define,

$$S_u := G_u \cap G_{u'}.$$

Also, define the unnormalized sum,

$$F(u) := \sum_{i \in S_u} \langle X_i, u\rangle^2 X_i X_i^T,$$

and note that we have assumed $\|F(u')\|_{\text{op}} \leqslant C_{est} L |S_{u'}| \leqslant C_{est} Ln$. Now, from triangle inequality

$$\|F(u)\|_{\text{op}} = \big\|F(u) - F(u') + F(u')\big\|_{\text{op}} \leqslant \big\|F(u) - F(u')\big\|_{\text{op}} + \big\|F(u')\big\|_{\text{op}}.$$

Therefore as long as

$$\big\|F(u) - F(u')\big\|_{\text{op}} \leqslant O(nL)$$

we get that $\|F(u)\|_{\text{op}} \leqslant O(nL)$ and are done. For the remainder of the proof we will use $\|\cdot\|$ to represent the operator norm. Now,

$$\begin{aligned}
\big\|F(u) - F(u')\big\| &= \left\| \sum_{i \in S_u} \langle X_i, u\rangle^2 X_i X_i^T - \sum_{i \in S_{u'}} \langle X_i, u'\rangle^2 X_i X_i^T \right\| \\
&\overset{(i)}{=} \left\| \sum_{i \in S_u} \big(\langle X_i, u\rangle^2 - \langle X_i, u'\rangle^2\big) X_i X_i^T - \sum_{i \in S_{u'} \setminus S_u} \langle X_i, u'\rangle^2 X_i X_i^T \right\| \\
&\overset{(ii)}{\leqslant} \left\| \sum_{i \in S_u} \big(\langle X_i, u\rangle^2 - \langle X_i, u'\rangle^2\big) X_i X_i^T \right\| + \left\| \sum_{i \in S_{u'} \setminus S_u} \langle X_i, u'\rangle^2 X_i X_i^T \right\| \\
&\overset{(iii)}{\leqslant} \left\| \sum_{i \in S_u} \big(\langle X_i, u\rangle^2 - \langle X_i, u'\rangle^2\big) X_i X_i^T \right\| + \left\| \sum_{i \in S_{u'}} \langle X_i, u'\rangle^2 X_i X_i^T \right\| \\
&\overset{(iv)}{\leqslant} \left\| \sum_{i \in S_u} \big(\langle X_i, u\rangle^2 - \langle X_i, u'\rangle^2\big) X_i X_i^T \right\| + O(nL).
\end{aligned}$$

In the above (i) follows from the definition of S_u and $S_{u'}$, (ii) is triangle inequality, (iii) is because we are adding more PSD matrices, and (iv) follows from the bound over the elements of the net. Now, again from triangle inequality,

$$\left\|\sum_{i\in S_u}\left(\langle X_i,u\rangle^2-\langle X_i,u'\rangle^2\right)X_iX_i^T\right\|\leqslant\sum_{i\in S_u}\left\|\left(\langle X_i,u\rangle^2-\langle X_i,u'\rangle^2\right)X_iX_i^T\right\|.$$

Consider each term in the sum,

$$\begin{aligned}\left\|\left(\langle X_i,u\rangle^2-\langle X_i,u'\rangle^2\right)X_iX_i^T\right\|&=\left\|\langle X_i,u-u'\rangle\langle X_i,u+u'\rangle X_iX_i^T\right\|\leqslant\|X_i\|_2^4\cdot\left\|u-u'\right\|_2\cdot\left\|u+u'\right\|_2\\&\leqslant\|X_i\|_2^4\cdot\|\Sigma^{-1/2}\|_2^2\cdot\|\Sigma^{1/2}(u-u')\|_2\cdot\|\Sigma^{1/2}\left(u+u'\right)\|_2\\&\leqslant\frac{800L^2d^2}{\varepsilon^4}\cdot\|\Sigma^{-1/2}\|_2^2\cdot\gamma=\frac{800Ld^2\kappa\gamma}{\varepsilon^4}.\end{aligned}$$

Since $\gamma=\Theta(\varepsilon^5/d^2)$ is small enough, and $\varepsilon\kappa<1$, we can bound the above by $O(L)$, and we have shown

$$\left\|F(u)-F(u')\right\|\leqslant O(n(L+L))=O(nL).$$

Since F was unnormalized (instead of average we took the sum), we get after normalization,

$$\frac{1}{n}\left\|F(u)-F(u')\right\|\leqslant O(L)$$

Therefore for any vector u with $\|u\|_\Sigma=1$, we get that

$$\begin{aligned}\left\|\frac{1}{|S_u|}\sum_{i\in S_u}\langle X_i,u\rangle^2X_iX_i^T\right\|&=\frac{n}{|S_u|}\left\|\frac{1}{n}\sum_{i\in S_u}\langle X_i,u\rangle^2X_iX_i^T\right\|\\&=\frac{n}{|S_u|}\left\|\frac{F(u)}{n}\right\|\leqslant O\left(\frac{nL}{|S_u|}\right)\leqslant O\left(\frac{nL}{n(1-O(\varepsilon^2))}\right)\leqslant O(L)\end{aligned}$$

which concludes that it is sufficient to prove our theorem for all the points on the net. It remains to prove that $|G_u|\geqslant(1-O(\varepsilon^2))n$ and that the result holds for all $u\in\mathcal{N}_\gamma$. From Lemma 10 and Lemma 11 we can conclude that for all $u\in\mathcal{N}_\gamma$, $|G_u|\geqslant(1-O(\varepsilon^2))n$ with probability at least 0.95. Furthermore, for any $u\in\mathcal{N}_\gamma$ from Lemma 12, with probability at least $1-\exp\left(-\Omega\left(d\cdot\log\left(d/\varepsilon^4\right)\right)\right)$, $\left\|\frac{1}{|S_u|}F(u)\right\|\leqslant O(L)$. With a union bound over all the $O\left(\left(d^2/\varepsilon^5\right)^d\right)$ vectors in the net, we get that with probability at least 0.95 for n as defined in the theorem, the result holds for all $u\in\mathcal{N}_\gamma$. Therefore, a final union bound over these two events concludes the proof.

□

B.3. Proof of Lemma 12

We restate the lemma and give its proof.

Lemma 12 *With probability at least* $1-\exp\left(-\Omega\left(d\cdot\log\left(d/\varepsilon^4\right)\right)\right)$ *over the choice of* $\{X_i\}_{i=1}^n$*, we have for a fixed* u *with* $\|u\|_\Sigma=1$ *that*

$$\sup_{v:\|v\|_2=1}\frac{1}{n}\sum_{i\in[n]}\langle X_i,u\rangle^2\langle X_i,v\rangle^2\mathbb{1}\left[\langle X_i,u\rangle^2\leqslant\frac{20}{\varepsilon}\right]\mathbb{1}\left[\|X_i\|^2\leqslant\frac{20Ld}{\varepsilon^2}\right]\leqslant L\cdot C''$$

for an absolute constant $C''>0$ *whenever* $n=\Omega\left(\frac{d}{\varepsilon^2}\cdot\log\frac{d}{\varepsilon^4}\right)$ *and* $\varepsilon\kappa\lesssim1$.

Proof. Since $X_i \sim \mathcal{N}(0,\Sigma)$, we have that $\langle X_i, u\rangle = \langle Z_i, \Sigma^{1/2}u\rangle$ where $Z_i \sim \mathcal{N}(0, I_d)$. As a consequence, for $u' := \Sigma^{1/2}u, v' := \Sigma^{1/2}v$

$$\langle X_i, u\rangle^2\langle X_i, v\rangle^2 = \langle Z_i, u'\rangle^2\langle Z_i, v'\rangle^2$$

and $\|u'\|_2 = 1$. The vector v' has norm $\|\Sigma^{1/2}v\|$ and we can define the unit vector $v'' = \frac{v'}{\|v'\|}$ which we can decompose along u' and $u'_\perp$ where $u'_\perp$ is the unit vector perpendicular to u' for this choice of v'[5]. Therefore, we can write $v'' = \beta\cdot u' + \sqrt{1-\beta^2}\cdot u'_\perp$ for some $\beta \in [-1,1]$ and we now have,

$$\begin{aligned}\langle Z_i, v''\rangle^2 = \langle Z_i, \beta\cdot u' + \sqrt{1-\beta^2}\cdot u'_\perp\rangle^2 &\leqslant 2\cdot\beta^2\cdot\langle Z_i, u'\rangle^2 + 2\cdot(1-\beta^2)\cdot\langle Z_i, u'_\perp\rangle^2\\ &\leqslant 2\cdot\left(\langle Z_i, u'\rangle^2 + \langle Z_i, u'_\perp\rangle^2\right)\end{aligned}$$

where we used that $(a+b)^2 \leqslant 2a^2 + 2b^2\ \forall a,b\in\mathbb{R}$. Therefore we have that

$$\begin{aligned}\langle Z_i, u'\rangle^2\langle Z_i, v'\rangle^2 = \langle Z_i, u'\rangle^2\langle Z_i, v''\rangle^2\cdot\|v'\|_2^2 &\leqslant L\cdot\langle Z_i, u'\rangle^2\langle Z_i, v''\rangle^2\\ &\leqslant 2L\cdot\left(\langle Z_i, u'\rangle^4 + \langle Z_i, u'\rangle^2\langle Z_i, u'_\perp\rangle^2\right)\end{aligned}$$

Putting it together, we want to bound,

$$\begin{aligned}&\frac{1}{n}\sum_{i\in[n]}\langle X_i, u\rangle^2\langle X_i, v\rangle^2\mathbb{1}\left[\langle X_i, u\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|X_i\|^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\\ &\leqslant 2L\cdot\frac{1}{n}\sum_{i=1}^n\langle Z_i, u'\rangle^4\mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\\ &+ 2L\cdot\frac{1}{n}\sum_{i=1}^n\langle Z_i, u'\rangle^2\langle Z_i, u'_\perp\rangle^2\cdot\mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\end{aligned}$$

Bounding the First Term

We now consider the first term up to the scaling factor.

$$\frac{1}{n}\sum_{i=1}^n\langle Z_i, u'\rangle^4\cdot\mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]$$

By applying Fact 6, we can bound this quantity, as it represents an average of independent, bounded random variables with bounded variance. To see this, we first define,

$$Y_i := \langle Z_i, u'\rangle^4\cdot\mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]$$

Now we have that

$$|Y_i| = \left|\langle Z_i, u'\rangle^4\cdot\mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\right|$$

5. Note that $u'_\perp$ indeed depends on v. Formally $u'_\perp$ is the unit vector proportional to $(I - u'u'^T)v''$. If $(I-u'u'^T)v'' = 0$, choose any unit vector $u'_\perp \perp u'$; the term multiplied by $\sqrt{1-\beta^2}$ is then zero.

$$\leqslant \left|\langle Z_i, u'\rangle^4 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\right| \leqslant \frac{400}{\varepsilon^2}$$

Furthermore,

$$\mathbb{E}[Y_i] = \mathbb{E}\left[\langle Z_i, u'\rangle^4 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\right] \leqslant \mathbb{E}\big[\langle Z_i, u'\rangle^4\big] \leqslant 3$$
$$\mathbb{E}[Y_i^2] = \mathbb{E}\left[\langle Z_i, u'\rangle^8 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\right] \leqslant \mathbb{E}\big[\langle Z_i, u'\rangle^8\big] \leqslant 105$$

Therefore, we have by Fact 6 we have that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) > t\right] \leqslant \exp\left(-\frac{n^2t^2}{2\big(105n + \frac{1}{3}\frac{400}{\varepsilon^2}nt\big)}\right)$$

By taking t to be a sufficiently large constant, we have that for an absolute constant $C' > 0$ that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \langle Z_i, u'\rangle^4 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] > C'\right] \leqslant \exp\big(-\Omega(n\varepsilon^2)\big)$$

Therefore, with probability at least $1 - \exp(-\Omega(n\varepsilon^2))$

$$\frac{1}{n}\sum_{i=1}^n \langle Z_i, u'\rangle^4 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \leqslant C'. \tag{4}$$

Bounding the Second Term

We now consider the second term up to the scaling factor,

$$\frac{1}{n}\sum_{i=1}^n \langle Z_i, u'\rangle^2\langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]$$

and observe that it suffices to bound,

$$\frac{1}{n}\sum_{i=1}^n \langle Z_i, u'\rangle^2\langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right].$$

We now observe that each summand is a product of independent random variables. In particular, for every $i \in [n]$, $\langle Z_i, u'\rangle$ is independent of $\langle Z_i, u'_\perp\rangle$ as Z_i is a standard Gaussian variable. We also observe that for any $i \neq j$, since X_i and X_j are independent, so are Z_i and Z_j and their respective functions.

To simplify, we define: $\alpha_i(u')^2 = \langle Z_i, u'\rangle^2 \cdot \mathbb{1}\big[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\big]$. Therefore, we now need to bound,

$$\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2,$$

which is the sum of the product of independent random variables. We observe that the second random variable in each summand is sub-exponential, since it is the square of a standard Gaussian random variable. We will exploit this crucial property by conditioning on an event that depends on the $\alpha_i(u')$ across all indices $i \in [n]$. We begin by defining a good set $\mathcal{G}$ depending on u' as follows:

$$\mathcal{G}_{u'} = \left\{ \alpha := \big[\alpha_1(u'), \alpha_2(u') \cdots \alpha_n(u')\big] \in \mathbb{R}^n : \sum_i \alpha_i(u')^2 = O(n) \wedge \sum_i \alpha_i(u')^4 = O(n) \right\}$$

For a fixed $u'_\perp$,

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2 > t\right] &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2 > t \mid \alpha \in \mathcal{G}_{u'}\right] \cdot \mathbb{P}[\alpha \in \mathcal{G}_{u'}] \\
&\quad + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2 > t \mid \alpha \notin \mathcal{G}_{u'}\right] \cdot \mathbb{P}[\alpha \notin \mathcal{G}_{u'}] \\
&\leqslant \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2 > t \mid \alpha \in \mathcal{G}_{u'}\right] + \mathbb{P}[\alpha \notin \mathcal{G}_{u'}]
\end{aligned}$$

where we used $\mathbb{P}[\alpha \in \mathcal{G}_{u'}] \leqslant 1$ and $\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2 > t \mid \alpha \notin \mathcal{G}_{u'}\right] \leqslant 1$. We abuse notation and will use $\mathcal{G}$ for $\mathcal{G}_{u'}$. Defining $\textsc{Avg}(\alpha, u', u'_\perp) := \frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2$, we have that

$$\begin{aligned}
\mathbb{P}\big[\textsc{Avg}(\alpha, u', u'_\perp) > t \mid \alpha \in \mathcal{G}\big] &= \frac{1}{\mathbb{P}[\alpha \in \mathcal{G}]} \cdot \mathbb{P}[\textsc{Avg}(\alpha, u', u'_\perp) > t \cap \alpha \in \mathcal{G}] \\
&= \frac{1}{\mathbb{P}[\alpha \in \mathcal{G}]} \cdot \int_{a \in \mathcal{G}} \mathbb{P}[\textsc{Avg}(\alpha, u', u'_\perp) > t | \alpha = a] \cdot p(a) \cdot da
\end{aligned}$$

where $p(a)$ is the probability density function at the vector a. We now observe that we can control $\mathbb{P}[\textsc{Avg}(\alpha, u', u'_\perp) > t | \alpha = a]$ whenever $a \in \mathcal{G}$ for fixed a as it is a weighted sum of sub-exponential random variables. Denoting this tail probability by B for some $B \in [0, 1]$, we have that

$$\begin{aligned}
\mathbb{P}\big[\textsc{Avg}(\alpha, u', u'_\perp) > t \mid \alpha \in \mathcal{G}\big] &= \frac{1}{\mathbb{P}[\alpha \in \mathcal{G}]} \cdot \int_{a \in \mathcal{G}} \mathbb{P}[\textsc{Avg}(\alpha, u', u'_\perp) > t | \alpha = a] \cdot p(a) \cdot da \\
&\leqslant B \cdot \frac{1}{\mathbb{P}[\alpha \in \mathcal{G}]} \int_{a \in \mathcal{G}} p(a) \cdot da = B
\end{aligned}$$

Putting it together, we have that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \alpha_i(u')^2\langle Z_i, u'_\perp\rangle^2 > t\right] \leqslant B + \mathbb{P}[\alpha \notin \mathcal{G}]$$

In the remainder of the proof, we will first show that $\alpha \in \mathcal{G}$ with probability at least $1 - 2\exp(-\Omega(n\varepsilon^2))$. Then we will apply the tail bound to complete the bound for the second term. Finally, we will put all the pieces together.

$\alpha \in \mathcal{G}$ WITH HIGH PROBABILITY

We will make essential use of Fact 6 once again to show that $\alpha \in \mathcal{G}$ with high probability. We recall that

$$\alpha_i(u')^2 = \langle Z_i, u' \rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u' \rangle^2 \leqslant \frac{20}{\varepsilon}\right]$$

Once again, observe that this is a bounded random variable with bounded variance since:

$$\begin{aligned}
&\mathbb{E}[\alpha_i(u')^2] \leqslant \mathbb{E}\big[\langle Z_i, u' \rangle^2\big] = 1 \\
&|\alpha_i(u')^2| \leqslant \frac{20}{\varepsilon} \\
&\mathbb{E}[\alpha_i(u')^4] \leqslant \mathbb{E}\big[\langle Z_i, u' \rangle^4\big] = 3
\end{aligned}$$

Therefore via an application of Fact 6, we have that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\big(\alpha_i(u')^2 - \mathbb{E}[\alpha_i(u')^2]\big) > O(1)\right] \leqslant \exp(-\Omega(n\varepsilon))$$

which implies that with probability at least $1 - \exp(-\Omega(n\varepsilon))$ that

$$\sum_{i=1}^{n} \alpha_i(u')^2 \leqslant a \cdot n$$

for some absolute constant $a > 0$. Similarly observe that

$$\begin{aligned}
&\mathbb{E}[\alpha_i(u')^4] \leqslant \mathbb{E}\big[\langle Z_i, u' \rangle^4\big] = 3 \\
&|\alpha_i(u')^4| \leqslant \frac{400}{\varepsilon^2} \\
&\mathbb{E}[\alpha_i(u')^8] \leqslant \mathbb{E}\big[\langle Z_i, u' \rangle^8\big] = 105
\end{aligned}$$

Therefore via an application of Fact 6, we have that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\big(\alpha_i(u')^4 - \mathbb{E}[\alpha_i(u')^4]\big) > O(1)\right] \leqslant \exp\big(-\Omega(n\varepsilon^2)\big)$$

and as a consequence have that

$$\sum_{i=1}^{n} \alpha_i(u')^4 \leqslant b \cdot n$$

for some absolute constant $b > 0$. Therefore, via a union bound, we have that with probability at least $1 - 2\exp(-\Omega(n\varepsilon^2))$

$$\sum_{i} \alpha_i(u')^2 \leqslant a \cdot n \text{ and } \sum_{i} \alpha_i(u')^4 \leqslant b \cdot n$$

from which we deduce,

$$\mathbb{P}[\alpha \in \mathcal{G}] \geqslant 1 - 2\exp(-\Omega(n\varepsilon^2)). \tag{5}$$

SUB-EXPONENTIAL TAIL BOUNDS FOR IMPROVED NORM BOUND.

Recall that it suffices for us to bound

$$\mathbb{P}[\textsc{Avg}(\alpha, u', u'_\perp) > t | \alpha = a]$$

for a fixed $a \in \mathcal{G}$, where $\textsc{Avg}(\alpha, u', u'_\perp) := \frac{1}{n}\sum_{i=1}^{n} \alpha_i (u')^2 \langle Z_i, u'_\perp\rangle^2$. Indeed, we will apply a sub-exponential weighted sum tail bound (Fact 7) to complete the proof. We have that $\langle Z_i, u'_\perp\rangle^2$ is a χ^2-random variable with a single degree of freedom, as it is the square of the standard Gaussian random variable. Abusing notation, for $a_i = \alpha_i^2$, we have that $\max_i a_i = O(1/\varepsilon)$ and further that $\sum_i a_i^2 = O(n)$. Taking $t = O(n)$ we have that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{n} \alpha_i (u')^2 \left(\langle Z_i, u'_\perp\rangle^2 - 1\right)\right| > O(n)\right) &\leqslant 2\exp\left(-\Omega\left(\min\left(\frac{n^2}{\sum_{i=1}^n 2a_i^2}, \frac{n}{\max_i |a_i| K}\right)\right)\right)\\ &= 2\exp\left(-\Omega\left(\min\left(n, \frac{n\varepsilon}{K}\right)\right)\right)\\ &= 2\exp(-\Omega(n\varepsilon))\end{aligned}$$

since $K = O(1)$ for our specific random variable. Putting the pieces together we have that

$$\mathbb{P}[\textsc{Avg}(\alpha, u', u'_\perp) > O(1) | \alpha = a] \leqslant \exp(-\Omega(n\varepsilon))$$

Therefore, to complete the entire second term bound, we have that

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \langle Z_i, u'\rangle^2 \langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right] > O(1)\right]\\ &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \alpha_i (u')^2 \langle Z_i, u'_\perp\rangle^2 > O(1)\right]\\ &\leqslant \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \alpha_i (u')^2 \langle Z_i, u'_\perp\rangle^2 > O(1) \mid \alpha \in \mathcal{G}\right] + \mathbb{P}[\alpha \notin \mathcal{G}]\\ &\leqslant 2\exp(-\Omega(n\varepsilon)) + 2\exp(-\Omega(n\varepsilon^2)) = 4\exp(-\Omega(n\varepsilon^2))\end{aligned}$$

Therefore, putting everything together, we have that for a fixed $u'_\perp$ that

$$\frac{1}{n}\sum_{i=1}^{n} \langle Z_i, u'\rangle^2 \langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right] \mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \leqslant C_2.$$

with probability at least $1 - 4\exp(-\Omega(n\varepsilon^2))$ for some absolute constant $C_2 > 0$.

Union Bound

To complete the bound, we now need a union bound over all choices of $u'_\perp$ since $u_\perp$ depends on v. We want to bound,

$$\mathbb{P}\left[\sup_{u'_\perp} \frac{1}{n}\sum_{i=1}^{n} \langle Z_i, u'\rangle^2 \langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right] \mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] > C_3\right]$$

for $C_3 \coloneqq C_2 + 1$. Now we will do a net argument over a γ net of the unit sphere's intersection with the set of all vectors perpendicular to u'. The Lipschitzness of the following function will be precisely what we will prove next.

$$f(z) := \frac{1}{n}\sum_{i=1}^{n}\langle Z_i, u'\rangle^2\langle Z_i, z\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right].$$

Since u' is fixed, we have that for any $z \in \mathcal{N}_\gamma$ with $\|z\|_2 = 1$ such that $\|z - u'_\perp\|_2 = \gamma$,

$$\begin{aligned}
\big|f(u'_\perp) - f(z)\big| =& \Big|\frac{1}{n}\sum_{i=1}^{n}\langle Z_i, u'\rangle^2\langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \\
&- \frac{1}{n}\sum_{i=1}^{n}\langle Z_i, u'\rangle^2\langle Z_i, z\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right]\Big| \\
&\leqslant \frac{20}{\varepsilon}\frac{1}{n}\sum_i \big|\langle Z_i, z - u'_\perp\rangle\langle Z_i, z + u'_\perp\rangle\big|\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \\
&\leqslant \frac{40\gamma}{\varepsilon n}\sum_i \|Z_i\|^2\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \\
&\leqslant \frac{40\gamma}{\varepsilon n}\sum_i \|\Sigma^{-1/2}\|_2^2\|Z_i\|_\Sigma^2\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \\
&\leqslant \frac{40\gamma}{\varepsilon}\frac{1}{\mu}\frac{20Ld}{\varepsilon^2} = \gamma \cdot \frac{800Ld}{\mu\varepsilon^3} \leqslant \gamma \cdot \frac{800d}{\varepsilon^4}
\end{aligned}$$

where in the last step we used $\varepsilon\kappa < 1$. Taking $\gamma = \frac{\varepsilon^4}{800d}$ suffices to establish the Lipschitzness of the function. We know that the γ-net $\mathcal{N}_\gamma$ defined as

$$\mathcal{N}_\gamma \subseteq \mathbb{S}^{d-1} \cap (u')^\perp.$$

has

$$|\mathcal{N}_\gamma| \leqslant \left(\frac{3}{\gamma}\right)^d$$

many elements. In particular our net has

$$|\mathcal{N}_\gamma| \leqslant \left(\frac{2400d}{\varepsilon^4}\right)^d$$

Now we do a union bound over this net and get that

$$\begin{aligned}
&\mathbb{P}\left[\sup_{u'_\perp}\frac{1}{n}\sum_{i=1}^{n}\langle Z_i, u'\rangle^2\langle Z_i, u'_\perp\rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u'\rangle^2 \leqslant \frac{20}{\varepsilon}\right]\mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] > C_3\right] \\
&\leqslant \left(\frac{2400d}{\varepsilon^4}\right)^d \cdot \exp(-\Omega(n\varepsilon^2)) \\
&= \exp\left(d\log\frac{2400d}{\varepsilon^4} - \Omega(n\varepsilon^2)\right).
\end{aligned}$$

By taking $n = 100d/\varepsilon^2 \log \frac{2400d}{\varepsilon^4}$, we can make the failure probability $\exp(-99d \log \frac{2400d}{\varepsilon^4})$. Therefore we have that with probability at least $1 - \exp(-99d \log \frac{2400d}{\varepsilon^4})$ that

$$\sup_{u'_\perp} \frac{1}{n} \sum_{i=1}^{n} \langle Z_i, u' \rangle^2 \langle Z_i, u'_\perp \rangle^2 \cdot \mathbb{1}\left[\langle Z_i, u' \rangle^2 \leqslant \frac{20}{\varepsilon}\right] \mathbb{1}\left[\|Z_i\|_\Sigma^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \leqslant C_3 \tag{6}$$

for some absolute constant C_3 whenever $n = \Omega\big(\frac{d}{\varepsilon^2} \cdot \log \frac{d}{\varepsilon^4}\big)$.

Completing the proof

From Eqs. (4), (6), and (5), we have with probability at least $1 - 2\exp(-\Omega(n\varepsilon^2)) - \exp(-\Omega(n\varepsilon^2)) - \exp\big(-\Omega(d \log \frac{d}{\varepsilon^4})\big) = 1 - \exp\big(-\Omega(d \log \frac{d}{\varepsilon^4})\big)$ that for a fixed u with $\|u\|_\Sigma = 1$,

$$\sup_{v:\|v\|_2=1} \frac{1}{n} \sum_{i \in [n]} \langle X_i, u \rangle^2 \langle X_i, v \rangle^2 \mathbb{1}\left[\langle X_i, u \rangle^2 \leqslant \frac{20}{\varepsilon}\right] \mathbb{1}\left[\|X_i\|^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \leqslant L \cdot C''$$

for some absolute constant $C'' > 0$. As a consequence,

$$\left\| \frac{1}{n} \sum_i \langle X_i, u \rangle^2 X_i X_i^T \mathbb{1}\left[\langle X_i, u \rangle^2 \leqslant \frac{20}{\varepsilon}\right] \mathbb{1}\left[\|X_i\|^2 \leqslant \frac{20Ld}{\varepsilon^2}\right] \right\|_{\text{op}} \leqslant L \cdot C'',$$

concluding the proof of Lemma 12. □

Proof of Theorem 2

The proof of Theorem 2 now follows similarly to Theorem 5 of Jambulapati et al. (2021) by replacing their Lemma 19 (Lemma 5) with our Theorem 9.

Appendix C. Statistical Query Lower Bounds for Robust Linear Regression

In this work, we rely on the SQ hardness for NGCA derived in Diakonikolas et al. (2017). More recently, these SQ hardness results have been refined in Diakonikolas et al. (2023b), which proves hardness under moment matching alone, without requiring bounded χ^2-divergence. However, these refinements appear to only provide quantitatively weaker lower bounds for fine-grained problems such as ours, and seem better suited to coarser separations such as between polynomial time and super-polynomial time SQ algorithms. We first give some background on the SQ Lower Bound framework. We begin by defining an SQ algorithm and the standard oracles.

Definition 14 *We define SQ Algorithms as algorithms that do not see samples from an underlying distribution D but instead have access to the following oracles.*

1. STAT(τ)*: For a tolerance parameter $\tau > 0$, and* any *bounded function $f : \mathbb{R}^d \to [-1, 1]$,* STAT($\tau$) *returns a value v such that $|v - \mathbb{E}_{x \sim D}[f(x)]| \leqslant \tau$.*

2. VSTAT(t)*: For a sample size parameter $t > 0$ and* any *bounded function $f : \mathbb{R}^d \to [0, 1]$,* VSTAT($t$) *returns a value v such that $|v - \mathbb{E}_{x \sim D}[f(x)]| \leqslant \tau$ where $\tau = \max\left\{\frac{1}{t}, \sqrt{\frac{Var_{x \sim D}[f(x)]}{t}}\right\}$ and $Var_{x \sim D}[f(x)] = \mathbb{E}_{x \sim D}[f(x)](1 - \mathbb{E}_{x \sim D}[f(x)])$.*

C.1. Search Problems and their SQ Hardness

Definition 15 (Search Problem over Distributions) *Let $\mathcal{D}$ be a set of distributions on $\mathbb{R}^d$, let $\mathcal{F}$ be a set of solutions, and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map from the set of distributions to subsets of solutions. The search problem $\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ is to find* a *solution $f \in \mathcal{Z}(D)$ given oracle access to an unknown $D \in \mathcal{D}$.*

The hardness of search problems is captured by the SQ Dimension. As a first step, we define the pairwise correlation between distributions.

Definition 16 (Pairwise Correlation) *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^d \to \mathbb{R}_+$ with respect to a reference distribution with density $S : \mathbb{R}^d \to \mathbb{R}_+$, where $\mathrm{supp}(D_1) \cup \mathrm{supp}(D_2) \subseteq \mathrm{supp}(S)$ is defined as $\chi_S(D_1, D_2) := \int_{\mathbb{R}^d} D_1(x)D_2(x)/S(x)dx - 1$. When $D_1 = D_2$, this is the χ^2-divergence between D_1 and S.*

Definition 17 *For $\gamma, \beta > 0$, the set of distributions $\{D_1, D_2, \ldots, D_m\}$ is called (γ, β)-correlated relative to the distribution S, if $|\chi_S(D_i, D_j)| \leqslant \gamma$ whenever $i \neq j$ and $|\chi_S(D_i, D_i)| \leqslant \beta$.*

The Statistical Query (SQ) dimension of a search problem is the largest set of (γ, β)-correlated distributions assigned to each solution.

Definition 18 (SQ Dimension Defn. 2.11 in Diakonikolas et al. (2017)) *For $\gamma, \beta > 0$, a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X, we define the statistical dimension of $\mathcal{Z}$ denoted by $SD(\mathcal{Z}, \gamma, \beta)$ to be the largest integer m such that there exists a reference distribution S over X and a finite set of distributions $\mathcal{D}_S \subset \mathcal{D}$ such that for any solution $f \in \mathcal{F}$, the set $\mathcal{D}_f = \mathcal{D}_S \backslash \mathcal{Z}^{-1}(f)$ is (γ, β) correlated with respect to S and $|\mathcal{D}_f| \geqslant m$.*

Lemma 19 (Corollary 3.12 in Feldman et al. (2017)) *Let $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over $\mathbb{R}^d$. For $\gamma, \beta > 0$, let $s = SD(\mathcal{Z}, \gamma, \beta)$ be the statistical dimension of the problem. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{Z}$ requires at least $s\gamma'/(\beta - \gamma)$ queries to the $\mathrm{STAT}(\sqrt{\gamma + \gamma'})$ or $\mathrm{VSTAT}(1/(3(\gamma + \gamma')))$ oracles to solve $\mathcal{Z}$.*

We next state a few standard techniques from Diakonikolas et al. (2017) that we will utilize.

Definition 20 (Hidden Direction Distribution) *For a unit vector $v \in \mathbb{R}^d$ and a distribution A on the real line with density function $A(x)$, we define $P_{A,v}$ to be the distribution that is A in the direction v and standard Gaussian in the $d-1$ dimensional complement of v. Formally, $P_{A,v}(x) = A(\langle v, x\rangle)\, \varphi_{v^\perp}(x)$ where $\varphi_{v^\perp}(x) = \exp(-\|x - \langle v, x\rangle v\|^2/2)/(2\pi)^{(d-1)/2}$.*

The distributions $P_{A,v}$ can be shown to be nearly uncorrelated as long as the directions in which A is embedded are pairwise almost orthogonal. We remark here that the problem of finding the hidden direction is also referred to as Non-Gaussian component analysis (NGCA) in contemporary literature.

Lemma 21 (Lemma 3.4 in Diakonikolas et al. (2017)) *Let $m \in \mathbb{Z}_+$. Let A be a distribution over $\mathbb{R}$ that matches the first m moments of $\mathcal{N}(0, 1)$. For any v, let $P_{A,v}$ be the distribution defined above. For all $v', v \in \mathbb{R}^d$, we have that $\left|\chi_{\mathcal{N}(0,I_d)}(P_{A,v}, P_{A,v'})\right| \leqslant |\langle v, v'\rangle|^{m+1} \cdot \chi^2(A, \mathcal{N}(0, 1))$.*

Lemma 22 (Lemma 3.7 in Diakonikolas et al. (2017)) *For any $0 < c < \frac{1}{2}$, there is a set S of at least $2^{\Omega(d^c)}$ unit vectors in $\mathbb{R}^d$ such that for each pair of distinct v, v' it is the case that $|\langle v, v'\rangle| \leqslant O(d^{c-1/2})$.*

Proof of Theorem 3

Utilizing the machinery described above, we now prove our main result of this section, Theorem 3.

Theorem 23 (Full Version of Theorem 3) *Let $0 < c < 1/2$ be an arbitrary sufficiently small constant, and κ, ε be given, satisfying $\varepsilon\kappa \lesssim 1$, κ sufficiently large and ε sufficiently small. For every vector $v \in \mathbb{R}^d$ with $\|v\| = 1$, there exists a covariance matrix Σ_v (with condition number κ), a direction β_v, and a corruption distribution E_v, such that the following holds.*

Let Q_v be a distribution over (X, y) where $X \sim \mathcal{N}(0, \Sigma_v)$ and $y \sim \langle X, \beta_v\rangle + \eta$ where $\eta \sim \mathcal{N}(0, \zeta^2)$, $\zeta^2 < 1$ unknown and η independent of X.

Let $Q'_v = (1-\varepsilon)Q_v + \varepsilon E_v$. Then, any SQ algorithm, which is given access to Q'_v (without knowing v) that outputs a vector $\widehat{\beta}_v$ such that $\|\widehat{\beta}_v - \beta_v\|_{\Sigma_v} = o(\sqrt{\varepsilon\kappa})$ either makes at least $2^{\Omega(d^c)}$ queries, or at least one query to the STAT oracle with tolerance

$$\tau \leqslant O\left(\sqrt{d^{4c-2} \cdot (\sqrt{\kappa} e^{O(1/\varepsilon)})}\right).$$

Proof of Theorem 23. Let $T \subset \mathcal{S}^{d-1}$ be the set of nearly-orthogonal vectors from Lemma 22 with $|T| = 2^{\Omega(d^c)}$. For each $v \in T$ we define:

$$Q'_v(x, y) = (1-\varepsilon) \cdot Q_v(x, y) + \varepsilon \cdot E_v(x, y),$$

where Q_v and Q'_v are as defined in Section C.2 and Section C.4 respectively. Our goal is to ensure the preconditions for applying Lemma 19. To that end we let the reference distribution be

$$S = \mathcal{N}(0, I_d) \times R(y),$$

where $R(y)$ is defined in Section C.4, and define $\mathcal{D}_S := \{Q'_v\}_{v\in T}$. Let $\beta_v = c_1\sqrt{\varepsilon}\kappa v$ denote the regression vector corresponding to Q'_v for a constant $c_1 > 0$ that will be picked later in Section C.4. The set of solutions $\mathcal{F}$ is defined as, $\mathcal{F} := \{u \in \mathbb{R}^d : \|u\|_2 \leqslant 2c_1\sqrt{\varepsilon}\kappa\}$, and the set of solutions $\mathcal{Z}(Q'_v)$ corresponding to Q'_v is

$$\mathcal{Z}(Q'_v) := \{u \in \mathcal{F} : \|u - \beta_v\|_{\Sigma_v} \leqslant r\},$$

for $r = o(\sqrt{\varepsilon\kappa})$. We claim that for any $u \in \mathcal{F}$, there is at most one element $Q'_v \in \mathcal{D}_S$ that satisfies $\|u - \beta_v\|_2 = o(\sqrt{\varepsilon}\kappa)$. To see why, observe that for some $u_v \in \mathcal{Z}(Q'_v)$,

$$\begin{aligned}\|u_v - \beta_v\|_2 &= \|\Sigma_v^{-1/2}\Sigma_v^{1/2}(u_v - \beta_v)\|_2 \leqslant \|\Sigma_v^{-1/2}\|_2\|\Sigma_v^{1/2}(u_v - \beta_v)\|_2 \\ &= \|\Sigma_v^{-1/2}\|_2\|(u_v - \beta_v)\|_{\Sigma_v} \leqslant \sqrt{\kappa} \cdot o(\sqrt{\varepsilon\kappa}) = o(\sqrt{\varepsilon}\kappa)\end{aligned}$$

However, for $v \neq v'$, $\|\beta_v - \beta_{v'}\|_2 = \Omega(\sqrt{\varepsilon}\kappa)$ since

$$\|\beta_v - \beta_{v'}\|_2 = \sqrt{2c_1^2\varepsilon\kappa^2(1 - \langle v, v'\rangle)} = \Omega(\sqrt{\varepsilon}\kappa).$$

We now claim that $u_v \notin \mathcal{Z}(Q'_{v'})$ for any $v' \neq v$. Note that

$$\begin{aligned}\Omega(\sqrt{\varepsilon}\kappa) &= \|\beta_v - \beta_{v'}\|_2 \\ &\leqslant \|u_v - \beta_v\|_2 + \|u_v - \beta_{v'}\|_2 \leqslant o(\sqrt{\varepsilon}\kappa) + \|u_v - \beta_{v'}\|_2.\end{aligned}$$

Now,

$$\Omega(\sqrt{\varepsilon\kappa}) \leqslant \|u_v - \beta_{v'}\|_2 \leqslant \sqrt{\kappa} \cdot \|u_v - \beta_{v'}\|_{\Sigma_{v'}}$$

which implies

$$\|u_v - \beta_{v'}\|_{\Sigma_{v'}} \geqslant \Omega(\sqrt{\varepsilon\kappa})$$

which is larger than $r = o(\sqrt{\varepsilon\kappa})$ concluding our claim. Therefore, $|\mathcal{Z}^{-1}(u)| \leqslant 1$, and $|\mathcal{D}_S \backslash \mathcal{Z}^{-1}(u)| \geqslant |\mathcal{D}_S| - 1 \geqslant 2^{\Omega(d^c)}$, where we use Lemma 22. We will prove in Lemma 32, that the pairwise correlations γ, β with respect to S are

$$\gamma \leqslant O(|\langle v, v' \rangle|^4 \cdot (\sqrt{\kappa} e^{O(1/\varepsilon)})) = O(d^{4c-2}\sqrt{\kappa} e^{O(1/\varepsilon)}), \quad \text{and,} \quad \beta \leqslant O(\sqrt{\kappa} e^{O(1/\varepsilon)}).$$

Taking $\gamma' = \gamma$ in Lemma 19 and using $\gamma/\beta \leqslant O(d^{4c-2}) = o(1)$ for $c < 1/2$, any algorithm using only queries with STAT tolerance larger than $O(\sqrt{\gamma})$ must make at least $2^{\Omega(d^c)}$ queries. Equivalently, any algorithm making fewer queries must issue a STAT query with tolerance $O(\sqrt{\gamma})$. As a result, our search problem $\mathcal{Z}$ has SQ dimension at least $2^{\Omega(d^c)}$ and is (γ, β) correlated. We finally apply Lemma 19 to conclude that any SQ algorithm that can output a vector u such that $\|u - \beta_v\|_{\Sigma_v} \leqslant o(\sqrt{\varepsilon\kappa})$ makes at least $2^{\Omega(d^c)}$ queries, or at least one query to STAT of tolerance

$$\tau \leqslant O\left(\sqrt{d^{4c-2} \cdot (\sqrt{\kappa} e^{O(1/\varepsilon)})}\right).$$

□

In the following sections, we will construct distributions Q_v and Q'_v and prove bounds on their pairwise correlations. In order to prove such correlation bounds we utilize Lemma 21. Therefore, we first define our distribution Q_v in Section C.2, then construct a moment matching distribution Q'_v of the required form in Section C.3. Further in Section C.4 we show that our constructed Q'_v is indeed a corruption of Q_v. Finally in Section C.5 we prove the required pairwise correlation bounds.

C.2. Lower Bound Instance

We describe the uncorrupted distribution Q_v. The main idea is similar to prior works and focuses on aligning the signal β with the direction of low variance of Σ, which is the same as the unknown direction v. More precisely, for every v, we define the parameters of the uncorrupted distribution Q_v as follows.

Definition 24 (Hard Instance for SQ Lower Bound)

$$\Sigma_v = I_d - (1 - 1/\kappa) v v^T$$
$$\beta_v = c_1 \sqrt{\varepsilon} \kappa v, \; c_1 > 0.$$

Similar to Diakonikolas et al. (2019c), we set y to have unit variance. That is, $\sigma_y^2 = c_1^2 \varepsilon\kappa + \zeta^2 = 1$. The main observation in the lower bound construction of Diakonikolas et al. (2019c) is that the conditional distribution $X|Y = y$ is an instance of the hidden direction detection problem. They then match the first three moments of the conditional distribution $X|Y = y$ with $\mathcal{N}(0, 1)$. This conditional distribution takes the following form

$$[X|Y = y] \sim \mathcal{N}\left(\frac{\Sigma\beta y}{\sigma_y^2}, \Sigma - \frac{(\Sigma\beta)(\Sigma\beta)^T}{\sigma_y^2}\right).$$

and is not Gaussian in the direction v. The one-dimensional conditional distribution along the direction v is

$$\mathcal{N}\left(\underbrace{c_1\sqrt{\varepsilon}y}_{:=\mu_s.}, \underbrace{\frac{1}{\kappa} - c_1^2\varepsilon}_{:=\sigma_s^2.}\right).$$

We pause here to remark that in Diakonikolas et al. (2019c), they pick κ to be a constant (at most 2). This choice made the resulting conditional distribution have constant variance in the direction of v. In that setting, the above Gaussian with large constant variance is corrupted to match moments with $\mathcal{N}(0,1)$. This sufficed for their goal of showing the hardness of achieving error $o(\sqrt{\varepsilon})$ in the sub-quadratic sample regime.

In contrast, in our setting, the variance of the one-dimensional conditional distribution can be arbitrarily small. Since our goal is to establish hardness results for achieving an error $o(\sqrt{\varepsilon\kappa})$, we cannot use the same moment matching construction as Diakonikolas et al. (2019c). Instead, we require a new moment matching construction, and the ill-conditioning introduces additional challenges that we address in the next section.

We also note that the conditional distribution considered above depends on y. Therefore, we must corrupt the distribution so that it matches the moments of $\mathcal{N}(0,1)$ for every possible $y \in \mathbb{R}$. Diakonikolas et al. (2019c) handled this by choosing a different corruption rate for each y while still ensuring a global corruption rate of ε on the joint distribution. We adopt a similar strategy.

We fix $c_1 > 0$ to be a sufficiently small absolute constant such that $1 - c_1^2\varepsilon\kappa$ is bounded below by an absolute constant, the normalization bound in Section C.4 holds, and $e^{O(c_1^2\varepsilon y^2)}G(y) \leqslant e^{-\Omega(y^2)}$.

C.3. Moment Matching

In this section, we will prove the following.

Lemma 25 *For $\varepsilon > 0$ sufficiently small, $\mu_s \in \mathbb{R}$, $\sigma_s^2 \in (0, 0.1]$, there exists a distribution A_{μ_s} such that A_{μ_s} agrees with the first three moments of $\mathcal{N}(0,1)$ and has the form $A_{\mu_s} = (1-\varepsilon_{\mu_s})\cdot\mathcal{N}(\mu_s,\sigma_s^2) + \varepsilon_{\mu_s}\cdot B_{\mu_s}$ for some distribution B_{μ_s} and ε_{μ_s} such that:*

- *If $|\mu_s| \geqslant \frac{\sqrt{\varepsilon}}{10000}$, then $\varepsilon_{\mu_s}/(1-\varepsilon_{\mu_s}) \leqslant O(\mu_s^2)$, and,*
$$\chi^2(A_{\mu_s},\mathcal{N}(0,1)) = O(\sqrt{\kappa}e^{O(\max\{\mu_s^2,1/\mu_s^2\})})$$
- *If $|\mu_s| < \frac{\sqrt{\varepsilon}}{10000}$, then $\varepsilon_{\mu_s} = \varepsilon$, and $\chi^2(A_{\mu_s},\mathcal{N}(0,1)) = \exp(O(\frac{1}{\varepsilon}))$.*

In the above we used that $1/\sigma_s^2 = O(\kappa)$.

Our result generalizes the moment matching construction of (Diakonikolas et al., 2019c, Lemma E.2) for $\kappa = \Omega(1)$. We will prove the lemma towards the end of the section. We now present our construction and then compute the χ^2-divergence bounds. Finally we put both pieces together to conclude the proof of Lemma 25.

Definition 26 *Define the distribution,*

$$A_{\mu_s} := \left\{\begin{array}{l} P_{1,\varepsilon_{\mu_s}} \text{ whenever } |\mu_s| \leqslant \sqrt{\varepsilon}/10000, \\ P_{2,\varepsilon_{\mu_s}} \text{ whenever } \sqrt{\varepsilon}/10000 < |\mu_s| < 0.65, \\ P_{3,\varepsilon_{\mu_s}} \text{ whenever } 0.65 \leqslant |\mu_s|. \end{array}\right\} \tag{7}$$

1. For $|\mu_s| < \frac{\sqrt{\varepsilon}}{10000}$, we use the following slight strengthening of Lemma 7.10 of Diakonikolas et al. (2025b), which follows from the proof of that lemma.

   Lemma 27 (Variant of Lemma 7.10, Diakonikolas et al. (2025b)) *There exists a positive constant η_0 such that for all $\eta \in (0, \eta_0)$, every $\xi \in (0, 1/2)$, and every δ satisfying $|\delta| \leqslant 0.001\sqrt{\eta}$, there exist univariate distributions A and Q satisfying*

   $$A = (1-\eta)\mathcal{N}(\delta, \xi^2) + \eta Q$$

   such that A matches the first three moments of $\mathcal{N}(0,1)$. Furthermore, $Q = \sum_{i=1}^{4} w_i \mathcal{N}(\mu_i, 1)$, where $|\mu_i| \leqslant 10/\sqrt{\eta}$, $w_i \geqslant 0$, and $\sum_i w_i = 1$.

   We apply Lemma 27 with $\eta = \varepsilon$, $\delta = \mu_s$, and $\xi = \sigma_s$. In this regime,

   $$|\mu_s| \leqslant \sqrt{\varepsilon}/10000 \leqslant 0.001\sqrt{\varepsilon},$$

   so the lemma yields

   $$P_{1,\varepsilon_{\mu_s}} = A_{\mu_s} = (1-\varepsilon)\mathcal{N}(\mu_s, \sigma_s^2) + \varepsilon Q_{\mu_s},$$

   where Q_{μ_s} is a mixture of at most four unit-variance Gaussians with component means bounded by $O(1/\sqrt{\varepsilon})$.

2. For $\frac{\sqrt{\varepsilon}}{10000} \leqslant |\mu_s| < 0.65$, we consider the following construction. Let ε_{μ_s} be a function of μ_s, to be specified later. We define a mixture of Gaussians whose first three moments match those of $\mathcal{N}(0,1)$:

   $$P_{2,\varepsilon_{\mu_s}} = (1-\varepsilon_{\mu_s}) \cdot \mathcal{N}(\mu_s, \sigma_s^2) + \varepsilon_{\mu_s}\left(\frac{1}{9} \cdot \mathcal{N}(-2\mu_N, \sigma_N^2) + \frac{8}{9} \cdot \mathcal{N}(\mu_N, \tau^2)\right)$$

   where $\tau^2 > 0$ is a fixed constant. This construction is inspired by Diakonikolas et al. (2019c), from which we adopt the mixing weights and the relationship between the noise means. The key difference is that our setting is ill-conditioned, as σ_s^2 may be arbitrarily close to zero. To address this, we allow μ_s to depend on σ_s^2 in a precise manner that ensures all component variances remain uniformly bounded (specifically in $(0,2)$), which is necessary to obtain finite χ^2 divergence with respect to $\mathcal{N}(0,1)$. We provide more intuition for this later.

   We next specify μ_N, σ_N^2, and τ^2, determine the relationship between ε_{μ_s} and μ_s, and show how to express μ_s, μ_N, and σ_N^2 in terms of σ_s^2, τ^2, and ε_{μ_s}. Finally, we prove that for every fixed σ_s^2, there is a bijection between values of μ_s and ε_{μ_s}.

   Using that the first moment must be 0 yields,

   $$\mu_s = -\frac{2\varepsilon_{\mu_s}}{3(1-\varepsilon_{\mu_s})}\mu_N.$$

   Substituting this into the second- and third-moment equations, we note that the third-moment constraint is linear in σ_N^2, while the second-moment constraint is linear in μ_N^2. We therefore solve first for σ_N^2, substitute into the variance equation to obtain μ_N^2 as a function of ε_{μ_s}, τ^2,

and σ_s^2, and then recover σ_N^2. These symbolic computations were carried out using SymPy Meurer et al. (2017)[6]. With this, we get the following values,

$$\mu_N^2 = \frac{27(1-\varepsilon_{\mu_s})^2\big(4\varepsilon_{\mu_s}\sigma_s^2 - 4\varepsilon_{\mu_s}\tau^2 - 3\sigma_s^2 + 3\big)}{4\varepsilon_{\mu_s}\big(17\varepsilon_{\mu_s}^2 - 45\varepsilon_{\mu_s} + 27\big)}.$$

$$\sigma_N^2 = \frac{-63\varepsilon_{\mu_s}^2\sigma_s^2 + 80\varepsilon_{\mu_s}^2\tau^2 + 144\varepsilon_{\mu_s}\sigma_s^2 - 180\varepsilon_{\mu_s}\tau^2 - 9\varepsilon_{\mu_s} - 81\sigma_s^2 + 108\tau^2}{17\varepsilon_{\mu_s}^2 - 45\varepsilon_{\mu_s} + 27}.$$

Using $\tau^2 = 0.2$ and taking the positive square root gives

$$\mu_s = \sqrt{\frac{3\varepsilon_{\mu_s}(20\varepsilon_{\mu_s}\sigma_s^2 - 4\varepsilon_{\mu_s} - 15\sigma_s^2 + 15)}{5\big(17\varepsilon_{\mu_s}^2 - 45\varepsilon_{\mu_s} + 27\big)}}, \tag{8}$$

and,

$$\sigma_N^2 = \frac{-315\varepsilon_{\mu_s}^2\sigma_s^2 + 80\varepsilon_{\mu_s}^2 + 720\varepsilon_{\mu_s}\sigma_s^2 - 225\varepsilon_{\mu_s} - 405\sigma_s^2 + 108}{5\big(17\varepsilon_{\mu_s}^2 - 45\varepsilon_{\mu_s} + 27\big)}.$$

As $\varepsilon_{\mu_s} \to 0$, the parameter μ_s scales as $\mu_s = O\Big(\sqrt{(1-\sigma_s^2)\,\varepsilon_{\mu_s}}\Big)$. While one might expect μ_s to depend only on ε_{μ_s}, this additional dependence on σ_s^2 is necessary in our ill-conditioned setting. This dependency can be seen by taking noise component variances to be a constant and expressing μ_s in terms of the other parameters in the second moment equations.

We claim that for every $\sigma_s^2 \in (0, 0.1)$, by varying the parameter ε_{μ_s} between $(0, 0.51]$ the parameter μ_s can be made to attain every value in $[0, 0.65)$. Moreover, for $\varepsilon_{\mu_s} \in (0, 0.51]$ we have for all $\sigma_s^2 \in (0, 0.1)$ that $\sigma_N^2 \in (0.01, 0.8)$.

These properties are verified symbolically using Mathematica Wolfram Research (2024); see Appendix H. Moreover, for each fixed $\sigma_s^2 \in (0, 0.1)$, the map $\varepsilon_{\mu_s} \mapsto \mu_s(\varepsilon_{\mu_s})$ is a bijection over $\varepsilon_{\mu_s} \in (0, 0.5]$; see Appendix H.

3. For $0.65 \leqslant |\mu_s| < \infty$, we use the following construction.

   Let

   $$\mu_s = \frac{1}{3\sqrt{1-\varepsilon_{\mu_s}}},$$

   which implies $0.7 \leqslant \varepsilon_{\mu_s} < 1$. Fix $k = \frac{4}{5}$ and define

   $$P_{3,\varepsilon_{\mu_s}} = (1-\varepsilon_{\mu_s}) \cdot \mathcal{N}(\mu_s, \sigma_s^2) + \varepsilon_{\mu_s}\left(\frac{(1/\varepsilon_{\mu_s} - 1)}{k^3}\,\mathcal{N}(-k\mu_s, v_2) + \left(1 - \frac{1-\varepsilon_{\mu_s}}{k^3\varepsilon_{\mu_s}}\right)\mathcal{N}(\mu_3, v_3)\right).$$

6. Our SymPy code is available in Appendix H.

where

$$\mu_3 = \frac{36(1-\varepsilon_{\mu_s})\mu_s}{189\varepsilon_{\mu_s} - 125}.$$

We claim that for every $\sigma_s^2 \in (0, 0.1]$ and $\varepsilon_{\mu_s} \in [0.7, 1)$,

$$v_2 \in (0.2, 1) \qquad \text{and} \qquad v_3 \in (0.7, 1.9),$$

so all the noise variances remain bounded by absolute constants. This is verified symbolically using Mathematica; see Appendix H.

In this construction we re-utilize symmetries from the second construction, such as scaling the means and the mixing weights together, and search over the appropriate scaling k, allowing simpler moment calculations. We then set up a linear system involving v_2 and v_3, expressed in terms of the weights and the signal mean parameterized by ε_{μ_s} using the second and third moments. Solving this linear system in SymPy gives us

$$v_2(\varepsilon_{\mu_s}, \sigma_s^2) = \frac{42525\,\varepsilon_{\mu_s}^2\sigma_s^2 - 58554\,\varepsilon_{\mu_s}\sigma_s^2 + 20125\,\sigma_s^2 + 5157\,\varepsilon_{\mu_s} - 3429}{25\,(9\varepsilon_{\mu_s} - 5)\,(189\varepsilon_{\mu_s} - 125)}$$

$$v_3(\varepsilon_{\mu_s}, \sigma_s^2) = \frac{3}{(9\varepsilon_{\mu_s} - 5)\,(189\varepsilon_{\mu_s} - 125)^2}\Big(107163\,\varepsilon_{\mu_s}^3\sigma_s^2 - 248913\,\varepsilon_{\mu_s}^2\sigma_s^2 + 188625\,\varepsilon_{\mu_s}\,\sigma_s^2 - 46875\,\sigma_s^2 + 36243\,\varepsilon_{\mu_s}^2 - 48102\,\varepsilon_{\mu_s} + 15955\Big).$$

which can be verified to be within the specified range using Mathematica, as mentioned earlier.

We remark that in the above constructions, to deal with negative means, we can simply reflect the constructions.

Lemma 28 *A_{μ_s} is well-defined for every $\mu_s \in \mathbb{R}$ and has first three moments $0, 1, 0$. Furthermore, whenever $|\mu_s| > \sqrt{\varepsilon}/10000$, we have*

$$\frac{\varepsilon_{\mu_s}}{1-\varepsilon_{\mu_s}} \leqslant 9\mu_s^2.$$

Proof. Based on our construction, we have two cases. In the first we have that

$$\mu_s^2 = \frac{3\varepsilon_{\mu_s}\big(20\varepsilon_{\mu_s}\sigma_s^2 - 4\varepsilon_{\mu_s} - 15\sigma_s^2 + 15\big)}{5\big(17\varepsilon_{\mu_s}^2 - 45\varepsilon_{\mu_s} + 27\big)},$$

which simplifies to

$$\frac{\varepsilon_{\mu_s}}{1-\varepsilon_{\mu_s}} \leqslant 5\mu_s^2,$$

for all choices of $\sigma_s^2 \leqslant 0.1$ and $\varepsilon_{\mu_s} \in [0, 0.5]$. We provide Mathematica code for verification in Appendix H. Now, for the last construction, we have that

$$\mu_s = \frac{1}{3(\sqrt{1-\varepsilon_{\mu_s}})}.$$

We conclude since

$$\frac{\varepsilon_{\mu_s}}{1-\varepsilon_{\mu_s}} \leqslant \frac{1}{1-\varepsilon_{\mu_s}} \leqslant 9\mu_s^2.$$

□

C.4. The Marginal and Corruption

In this section we formally define the marginal distribution over y. We then prove that the construction in Section C.3 can indeed be viewed as an ε-corruption of Q_v for every $v \in T$.

Following Diakonikolas et al. (2019c), we define

$$R(y) \propto \frac{G(y)}{1-\varepsilon_{\mu_s}},$$

where $G(y)$ is the standard 1-dimensional Gaussian probability density function. To see that this is an actual distribution, it suffices to show that the normalization factor is bounded. Let $a = 1/(10000c_1)$. Since $\mu_s = c_1\sqrt{\varepsilon}\,y$, Lemma 28 gives, on $|y| > a$,

$$\frac{\varepsilon_{\mu_s}}{1-\varepsilon_{\mu_s}} \leqslant 9c_1^2\varepsilon y^2.$$

Therefore for

$$C \coloneqq \int_{\mathbb{R}} \frac{G(y)}{1-\varepsilon_{\mu_s}}dy = \int_{\mathbb{R}} \frac{G(y)\varepsilon_{\mu_s}}{1-\varepsilon_{\mu_s}}dy + \int_{\mathbb{R}} G(y)dy$$

we have that

$$C \leqslant 1 + \frac{\varepsilon}{1-\varepsilon}\,\mathbb{P}(|Y| \leqslant a) + 9c_1^2\varepsilon\,\mathbb{E}[Y^2\mathbf{1}_{|Y|>a}].$$

By the standard Gaussian tail bound $\mathbb{E}[Y^2\mathbf{1}_{|Y|>a}] \leqslant (a^2+2)\,\mathbb{P}(|Y| > a)$ for $a \geqslant 1$. Since $c_1^2a^2 = 10^{-8}$, choosing c_1 and ε as fixed sufficiently small constants gives

$$9c_1^2\mathbb{E}[Y^2\mathbf{1}_{|Y|>a}] \leqslant \frac{1}{1-\varepsilon}\,\mathbb{P}(|Y| > a),$$

and hence $C \leqslant 1/(1-\varepsilon)$. We further note that $C \geqslant 1$ since

$$G(y) \leqslant G(y)/(1-\varepsilon_{\mu_s}),$$

which in particular implies

$$R(y) \leqslant G(y)/(1-\varepsilon_{\mu_s}).$$

Therefore we have

$$R(y) = \frac{G(y)}{(1-\varepsilon_{\mu_s})\cdot C} \geqslant \frac{(1-\varepsilon)\cdot G(y)}{(1-\varepsilon_{\mu_s})}.$$

Remark 29 *A careful reader might have noticed that the condition $\varepsilon\kappa \lesssim 1$ was crucial in the above proof. Indeed, if $\varepsilon\kappa = \omega(1)$, then $\sigma_y^2 = \omega(1)$ in our construction and the corresponding choice of $R(y) \propto G(\sigma_y y)/(1-\varepsilon_{\mu_s})$ would no longer produce a Huber contamination.*

Definition 30 *For every $v \in T$, define the corrupted distribution as*

$$Q'_v \coloneqq \frac{1}{(2\pi)^{(d-1)/2}} A_{\mu_s}(\langle v, x\rangle)\exp(-\|x - \langle v, x\rangle v\|^2/2)\cdot R(y).$$

Denoting the d-dimensional conditional distribution $Q'_v(X|Y=y)$ as $P_{\mu_s,v}$, we will now show that Q'_v is indeed an ε Huber contamination of Q_v (from Section C.2).

Lemma 31 *There exists E_v such that $Q'_v = (1-\varepsilon)Q_v + \varepsilon E_v$.*

Proof. By construction,

$$A_{\mu_s} \geqslant (1-\varepsilon_{\mu_s})\mathcal{N}(\mu_s, \sigma_s^2).$$

This implies that

$$P_{\mu_s,v} \geqslant \big(1-\varepsilon_{\mu_s}\big)\mathcal{N}(\mu_s v, I_d - vv^T + \sigma_s^2 vv^T\big).$$

Since $\mu_s = c_1\sqrt{\varepsilon}y$ and $\sigma_s^2 = \frac{1}{\kappa} - c_1^2\varepsilon$, we have

$$Q'_v \coloneqq P_{\mu_s,v}\cdot R(y) \geqslant (1-\varepsilon_{\mu_s})\mathcal{N}\left(c_1\sqrt{\varepsilon}yv, I_d - vv^T + \left(\frac{1}{\kappa} - c_1^2\varepsilon\right)vv^T\right)\cdot R(y).$$

Recall that by construction, we have that

$$R(y) \geqslant \frac{(1-\varepsilon)\cdot G(y)}{(1-\varepsilon_{\mu_s})}.$$

Using this, we have

$$Q'_v \geqslant (1-\varepsilon)\cdot\mathcal{N}\left(c_1\sqrt{\varepsilon}yv, I_d - vv^T + \left(\frac{1}{\kappa} - c_1^2\varepsilon\right)vv^T\right)\cdot G(y).$$

Now the RHS is precisely $(1-\varepsilon)Q_v(X,y)$ expressed as $(1-\varepsilon)Q_v(X|Y=y)Q_v(y)$. Therefore we conclude that the following inequality holds point wise.

$$Q'_v(X,y) \geqslant (1-\varepsilon)Q_v(X,y).$$

Now define

$$E_v(X,y) \coloneqq \frac{1}{\varepsilon}\big(Q'_v(X,y) - (1-\varepsilon)Q_v(X,y)\big).$$

Observe that $E_v(X,y) \geqslant 0$ for all (X,y) and that $E_v(X,y)$ integrates to 1. Therefore $E_v(X,y)$ is a probability distribution, concluding the proof. □

C.5. Bounds on the Pairwise Correlation

Recall that we need to bound $\chi_S(Q'_v, Q'_{v'})$ where $S = R(y)\cdot G(x)$. To that end, we have the following Lemma.

Lemma 32

$$\gamma = \chi_S(Q'_v, Q'_{v'}) \leqslant O(|\langle v, v'\rangle|^4\sqrt{\kappa}e^{O(1/\varepsilon)}),$$
$$\beta = \chi_S(Q'_v, Q'_v) \leqslant O(\sqrt{\kappa}e^{O(1/\varepsilon)}).$$

Proof. We begin by expanding out the expression for the χ^2 divergence.

$$\chi_S(Q'_v, Q'_{v'}) = \int_{x,y}\frac{Q'_v(x,y)}{S(x,y)}\frac{Q'_{v'}(x,y)}{S(x,y)}S(x,y)dxdy - 1 = \int_{x,y}\frac{Q'_v(x,y)}{S(x,y)}Q'_{v'}(x,y)dxdy - 1$$
$$= \int_{x,y}\left(\frac{P_{\mu_s,v}(x)R(y)}{G(x)R(y)}\right)\cdot\big(P_{\mu_s,v'}(x)R(y)\big)dxdy - 1$$

$$= \int_{\mathbb{R}} \chi_{\mathcal{N}(0,I_d)}\big(P_{\mu_s,v}, P_{\mu_s,v'}\big) R(y)dy$$
$$\leqslant |\langle v, v'\rangle|^4 \int_{\mathbb{R}} \big(\chi_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s})\big) R(y)dy.$$

where the last inequality follows from applying Lemma 21. From Lemma 25 we have that

$$\int_{\mathbb{R}} \Big(\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s})\Big) R(y)dy = \int_{|\mu_s|\leqslant\sqrt{\varepsilon}/10000} \Big(\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s})\Big) R(y)dy + \int_{|\mu_s|>\sqrt{\varepsilon}/10000} \Big(\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s})\Big) R(y)dy.$$

In the first case,

$$\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s}) \leqslant e^{O(1/\varepsilon)}.$$

And in the second case, we have

$$\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s}) \leqslant O(\sqrt{\kappa} e^{O(\max\{\mu_s^2, 1/\mu_s^2\})}).$$

Using this, we have

$$\int_{|\mu_s|\leqslant\sqrt{\varepsilon}/10000} \Big(\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s})\Big) R(y)dy \leqslant O(\sqrt{\kappa}) e^{O(1/\varepsilon)},$$

and,

$$\int_{|\mu_s|>\sqrt{\varepsilon}/10000} \Big(\chi^2_{\mathcal{N}(0,1)}(A_{\mu_s}, A_{\mu_s})\Big) R(y)dy \leqslant O(\sqrt{\kappa}) \int_{|\mu_s|>\sqrt{\varepsilon}/10000} e^{O(\max\{1/\varepsilon, \mu_s^2\})} R(y)dy$$
$$\leqslant O(\sqrt{\kappa} e^{O(1/\varepsilon)}) + O(\sqrt{\kappa}) \cdot \int_{|\mu_s|>\sqrt{\varepsilon}/10000} e^{O(\mu_s^2)} R(y)dy.$$

Now,

$$\int_{|\mu_s|>\sqrt{\varepsilon}/10000} e^{O(\mu_s^2)} R(y)dy \leqslant \int_{|\mu_s|>\sqrt{\varepsilon}/10000} e^{O(c_1^2\varepsilon y^2)} \frac{G(y)}{(1-\varepsilon_{\mu_s})} dy$$
$$\leqslant \int_{|\mu_s|>\sqrt{\varepsilon}/10000} \frac{e^{-\Omega(y^2)} \cdot (1-\varepsilon_{\mu_s}+\varepsilon_{\mu_s})}{(1-\varepsilon_{\mu_s})} dy$$
$$\leqslant \int_{|\mu_s|>\sqrt{\varepsilon}/10000} e^{-\Omega(y^2)} dy + \int_{|\mu_s|>\sqrt{\varepsilon}/10000} 9y^2 e^{-\Omega(y^2)} dy \leqslant O(1),$$

where we used the definition of $R(y)$, and that for c_1, ε sufficiently small we can indeed take $O(c_1^2\varepsilon) < \frac{1}{2}$. Therefore, putting it all together,

$$\chi_S(Q'_v, Q'_{v'}) \leqslant O(|\langle v, v'\rangle|^4 \sqrt{\kappa} e^{O(1/\varepsilon)}).$$

Taking $v = v'$ gives the result for β. □

To prove the bounds on the divergences, we will require the following facts.

Fact 33 *For distributions $\{X_i\}_{i=1}^n$ absolutely continuous with respect to D and for convex weights $\{w_i\}_{i=1}^n$ we have that*

$$\chi^2\left(\sum_{i=1}^n w_i X_i, D\right) = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \chi_D(X_i, X_j) \leqslant \sum_{i=1}^n \sum_{j=1}^n |\chi_D(X_i, X_j)|.$$

Fact 34

$$\chi^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = \frac{\sigma_2^2}{\sigma_1\sqrt{2\sigma_2^2 - \sigma_1^2}} e^{\frac{(\mu_1-\mu_2)^2}{2\sigma_2^2-\sigma_1^2}} - 1.$$

whenever $2\sigma_2^2 > \sigma_1^2$. In particular with respect to the standard Gaussian we obtain

$$\chi^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(0, 1)) = \frac{1}{\sigma_1\sqrt{2 - \sigma_1^2}} e^{\frac{\mu_1^2}{2-\sigma_1^2}} - 1.$$

Note that this places a bound on the variance σ_1^2.

Fact 35

$$\chi^2_{\mathcal{N}(0,1)}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = \frac{\exp\left(-\frac{\mu_1^2(\sigma_2^2-1)+2\mu_1\mu_2+\mu_2^2(\sigma_1^2-1)}{2\sigma_1^2(\sigma_2^2-1)-2\sigma_2^2}\right)}{\sqrt{\sigma_1^2 + \sigma_2^2 - \sigma_1^2\sigma_2^2}} - 1.$$

We are now ready to prove the following lemma. Since our constructions for A_{μ_s} are mixtures of Gaussians it suffices to control the pairwise correlations between the mixture components with respect to the standard Gaussian using Fact 33. We next prove the following bound on the pairwise correlation for each part of our construction.

Lemma 36 *For $P_{1,\varepsilon_{\mu_s}}$, $P_{2,\varepsilon_{\mu_s}}$ and $P_{3,\varepsilon_{\mu_s}}$ as defined in Definition 26,*

$$\chi^2(P_{1,\varepsilon_{\mu_s}}, \mathcal{N}(0,1)) \leqslant O(\exp(O(1/\varepsilon))),$$

$$\chi^2(P_{2,\varepsilon_{\mu_s}}, \mathcal{N}(0,1)) \leqslant O\left(\sqrt{\kappa} e^{O\left(\max\{\mu_s^2, 1/\mu_s^2\}\right)}\right).$$

$$\chi^2(P_{3,\varepsilon_{\mu_s}}, \mathcal{N}(0,1)) \leqslant O\left(\sqrt{\kappa} e^{O\left(\max\{\mu_s^2, 1/\mu_s^2\}\right)}\right).$$

Proof. From Fact 34 first note that the signal component in our case has χ^2 divergence

$$\chi^2\left(\mathcal{N}(\mu_s, \sigma_s^2), \mathcal{N}(0,1)\right) = \frac{1}{\sigma_s\sqrt{2-\sigma_s^2}} e^{\frac{\mu_s^2}{2-\sigma_s^2}} - 1 = \frac{1}{\sigma_s\sqrt{2-\sigma_s^2}} e^{O(\mu_s^2)} - 1 = O\left(\sqrt{\kappa} e^{O(\mu_s^2)}\right) - 1,$$

since $\sigma_s^2 \leqslant 0.1$ and $\frac{1}{\sigma_s^2} = \frac{1}{1/\kappa - c_1^2\varepsilon} \leqslant \frac{\kappa}{1-c_1^2\varepsilon\kappa} \leqslant O(\kappa)$. We first consider $P_{1,\varepsilon_{\mu_s}}$. The noise components all have variance 1. So the χ^2 divergence with respect to the standard Gaussian is at most $e^{O(\mu_i^2)}$ for each of these noise components. There are two types of pairwise correlations: one for those with the signal component and the noise component, and between the noise components. From Fact 35, the correlation between the signal component and the noise components is at most $O(e^{O(\mu_i^2)})$ and

within the noise components it is $O(e^{O(\mu_i\mu_j)})$. Since we know that $\max_i |\mu_i| = O(\frac{1}{\sqrt{\varepsilon}})$, the total χ^2 divergence for this component is,

$$O\Big(\sqrt{\kappa}e^{O(\mu_s^2)}\Big) + O(e^{O(1/\varepsilon)}) = O(\sqrt{\kappa}) + O(e^{O(1/\varepsilon)}).$$

However since $\varepsilon\kappa \lesssim 1$ we have that

$$O(\sqrt{\kappa}) < O(\sqrt{1/\varepsilon}) < O(e^{O(1/\varepsilon)}),$$

and hence the χ^2 divergence for $P_{1,\varepsilon_{\mu_s}}$ is bounded by $e^{O(1/\varepsilon)}$.

In $P_{2,\varepsilon_{\mu_s}}$ the noise component means scale as $O(1/\mu_s)$ since $\varepsilon_{\mu_s} \approx \mu_s^2$ in this construction. In terms of variances, we have $\tau^2 = 0.2$ and $\sigma_N^2 > 0$ is always a constant bounded away from 0. This implies that the individual χ^2 divergence contributions for the noise components scales with the signal mean μ_s as $O(e^{O(1/\mu_s^2)})$. The pairwise correlation between the noise components scales similarly as $O(e^{O(1/\mu_s^2)})$ and between the signal and the noise component scales as $O(e^{O\left(\max\{\mu_s^2, 1/\mu_s^2\}\right)})$. Putting it together, we have that the χ^2 divergence for this component scales as

$$O\Big(\sqrt{\kappa}e^{O(\mu_s^2)}\Big) + O(e^{O(1/\mu_s^2)}) + O\Big(e^{O\left(\max\{\mu_s^2, 1/\mu_s^2\}\right)}\Big) = O\Big(\sqrt{\kappa}e^{O\left(\max\{\mu_s^2, 1/\mu_s^2\}\right)}\Big).$$

In $P_{3,\varepsilon_{\mu_s}}$ the two noise components have means one of which scales as μ_s and the other is much smaller than μ_s and closer to the origin. Furthermore, by construction the noise components have constant variances. This implies that the pairwise correlation among the noise components and the individual χ^2 divergences for the noise terms scales as $O(e^{O(\mu_s^2)})$. The signal component is the dominant mean term here as $P_{3,\varepsilon_{\mu_s}}$ deals with large y. The pairwise correlation between the signal and the noise components also scales as $O(e^{O(\mu_s^2)})$ since the dominant mean term is that of the signal. Therefore putting all pieces together we have that the χ^2 divergence for this component scales as

$$O\Big(\sqrt{\kappa}e^{O(\mu_s^2)}\Big) + O(e^{O(\mu_s^2)}) = O\Big(\sqrt{\kappa}e^{O(\mu_s^2)}\Big).$$

□

C.6. Putting the pieces together

Proof of Lemma 25. The construction in Definition 26, together with Lemma 27, the symbolic moment calculations for P_2 and P_3, and reflection for negative μ_s, shows that A_{μ_s} has the desired mixture form and matches the first three moments of $\mathcal{N}(0,1)$. The bound on $\varepsilon_{\mu_s}/(1-\varepsilon_{\mu_s})$ follows from Lemma 28. The stated χ^2 bounds follow from Lemma 36. □

Appendix D. Low-Degree Lower Bound

We begin with some background on the low-degree polynomial tests which are very closely related to Statistical Query algorithms Brennan et al. (2021).

Definition 37 *Consider a testing problem of distinguishing $H_0 : y \sim \mathcal{Q}$ and $H_1 : y \sim \mathcal{P}$ for input $y \in \mathbb{R}^{n\times d}$. The degree-$D$ advantage of the testing problem is defined as*

$$\mathrm{Adv}^{\leqslant D} = \max_{f\in\mathbb{R}[z]_{\leqslant D}} \frac{\mathbb{E}_{\mathcal{P}}[f(z)]}{\sqrt{\mathbb{E}_{\mathcal{Q}}[f(z)^2]}} = \sqrt{1 + \left(\max_{f\in\mathbb{R}[z]_{\leqslant D}} \frac{\mathbb{E}_{\mathcal{P}}[f(z)] - \mathbb{E}_{\mathcal{Q}}[f(z)]}{\sqrt{\mathbb{V}_{\mathcal{Q}}[f(z)]}}\right)^2}$$

where the maximum is taken over all degree-D polynomials f over the $n \times d$ dimensional input y.

This method for proving lower bounds against low-degree polynomials is supported by the low-degree conjecture of Hopkins Hopkins (2018). The conjecture posits that if $\mathrm{Adv}^{\leqslant D} = O(1)$ for $D = \Omega(\mathrm{poly}\log n)$, then no polynomial-time algorithm exists for the above distinguishing task. We remark that a recent work of Buhai et al. (2025) finds a counterexample for this conjecture. Despite the caution offered by this work, the low-degree method remains widely useful for obtaining initial evidence of computational hardness. We refer the reader to Kunisky et al. (2019); Wein (2025) for more details about this method and the wide applicability of the method in providing evidence for hardness in statistical problems. In the rest of this section, we shall show that the degree-D advantage is bounded for our testing problem. We start with some facts about Hermite polynomials that we will make extensive use of throughout the rest of this section.

Hermite polynomials. Hermite polynomials are orthogonal polynomials that form a complete orthogonal basis of the vector space $\mathcal{L}^2(\mathbb{R}, \mathcal{N}(0,1))$ for all functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[f^2(X)] < \infty$ (See for example Szegö (1939)). We will make use of the normalized version of the probabilist's Hermite polynomials. We denote the k^{th} normalized probabilist's Hermite polynomials as $h_k, k \in \mathbb{N}$. These normalized polynomials h_k are defined using the probabilist's Hermite polynomials He_k as follows

$$h_k(x) := \frac{\mathrm{He}_k(x)}{\sqrt{k!}}.$$

Following Mao and Wein (2025) we will also make use of multivariate versions of the above polynomials: $H_\alpha : \mathbb{R}^{n\times d} \to \mathbb{R}$, indexed by the multi-index $\alpha \in \mathbb{N}^{n\times d}$ and defined as

$$H_\alpha(z) = \prod_{i=1}^{n}\prod_{j=1}^{d} h_{\alpha_{i,j}}(z_{i,j})$$

Taking $|\alpha| = \sum_{i=1}^{n}\sum_{j=1}^{d}\alpha_{i,j}$ we have that $\{H_\alpha\}_{|\alpha|\leqslant D}$ is a basis for the subspace of polynomials $\mathbb{R}^{n\times d} \to \mathbb{R}$ of degree at most D. If z has iid $\mathcal{N}(0,1)$ entries we further have that these polynomials are orthonormal in the following sense

$$\mathbb{E}[H_\alpha(z)H_\beta(z)] = \mathbb{1}[\alpha = \beta]$$

We are now ready to state our lower-bound instance and prove the result.

D.1. Main result

Problem 38 (Robust Linear Testing in Mahalanobis Norm) *Given corruption rate $\varepsilon \in (0, 1/2)$, $\kappa \geqslant 1$, signal strength $\alpha \in \mathbb{R}_+$, sample size $n \in \mathbb{N}$, noise variance $\sigma^2 > 0$, dimension $d \in \mathbb{N}$, define $\sigma_y^2 := \alpha^2 + \sigma^2$ and consider the following hypothesis testing problem with input samples $\{z_i\}_{i=1}^n \in \mathbb{R}^{n\times(d+1)}$.*

1. *H_0: Null $\mathcal{Q}$: Let $X \sim \mathcal{N}(0, I_d)$ and $Y \sim \mathcal{N}(0, \sigma_y^2)$ be independent. Define $Z := (X, Y) \in \mathbb{R}^{d+1}$. Then*

$$z_1, \ldots, z_n \overset{\text{iid}}{\sim} \mathcal{Q} \overset{\text{def}}{=} \mathcal{N}\big(0, \mathrm{diag}\big(I_d, \sigma_y^2\big)\big).$$

2. *H_1: Alternative $\mathcal{P}$: $v \sim \mathrm{Unif}(\mathcal{S}^{d-1})$. Conditioned on v,*

$$z_1, \dots, z_n \overset{\text{iid}}{\sim} \mathcal{P} \overset{\text{def}}{=} (1-\varepsilon) \cdot D_v + \varepsilon \cdot E$$

where D_v is the distribution of (X, Y) for $Y = \langle X, \beta_v \rangle + \eta$ for $X \sim \mathcal{N}(0, \Sigma_v)$ for $\Sigma_v \succ 0$ with condition number κ and $\eta \sim N(0, \sigma^2)$ independent of X. Here E is an arbitrary adversarial distribution on $\mathbb{R}^{d+1}$ that may depend on v. The Mahalanobis norm of β_v satisfies

$$\|\Sigma_v^{1/2} \beta_v\| = \alpha.$$

Furthermore, the marginal variance of Y under the inlier distribution D_v matches that under the null:

$$\mathrm{Var}_{D_v}(Y) = \alpha^2 + \sigma^2 = \sigma_y^2$$

We now define a hard instance for the above problem.

Definition 39 (Hard Instance for Robust Testing) *We define a hard instance for the Robust Testing in Mahalanobis Norm Problem. For $\varepsilon > 0, \kappa, \alpha, n, d, \sigma$, the distribution in $\mathcal{Q}$ is uniquely defined. We define the distribution $\mathcal{P}$ as follows: for $v \sim \mathrm{Unif}(S^{d-1})$ as, $\beta_v = \delta v$, $\Sigma_v = I_d - (1 - 1/\kappa) v v^T$ where $\delta = \alpha\sqrt{\kappa}$. Let $E(X, y) = \mathcal{N}(0_{d+1}, \Sigma)$ where,*

$$\Sigma = \begin{bmatrix} I_d + \frac{(1-\varepsilon)}{\varepsilon} \cdot (1 - 1/\kappa) v v^T & -\frac{(1-\varepsilon)}{\varepsilon} \frac{\delta}{\kappa} v \\ -\frac{(1-\varepsilon)}{\varepsilon} \frac{\delta}{\kappa} v^T & \frac{\delta^2}{\kappa} + \sigma^2 \end{bmatrix}.$$

Remark 40 *In Lemma 52 we show that Σ defined above is positive semi-definite. We further note that Σ is chosen so that the distribution P has the same first three moments as that of Q. We show this in Lemma 53.*

We will now prove the main result for this section.

Theorem 41 *Let $\kappa \geqslant 1$ and $\varepsilon \gg \frac{1}{\sqrt{d}}$ be such that $\varepsilon\kappa \geqslant 1 - \varepsilon$. For every $\sigma^2 > 0$ and $\alpha > 0$, there exists a choice of $\beta_v, \Sigma_v, E(X, y)$ such that the degree-D advantage of Problem 38 is $1 + o(1)$ for*

$$n \ll \frac{1}{\mathrm{poly}(D)} \min\left(d\varepsilon^2\kappa^2, \varepsilon^2 d^2\right).$$

In particular, the choice for β_v, Σ_v, E is the same as the one in Definition 39.

Remark 42 *Our joint assumption on ε, κ is a direct result of requiring PSD-ness of the covariance matrix of the corruption distribution in our hard instance.*

We claim now that the proof of Theorem 41 follows from the following lemma.

Lemma 43 *The squared degree D advantage of Problem 38 is at most the following for the instance in Definition 39.*

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 - 1 \lesssim \sum_{\substack{p=4 \\ p\text{even}}}^{D} \sum_{m=1}^{p/4} n^m p^{p/2} \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{2m-p} p^{p/2} \sum_{L=0}^{p/2} \left(\frac{\sqrt{d}}{\kappa}\right)^L.$$

We first prove Theorem 41.

Proof of Theorem 41. We consider two cases depending on which of $d\varepsilon^2\kappa^2$ and $\varepsilon^2 d^2$ is smaller.

- Case 1: $\varepsilon^2 d^2 > d\varepsilon^2\kappa^2$, or equivalently $\kappa < \sqrt{d}$. In this case, the innermost geometric sum is diverging and is dominated by the largest term. Therefore the geometric sum can be bounded by $O\left(\left(\frac{\sqrt{d}}{\kappa}\right)^{p/2}\right)$. Using this, the sum reduces to

$$\begin{aligned}
\left(\mathrm{Adv}^{\leqslant D}\right)^2 - 1 &\lesssim \sum_{\substack{p=4\\ p \text{ even}}}^{D} \sum_{m=1}^{p/4} n^m p^{p/2} \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{2m-p} p^{p/2} \left(\frac{\sqrt{d}}{\kappa}\right)^{p/2} \\
&= \sum_{\substack{p=4\\ p \text{ even}}}^{D} p^p \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{-p} \left(\frac{\sqrt{d}}{\kappa}\right)^{p/2} \sum_{m=1}^{p/4} \left(n\varepsilon^2\right)^m \\
&\lesssim \sum_{\substack{p=4\\ p \text{ even}}}^{D} p^p \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{-p} \left(\frac{\sqrt{d}}{\kappa}\right)^{p/2} \left(n\varepsilon^2\right)^{p/4} \\
&= \sum_{\substack{p=4\\ p \text{ even}}}^{D} \left(\frac{16p^6 dn\varepsilon^2}{d^2\varepsilon^4\kappa^2}\right)^{p/4} = \sum_{\substack{p=4\\ p \text{ even}}}^{D} \left(\frac{16p^6 n}{d\varepsilon^2\kappa^2}\right)^{p/4}.
\end{aligned}$$

  When $n \ll \frac{1}{\mathrm{poly}(D)}\left(d\varepsilon^2\kappa^2\right)$, the right hand side above is $o(1)$, thus enabling us to bound the advantage by $1 + o(1)$.

- Case 2: $\varepsilon^2 d^2 \leqslant d\varepsilon^2\kappa^2$, or equivalently $\kappa \geqslant \sqrt{d}$ [7]. In this case, the inner most geometric series is bounded by at most a constant. This implies that we have

$$\begin{aligned}
\left(\mathrm{Adv}^{\leqslant D}\right)^2 - 1 &\lesssim \sum_{\substack{p=4\\ p \text{ even}}}^{D} \sum_{m=1}^{p/4} n^m p^{p/2} \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{2m-p} p^{p/2} \\
&\lesssim \sum_{\substack{p=4\\ p \text{ even}}}^{D} p^p \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{-p} \left(n\varepsilon^2\right)^{p/4} \\
&= \sum_{\substack{p=4\\ p \text{ even}}}^{D} \left(\frac{16p^6 n\varepsilon^2}{d^2\varepsilon^4}\right)^{p/4} = \sum_{\substack{p=4\\ p \text{ even}}}^{D} \left(\frac{16p^6 n}{d^2\varepsilon^2}\right)^{p/4}.
\end{aligned}$$

  which is $o(1)$ whenever $n \ll \frac{1}{\mathrm{poly}(D)}\left(d^2\varepsilon^2\right)$. Taking the minimum concludes the proof.

□

7. We carry out the calculations for $\kappa > \sqrt{d}$. The $\kappa = \sqrt{d}$ case only adds a polynomial factor of $O(p)$, so the same conclusion holds.

Proof of Lemma 43

Lemma 43 *The squared degree D advantage of Problem 38 is at most the following for the instance in Definition 39.*

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 - 1 \lesssim \sum_{\substack{p=4\\ p\text{even}}}^{D} \sum_{m=1}^{p/4} n^m p^{p/2} \left(\frac{4p}{d}\right)^{p/2} \varepsilon^{2m-p} p^{p/2} \sum_{L=0}^{p/2} \left(\frac{\sqrt{d}}{\kappa}\right)^L.$$

The proof of the above lemma is obtained by combining multiple intermediate lemmas. We first outline the proof and then present the mathematical details. As a first step we will express the squared advantage in terms of expectations over Hermite polynomials. To do so, we introduce a new, more general result, stated as Lemma 49. This result can be seen as a generalization of (Mao and Wein, 2025, Lemma 6.4). Then, we will compute a closed form expression for these expectations using exponential generating series for our case. The rest of the argument will be primarily combinatorial and will concern the split of the degrees among the samples. We are now ready to prove Lemma 43. We split our proof internally into different sections to make it more amenable for the reader to parse the proof.

Proof.

We begin by relating the advantage to the Hermite coefficients. We first apply Lemma 49 to obtain that the squared advantage is

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 = \sum_{p=0}^{D} \sum_{\substack{|\alpha|=p\\ \alpha \text{ supp. on } \{1,d+1\}}} \mathop{\mathbb{E}}_{u,u'} [\langle u, u'\rangle^{\sum_{i=1}^n \alpha_{i,1}}] \cdot \prod_{i=1}^{n} \left(\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{P}'_u} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right)\right]\right)^2.$$

where u, u' are random vectors drawn from $\mathcal{S}^{d-1}$ and $\mathcal{P}'_u$ is the two-dimensional distribution along u and y. In our case, u is the hidden direction (which we represent using v in the hard instance) and $s = \delta^2/\kappa + \sigma^2$. Using Fact 50 about correlation between random vectors we bound the advantage as

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 \leqslant \sum_{p=0}^{D} \sum_{\substack{|\alpha|=p\\ \alpha \text{ supp. on } \{1,d+1\}}} \left(\frac{p}{d}\right)^{(\sum_{i=1}^n \alpha_{i,1})/2} \cdot \prod_{i=1}^{n} \left(\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{P}'_u} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right)\right]\right)^2.$$

whenever $\sum_{i=1}^n \alpha_{i,1}$ is even. We will use this implicitly in the remainder of the proof.

By our construction the first three moments of the alternate distribution agree with those of the null distribution. Therefore the expectation vanishes for any polynomial up to degree 3. In particular, this implies that the squared advantage is non-zero only when each sample that receives a non-zero degree has an even degree of at least 4 as the Hermite coefficient factorizes across samples and any sample with degree $1, 2, 3$ has zero coefficient. Furthermore, the alternate distribution is symmetric by construction, ensuring that all odd moments also match those of the null distribution and equal zero. Therefore, to bound the advantage, it is sufficient to consider even degrees $p \geqslant 4$. Thus we have

$$\begin{aligned}&\left(\mathrm{Adv}^{\leqslant D}\right)^2\\ &\leqslant 1 + \sum_{\substack{p=4\\ p \text{ even}}}^{D} \sum_{\substack{|\alpha|=p\\ \alpha \text{ supp. on } \{1,d+1\}}} \left(\frac{p}{d}\right)^{(\sum_{i=1}^n \alpha_{i,1})/2} \cdot \prod_{i=1}^{n} \left(\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{P}'_u} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right)\right]\right)^2.\end{aligned}$$

Henceforth we will use the following notation for convenience: We will replace $\alpha_{i,1}$ by k_i and $\alpha_{i,d+1}$ by ℓ_i and the total degree of sample i by $d_i = k_i + \ell_i$ and define $L = \sum_{i=1}^n \ell_i$ and $K = \sum_{i=1}^n k_i$. We will now make use of the following lemma that enables us to obtain a closed form expression for the expectation of Hermite polynomials.

A closed form for the Hermite coefficients.

Lemma 45 *Let $(x, y) \sim D$ for $D = \mathcal{N}\left(0, \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}\right)$. Then, we have*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\left[h_k(x) \cdot h_\ell\left(\frac{y}{\sigma_y}\right)\right] = \sqrt{\frac{k!}{\ell!}} \cdot \frac{1}{((k-\ell)/2)! \cdot 2^{(k-\ell)/2}} \cdot (\sigma_x^2 - 1)^{(k-\ell)/2} \cdot \left(\frac{\sigma_{xy}}{\sigma_y}\right)^\ell.$$

for $k \geqslant \ell$ and $k - \ell$ even and 0 otherwise.

Lemma 45 enables us to compute a closed form equation for the expectation of the Hermite polynomials for our two-dimensional mixture. We provide a proof for this lemma in Section D.3 where we utilize exponential generating functions for Hermite polynomials. We first focus on the terms in the lemma that depend only on the variances and then later deal with the combinatorial coefficients. From Definition 39 we recall that we have for the component with weight $(1-\varepsilon)$ that

$$\sigma_x^2 = \frac{1}{\kappa}, \quad \sigma_{xy} = \frac{\delta}{\kappa}, \quad \sigma_y^2 = \frac{\delta^2}{\kappa} + \sigma^2,$$

which implies

$$\frac{\sigma_{xy}}{\sigma_y} = \frac{\delta}{\kappa\sqrt{\delta^2/\kappa + \sigma^2}} \leqslant \frac{\delta}{\kappa\sqrt{\delta^2/\kappa}} = \frac{1}{\sqrt{\kappa}}.$$

Now for the component with weight ε we have

$$\sigma_x^2 = \frac{1}{\varepsilon} \cdot \left(1 - \frac{1}{\kappa}(1-\varepsilon)\right), \quad \sigma_{xy} = -\frac{1-\varepsilon}{\varepsilon}\frac{\delta}{\kappa}, \quad \sigma_y^2 = \frac{\delta^2}{\kappa} + \sigma^2$$

which implies that

$$\left|\frac{\sigma_{xy}}{\sigma_y}\right| = \frac{(1-\varepsilon)\cdot\delta}{\varepsilon \cdot \kappa \cdot \sqrt{\delta^2/\kappa + \sigma^2}} \leqslant \frac{\delta}{\varepsilon\kappa\sqrt{\delta^2/\kappa}} = \frac{1}{\varepsilon\sqrt{\kappa}}.$$

We further note that $\sigma_x^2 \leqslant \frac{1}{\varepsilon}$ for the second component and we will use this bound going forward. Putting it together we have the contribution from only the variance terms being

$$\begin{aligned} &= (1-\varepsilon) \cdot \left(\frac{1}{\kappa} - 1\right)^{(k-\ell)/2} \cdot \left(\frac{1}{\sqrt{\kappa}}\right)^\ell + \varepsilon \cdot \left(\frac{1}{\varepsilon}\right)^{(k-\ell)/2} \cdot \left(\frac{1}{\varepsilon\sqrt{\kappa}}\right)^\ell \\ &\lesssim \varepsilon \cdot \left(\frac{1}{\varepsilon}\right)^{(k-\ell)/2} \cdot \left(\frac{1}{\varepsilon\sqrt{\kappa}}\right)^\ell \end{aligned}$$

assuming $\kappa \geqslant 1$ and ε sufficiently small and by using that the absolute value of $1/\kappa - 1$ is at most a constant. Combining this with the combinatorial coefficients, we have

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}'_u}\left[h_k(x) \cdot h_\ell\left(\frac{y}{\sqrt{s}}\right)\right]\right| \lesssim \underbrace{\sqrt{\frac{k!}{\ell!}} \cdot \frac{1}{((k-\ell)/2)! \cdot 2^{(k-\ell)/2}}}_{:=C_{k,\ell}} \cdot \varepsilon \cdot \left(\frac{1}{\varepsilon}\right)^{(k-\ell)/2} \cdot \left(\frac{1}{\varepsilon\sqrt{\kappa}}\right)^\ell$$

Squaring and substituting into the advantage equation, we obtain

$$\begin{aligned}&\left(\mathrm{Adv}^{\leqslant D}\right)^2 - 1\\ &\lesssim \sum_{\substack{p=4\\ p\text{ even}}}^{D} \sum_{\substack{\sum_{i=1}^n d_i = p\\ \forall i\in[n]\colon d_i = k_i+\ell_i\\ \forall i\in[n]\colon k_i\geqslant \ell_i\\ \forall i\in[n]\colon k_i-\ell_i \text{ even}}} \left(\frac{p}{d}\right)^{\left(\sum_{i=1}^n k_i\right)/2} \cdot \left(\prod_{i=1}^n (C_{k_i,\ell_i})^2 \cdot \left(\varepsilon^2 \cdot \left(\frac{1}{\varepsilon}\right)^{k_i-\ell_i} \cdot \left(\frac{1}{\varepsilon^2\kappa}\right)^{\ell_i}\right)\right)\end{aligned}$$

Degree split, done globally. We will now concern ourselves with splitting the degrees globally among the samples. We now count the number of ways of dividing a given degree p among the samples so that each sample with a non-zero degree receives degree at least 4. We observe that this is the same as the cardinality of the following set.

Lemma 46 (Lemma 6.7 in Mao and Wein (2025)) *Consider the following set.*

$$S(p,m) = \left\{\beta \in \mathbb{N}^n : \sum_{i=1}^n \beta_i = p, \|\beta\|_0 = m, \beta_i \in \{0,4,6,8\dots\} \forall\, i \in [n]\right\}$$

Then $|S(p,m)| \leqslant n^m \cdot p^{p/2}$

$|S(p,m)|$ counts the number of ways of distributing degree p among m samples from a total of n samples. For the next parts of the computations, we will fix a specific subset of samples S that has cardinality m. The subset S satisfies that $\forall i \in S : d_i > 0$ and $\forall i \notin S : d_i = 0$. We will finally take a summation over all choices of m to finish our computations. We will now bound the advantage for a fixed choice of p, m and $\{d_i\}_{i\in S}$. Therefore for a fixed $p, m, \{d_i\}_{i\in S}$ we have the contribution to the advantage being at most

$$\begin{aligned}&\sum_{\substack{\forall i\in S: k_i+\ell_i = d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}} \left(\frac{p}{d}\right)^{\left(\sum_{i=1}^n k_i\right)/2} \cdot \left(\prod_{i\in S} (C_{k_i,\ell_i})^2 \cdot \left(\varepsilon^2 \cdot \left(\frac{1}{\varepsilon}\right)^{k_i-\ell_i} \cdot \left(\frac{1}{\varepsilon^2\kappa}\right)^{\ell_i}\right)\right)\\ &= \sum_{\substack{\forall i\in S: k_i+\ell_i = d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}} \left(\frac{p}{d}\right)^{(p-L)/2} \cdot \left(\varepsilon^{2m} \cdot \left(\frac{1}{\varepsilon}\right)^{p-2L} \cdot \left(\frac{1}{\varepsilon^2\kappa}\right)^{L}\right) \cdot \left(\prod_{i\in S} (C_{k_i,\ell_i})^2\right)\\ &= \left(\frac{p}{d}\right)^{p/2} \varepsilon^{2m-p} \sum_{\substack{\forall i\in S: k_i+\ell_i = d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}} \left(\frac{\sqrt{d}}{\sqrt{p}\kappa}\right)^{L} \cdot \left(\prod_{i\in S} (C_{k_i,\ell_i})^2\right)\\ &\leqslant \left(\frac{p}{d}\right)^{p/2} \varepsilon^{2m-p} \sum_{\substack{\forall i\in S: k_i+\ell_i = d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}} \left(\frac{\sqrt{d}}{\kappa}\right)^{L} \cdot \left(\prod_{i\in S} (C_{k_i,\ell_i})^2\right).\end{aligned}$$

where the last inequality follows since $p \geqslant 4$.

Bounding the Combinatorial Coefficients. We will now compute a closed form for the combinatorial coefficients that appear in our computations.

Lemma 47 *For a fixed choice of $p, m, \{d_i\}_{i\in S}$, we have*

$$\prod_{i\in S}(C_{k_i,\ell_i})^2 = \prod_{i\in S}\left(\sqrt{\frac{k_i!}{\ell_i!}}\cdot\frac{1}{((k_i-\ell_i)/2)!\cdot 2^{(k_i-\ell_i)/2}}\right)^2 \leqslant 2^p$$

We provide a proof of Lemma 47 in Section D.3. The proof follows by applying Stirling's approximation and univariate calculus. By an application of Lemma 47 the contribution to the advantage for a fixed choice of $p, m, \{d_i\}_{i\in S}$ is at most

$$\left(\frac{p}{d}\right)^{p/2}\varepsilon^{2m-p}\sum_{\substack{\forall i\in S: k_i+\ell_i=d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}}\left(\frac{\sqrt{d}}{\kappa}\right)^L\cdot 2^p = \left(\frac{4p}{d}\right)^{p/2}\varepsilon^{2m-p}\sum_{\substack{\forall i\in S: k_i+\ell_i=d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}}\left(\frac{\sqrt{d}}{\kappa}\right)^L$$

Degree split, within the samples. We now observe that the terms inside the summation only depends on L and not the individual $\{\ell_i\}_{i=1}^n$ themselves. Recall that since $k_i \geqslant \ell_i$, each $\ell_i \leqslant d_i/2$ and thus $L \leqslant p/2$. Using this we rewrite the summation in terms of L as follows:

$$\sum_{\substack{\forall i\in S: k_i+\ell_i=d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}}\left(\frac{\sqrt{d}}{\kappa}\right)^L = \sum_{L=0}^{p/2}\sum_{\substack{\sum_{i\in S}\ell_i=L\\ k_i+\ell_i=d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}}\left(\frac{\sqrt{d}}{\kappa}\right)^L = \sum_{L=0}^{p/2}\left(\left(\frac{\sqrt{d}}{\kappa}\right)^L\cdot\left(\sum_{\substack{\sum_{i\in S}\ell_i=L\\ k_i+\ell_i=d_i\\ k_i\geqslant\ell_i\\ k_i-\ell_i \text{ even}}}1\right)\right)$$

The contribution of the innermost summation is at most the number of ways of splitting a fixed $\{d_i\}_{i\in S}$ into k_i and ℓ_i subject to the constraints on k_i and ℓ_i. Since ℓ_i is determined once k_i is fixed and $k_i \geqslant d_i/2$, the number of choices for each chosen sample i is at most $(d_i/2+1)$ and we have totally m samples. Therefore, across all m chosen samples, the number of choices is at most

$$\prod_{i\in S}\left(\frac{d_i}{2}+1\right) \leqslant \prod_{i\in S} p \leqslant p^m \leqslant p^{p/4}$$

where we used $p \geqslant 4m$. Therefore for a fixed choice of $p, m, \{d_i\}_{i\in S}$ we have that the above quantity is bounded by

$$\left(\frac{4p}{d}\right)^{p/2}\varepsilon^{2m-p}p^{p/2}\sum_{L=0}^{p/2}\left(\frac{\sqrt{d}}{\kappa}\right)^L.$$

Putting it together. Combining the above steps, we have that the total squared advantage over all choices of p, m and the $\{d_i\}_{i\in S}$ as follows.

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 - 1 \lesssim \sum_{\substack{p=4\\ p\text{even}}}^{D}\sum_{m=1}^{p/4} n^m p^{p/2}\left(\frac{4p}{d}\right)^{p/2}\varepsilon^{2m-p}p^{p/2}\sum_{L=0}^{p/2}\left(\frac{\sqrt{d}}{\kappa}\right)^L.$$

This concludes the proof of Lemma 43.

□

D.2. A general testing problem.

We begin with a more general testing problem, of which our original problem is a special case.

Problem 48 *Let ν be a distribution over $\mathbb{R}^2$. Let $s > 0$. We define the following null and planted distributions.*

- *Under $\mathcal{Q}$, we observe i.i.d. samples $z_1, \dots, z_n \in \mathbb{R}^d \times \mathbb{R}$ with*

$$z_i = (x_i, y_i), \qquad x_i \sim \mathcal{N}(0, I_d), \quad y_i \sim \mathcal{N}(0, s),$$

  where x_i and y_i are independent.

- *Under $\mathcal{P}$, we first draw $v \sim \mathrm{Unif}(\mathcal{S}^{d-1})$. Conditional on v, we then draw i.i.d. samples $z_1, z_2, \dots, z_n$ as follows:*

$$\begin{aligned} (a_i, b_i) &\sim \nu \\ g_i &\sim \mathcal{N}(0, I_d) \quad \textit{independently,} \\ x_i &= a_i\, v + (I_d - vv^\top) g_i, \\ y_i &= b_i, \end{aligned}$$

  and set $z_i = (x_i, y_i)$.

Equivalently, under $\mathcal{P}$, the pair $(\langle x_i, v\rangle, y_i)$ has distribution ν, while the component of x_i orthogonal to v is standard Gaussian.

Lemma 49 *Consider the distribution $\mathcal{P}$ in Problem 48. Suppose the first D moments of ν are finite. For $\alpha \in \mathbb{N}^{n\times(d+1)}$, let $|\alpha| = \sum_{i=1}^n \sum_{j=1}^{d+1} \alpha_{i,j}$. Then we have*

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 = \sum_{p=0}^{D} \sum_{\substack{|\alpha|=p \\ \alpha \text{ supp. on } \{1,d+1\}}} \mathop{\mathbb{E}}_{u,u'} [\langle u, u'\rangle^{\sum_{i=1}^n \alpha_{i,1}}] \cdot \prod_{i=1}^n \left(\mathop{\mathbb{E}}_{(x,y)\sim\nu} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right)\right]\right)^2.$$

for u, u' uniformly chosen random vectors from the unit sphere.

Proof. We begin by recalling that the degree D advantage is defined as

$$\mathrm{Adv}^{\leqslant D} = \max_{f \in \mathbb{R}[z]_{\leqslant D}} \frac{\mathbb{E}_{\mathcal{P}}[f(z)]}{\sqrt{\mathbb{E}_{\mathcal{Q}}[f(z)^2]}}$$

It is a well-known fact that the orthogonal projection of the likelihood ratio L to the set of all polynomials of degree at most D maximizes the above quantity and that the optimum objective value is $\|L^{\leqslant D}\|$. Here the norm is with respect to the following inner product space, defined using the null distribution. Below, we follow the techniques introduced in Mao and Wein (2025) for bounding the advantage. In particular, our result can be viewed as a generalization of their result for broader classes of testing problems.

For functions $f, g : \mathbb{R}^{n\times(d+1)} \to \mathbb{R}$, we define the inner product $\langle f, g\rangle := \mathbb{E}_{z\sim\mathcal{Q}}[f(z)g(z)]$ and the associated norm $\|f\| = \sqrt{\langle f, f\rangle}$. We then have that

$$\mathrm{Adv}^{\leqslant D} = \max_{f\in\mathbb{R}[z]_{\leqslant D}} \frac{\langle f, L\rangle}{\|f\|} = \frac{\langle L^{\leqslant D}, L\rangle}{\|L^{\leqslant D}\|} = \|L^{\leqslant D}\|,$$

where $f^{\leqslant D}$ is the orthogonal projection (with respect to $\langle\cdot,\cdot\rangle$) of a function f onto $\mathbb{R}[z]_{\leqslant D}$ (the subspace of polynomials $\mathbb{R}^{n\times(d+1)} \to \mathbb{R}$ of degree at most D), and $L(z) := \frac{d\mathcal{P}}{d\mathcal{Q}}(z)$ is the likelihood ratio. Let $\mathcal{P}_u$ denote the distribution of $z \sim \mathcal{P}$ conditioned on a particular choice of u, and let $L_u(z) := \frac{d\mathcal{P}_u}{d\mathcal{Q}}(z)$ so that $L(z) = \mathbb{E}_{u\sim\mathrm{Unif}(\mathcal{S}^{d-1})} L_u(z)$. We have the squared advantage

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 = \|L^{\leqslant D}\|^2 = \langle L^{\leqslant D}, L^{\leqslant D}\rangle = \langle \mathop{\mathbb{E}}_{u} L_u^{\leqslant D}, \mathop{\mathbb{E}}_{u'} L_{u'}^{\leqslant D}\rangle = \mathop{\mathbb{E}}_{u,u'} \langle L_u^{\leqslant D}, L_{u'}^{\leqslant D}\rangle \tag{9}$$

where $u, u' \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ are independent. We fix u, u' and then focus on computing $\langle L_u^{\leqslant D}, L_{u'}^{\leqslant D}\rangle$. As in Mao and Wein (2025), we will work in the orthonormal basis where $u = e_1$ and $u' = \tau e_1 + \sqrt{1-\tau^2}e_2$ for $\tau := \langle u, u'\rangle$. Defining

$$H_\alpha(z) := \prod_{i=1}^{n}\left(h_{\alpha_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)\cdot\left(\prod_{j=1}^{d} h_{\alpha_{i,j}}(z_{i,j})\right)\right),$$

we expand the projected likelihood in the Hermite basis as follows.

$$\langle L_u^{\leqslant D}, L_{u'}^{\leqslant D}\rangle = \sum_{|\alpha|\leqslant D}\sum_{|\beta|\leqslant D}\langle c_{\alpha,u}H_\alpha(z), c_{\beta,u'}H_\beta(z)\rangle = \sum_{|\alpha|\leqslant D}\sum_{|\beta|\leqslant D} c_{\alpha,u}c_{\beta,u'}\langle H_\alpha(z), H_\beta(z)\rangle$$

for multi-indices α and β. We claim that $H_\alpha(.)$ is orthonormal with respect to $\mathcal{Q}$ (See Lemma 54 for a proof). As a consequence, we have

$$\langle L_u^{\leqslant D}, L_{u'}^{\leqslant D}\rangle = \sum_{|\alpha|\leqslant D} c_{\alpha,u}c_{\alpha,u'}.$$

We observe that for α such that $|\alpha| \leqslant D$ we have

$$c_{\alpha,u} := \langle L_u, H_\alpha\rangle = \mathop{\mathbb{E}}_{z\sim\mathcal{P}_u}[H_\alpha(z)].$$

Using this, we therefore have

$$\left(\mathrm{Adv}^{\leqslant D}\right)^2 = \mathop{\mathbb{E}}_{u,u'}\langle L_u^{\leqslant D}, L_{u'}^{\leqslant D}\rangle = \mathop{\mathbb{E}}_{u,u'}\left[\sum_{|\alpha|\leqslant D} c_{\alpha,u}c_{\alpha,u'}\right] = \sum_{|\alpha|\leqslant D}\mathop{\mathbb{E}}_{u,u'}\left[c_{\alpha,u}c_{\alpha,u'}\right].$$

Recalling the definition of H_α we have

$$c_{\alpha,u} = \mathop{\mathbb{E}}_{z\sim\mathcal{P}_u}\left[\prod_{i=1}^{n}\left(h_{\alpha_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)\cdot\left(\prod_{j=1}^{d} h_{\alpha_{i,j}}(z_{i,j})\right)\right)\right].$$

Since $u = e_1$, observe that the above term is only non-zero if the multi-indices are supported on the first and the last coordinate (respectively the first and last columns of α)[8] since the expectation under the standard Gaussian of $h_k(x)$ vanishes for $k \neq 0$. This allows for further simplification and we can express $c_{\alpha,u}$ in this more compact form.

$$c_{\alpha,u} = \prod_{i=1}^{n} \left(\mathop{\mathbb{E}}_{(x,y)\sim\nu} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \right] \right),$$

where $\mathcal{P}'_u$ is the two-dimensional distribution of $\mathcal{P}$ along u (and thus e_1 in this case) and the last coordinate. That is, this expression for α is supported only on the two relevant columns and for the remaining α the Hermite coefficient is zero.

We notice that as a direct consequence, even for $c_{\alpha,u'}$ we only need to calculate the coefficients where α is supported on the first and last columns. There could be more non-zero coefficients for u': however they do not contribute to the advantage since they are multiplied by zero in the above inner product expression. Thus, expressing u' in terms of its component along $u = e_1$ and e_2, we obtain

$$c_{\alpha,u'} = \prod_{i=1}^{n} \left(\mathop{\mathbb{E}}_{\substack{(x,y)\sim\nu \\ z\sim\mathcal{N}(0,1)}} \left[h_{\alpha_{i,1}}\left(\tau x + \sqrt{1-\tau^2}\cdot z\right) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \right] \right).$$

Now,

$$\begin{aligned} &\mathop{\mathbb{E}}_{\substack{(x,y)\sim\nu \\ z\sim\mathcal{N}(0,1)}} \left[h_{\alpha_{i,1}}\left(\tau x + \sqrt{1-\tau^2} z\right) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \right] \\ &= \mathbb{E}_{(x,y)\sim\nu} \left[h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)} \left[h_{\alpha_{i,1}}\left(\tau x + \sqrt{1-\tau^2} z\right) \mid x, y \right] \right] \end{aligned}$$

by the law of iterated expectations. Now, since $\tau \in [-1, 1]$ and because Hermite polynomials are eigenfunctions of the Gaussian Noise operator we obtain that

$$\begin{aligned} &\mathbb{E}_{(x,y)\sim\nu} \left[h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)} \left[h_{\alpha_{i,1}}\left(\tau x + \sqrt{1-\tau^2} z\right) \mid x, y \right] \right] \\ &= \mathbb{E}_{(x,y)\sim\nu} \left[h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \tau^{\alpha_{i,1}} h_{\alpha_{i,1}}(x) \right] \\ &= \tau^{\alpha_{i,1}} \cdot \mathbb{E}_{(x,y)\sim\mathcal{P}'_u} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \right]. \end{aligned}$$

Therefore we have

$$c_{\alpha,u'} = \prod_{i=1}^{n} \left(\tau^{\alpha_{i,1}} \cdot \mathop{\mathbb{E}}_{(x,y)\sim\nu} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \right] \right),$$

and thus

$$c_{\alpha,u} \cdot c_{\alpha,u'} = \left(\langle u, u' \rangle^{\sum_{i=1}^{n} \alpha_{i,1}} \right) \cdot \prod_{i=1}^{n} \mathop{\mathbb{E}}_{(x,y)\sim\nu} \left[h_{\alpha_{i,1}}(x) \cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right) \right]^2.$$

8. In the alternate distribution, the signal resides in this two-dimensional subspace. The remaining $d-1$ dimensional subspace is Gaussian noise.

Putting it together, we have that the squared advantage is:

$$\begin{aligned}
&\sum_{|\alpha|\leqslant D} \mathop{\mathbb{E}}_{u,u'}\left[c_{\alpha,u}c_{\alpha,u'}\right] \\
&= \sum_{|\alpha|\leqslant D} \mathop{\mathbb{E}}_{u,u'}\left[\langle u,u'\rangle^{\sum_{i=1}^n \alpha_{i,1}}\right] \cdot \left(\prod_{i=1}^n \mathop{\mathbb{E}}_{(x,y)\sim\nu}\left[h_{\alpha_{i,1}}(x)\cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right)\right]^2\right) \\
&= \sum_{p=0}^{D} \sum_{\substack{|\alpha|=p \\ \alpha \text{ supp. on } \{1,d+1\}}} \mathop{\mathbb{E}}_{u,u'}\left[\langle u,u'\rangle^{\sum_{i=1}^n \alpha_{i,1}}\right] \cdot \prod_{i=1}^n \left(\mathop{\mathbb{E}}_{(x,y)\sim\nu}\left[h_{\alpha_{i,1}}(x)\cdot h_{\alpha_{i,d+1}}\left(\frac{y}{\sqrt{s}}\right)\right]\right)^2.
\end{aligned}$$

Observing that $\mathcal{P}'_u$ is ν concludes our proof. □

D.3. Missing facts and proofs for low-degree lower bound lemmas

Fact 50 (Almost orthogonality of random vectors (Lemma 6.5 in Mao and Wein (2025))) *Let u, u' be independent random vectors drawn from $\mathcal{S}^{d-1}$. For $q \in \mathbb{N}$, if q is odd, then $\mathbb{E}_{u,u'}[\langle u,u'\rangle^q] = 0$ and if q is even*

$$\mathop{\mathbb{E}}_{u,u'}[\langle u,u'\rangle^q] \leqslant \left(\frac{q}{d}\right)^{q/2}$$

Ornstein-Uhlenbeck Operator. For some $\rho > 0$, the Ornstein-Uhlenbeck operator U_ρ maps some distribution F on $\mathbb{R}$ to the distribution of the random variable $\rho X + \sqrt{1-\rho^2}Z$, where $X \sim F$ and $Z \sim \mathcal{N}(0,1)$ is independent of X. A useful property of Ornstein–Uhlenbeck operator is that it operates diagonally with respect to Hermite polynomials.

Fact 51 (See e.g., O'Donnell (2014)) *For a Hermite polynomial h_k, and for any distribution F on $\mathbb{R}$, $\rho \in [-1,1]$, we have*

$$\mathop{\mathbb{E}}_{X\sim U_\rho F}[h_k(x)] = \rho^k \mathop{\mathbb{E}}_{X\sim F}[h_k(X)]$$

Lemma 52 (PSD-ness of the corruption covariance) *For $\varepsilon\kappa \geqslant 1-\varepsilon$, we have*

$$\begin{bmatrix} \frac{1}{\varepsilon}\cdot\left(1-\frac{1}{\kappa}(1-\varepsilon)\right) & -\frac{(1-\varepsilon)}{\varepsilon}\frac{\delta}{\kappa} \\ -\frac{(1-\varepsilon)}{\varepsilon}\frac{\delta}{\kappa} & \frac{\delta^2}{\kappa}+\sigma^2 \end{bmatrix} \succeq 0.$$

As a result Σ defined in Definition 39 is a valid covariance matrix.

Proof. For Σ to be a valid covariance matrix, it needs to be positive semi-definite. Our choice of Σ is I_{d-1} in the subspace orthogonal to v and the last coordinate. Therefore, it suffices for the following 2×2 sub-matrix to be PSD:

$$\begin{bmatrix} \frac{1}{\varepsilon}\cdot\left(1-\frac{1}{\kappa}(1-\varepsilon)\right) & -\frac{(1-\varepsilon)}{\varepsilon}\frac{\delta}{\kappa} \\ -\frac{(1-\varepsilon)}{\varepsilon}\frac{\delta}{\kappa} & \frac{\delta^2}{\kappa}+\sigma^2 \end{bmatrix}$$

For any symmetric 2×2 matrix to be PSD, it suffices that both diagonal elements and the determinant are non-negative. We note that $\kappa \geqslant 1$ and $\varepsilon > 0$ ensures non-negativity of the diagonal

elements. We will prove the result for the case $\sigma^2 = 0$ as $\sigma^2 > 0$ only makes it easier to satisfy the non-negativity of the determinant. For this case by a direct calculation we have that

$$\begin{aligned}
&\frac{1}{\varepsilon}\left(1 - \frac{1-\varepsilon}{\kappa}\right)\frac{\delta^2}{\kappa} \geqslant \left(\frac{1-\varepsilon}{\varepsilon}\right)^2 \frac{\delta^2}{\kappa^2} \iff \\
&\left(1 - \frac{1-\varepsilon}{\kappa}\right) \geqslant \frac{(1-\varepsilon)^2}{\varepsilon\kappa} \iff \\
&\left(1 - \frac{1-\varepsilon}{\kappa}\right) - \frac{(1-\varepsilon)^2}{\varepsilon\kappa} \geqslant 0 \iff \\
&\frac{1}{\varepsilon\kappa}\big(\varepsilon(\kappa - (1-\varepsilon)) - (1-\varepsilon)^2\big) \geqslant 0 \iff \\
&\varepsilon\kappa - \varepsilon + \varepsilon^2 - 1 - \varepsilon^2 + 2\varepsilon \geqslant 0 \iff \\
&\varepsilon\kappa \geqslant 1 - \varepsilon
\end{aligned}$$

□

Lemma 53 (Moment Matching) *For distributions $\mathcal{P}$ and $\mathcal{Q}$ as defined in Definition 39, the first three moments of $\mathcal{P}$ agree with $\mathcal{Q}$.*

Proof. For $y = \langle X, \beta\rangle + \eta$ where $X \sim \mathcal{N}(0, \Sigma), \eta \sim \mathcal{N}(0, \sigma^2)$ independent of X, the joint distribution $(X, y) \sim N(0, \Sigma_{Xy})$ for

$$\Sigma_{Xy} = \begin{bmatrix} \Sigma & \Sigma\beta \\ \beta^T\Sigma & \beta^T\Sigma\beta + \sigma^2 \end{bmatrix}.$$

Therefore, our distribution looks like,

$$\mathcal{P}(X, y) = (1-\varepsilon)\cdot\mathcal{N}\left(0, \begin{bmatrix} I_d - (1-1/\kappa)vv^T & \frac{\delta}{\kappa}v \\ \frac{\delta}{\kappa}v^T & \frac{\delta^2}{\kappa} + \sigma^2 \end{bmatrix}\right) + \varepsilon\cdot\mathcal{N}(0, \Sigma_2).$$

We pick Σ to match the first three moments with the null distribution $\mathcal{Q}$. Our first observation is that we require moment matching only in the subspace spanned by the last coordinate y and the direction v. Furthermore, since $E(X, y)$ is a centered Gaussian, the first and third moments already match with $\mathcal{Q}$. Therefore it suffices to match only the second moment and pick Σ_2 such that

$$(1-\varepsilon)\cdot\begin{bmatrix} I_d - (1-1/\kappa)vv^T & \frac{\delta}{\kappa}v \\ \frac{\delta}{\kappa}v^T & \frac{\delta^2}{\kappa} + \sigma^2 \end{bmatrix} + \varepsilon\cdot\Sigma = \begin{bmatrix} I_d & 0_d \\ 0_d^T & \frac{\delta^2}{\kappa} + \sigma^2 \end{bmatrix}$$

By direct calculation, we have

$$\Sigma = \begin{bmatrix} I_d + \frac{(1-\varepsilon)}{\varepsilon}\cdot(1-1/\kappa)vv^T & -\frac{(1-\varepsilon)}{\varepsilon}\frac{\delta}{\kappa}v \\ -\frac{(1-\varepsilon)}{\varepsilon}\frac{\delta}{\kappa}v^T & \frac{\delta^2}{\kappa} + \sigma^2 \end{bmatrix}$$

This is a valid covariance for $\varepsilon\kappa \geqslant 1 - \varepsilon$ as established in Lemma 52.

□

Lemma 54 (Orthonormality of the Hermite Basis) *$H_\alpha(z)$ is orthonormal with respect to $\mathcal{Q}$.*

Proof.

$$
\begin{aligned}
&\langle H_\alpha(z), H_\beta(z)\rangle \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{Q}}\left[\prod_{i=1}^{n}\left(h_{\alpha_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)\cdot\prod_{j=1}^{d}h_{\alpha_{i,j}}(z_{i,j})\right)\cdot\qquad\prod_{i=1}^{n}\left(h_{\beta_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)\cdot\prod_{j=1}^{d}h_{\beta_{i,j}}(z_{i,j})\right)\right] \\
&= \mathop{\mathbb{E}}_{z\sim\mathcal{Q}}\left[\prod_{i=1}^{n}\left(h_{\alpha_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)h_{\beta_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)\cdot\prod_{j=1}^{d}h_{\alpha_{i,j}}(z_{i,j})h_{\beta_{i,j}}(z_{i,j})\right)\right] \\
&= \mathop{\mathbb{E}}_{z_{i,d+1}\sim\mathcal{N}(0,s)}\left[\prod_{i=1}^{n}h_{\alpha_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)h_{\beta_{i,d+1}}\left(\frac{z_{i,d+1}}{\sqrt{s}}\right)\right]\cdot\qquad\mathop{\mathbb{E}}_{z_{i,j}\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^{n}\prod_{j=1}^{d}h_{\alpha_{i,j}}(z_{i,j})h_{\beta_{i,j}}(z_{i,j})\right] \\
&= \mathop{\mathbb{E}}_{z'_{i,d+1}\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^{n}h_{\alpha_{i,d+1}}(z'_{i,d+1})h_{\beta_{i,d+1}}(z'_{i,d+1})\right]\cdot\qquad\mathop{\mathbb{E}}_{z_{i,j}\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^{n}\prod_{j=1}^{d}h_{\alpha_{i,j}}(z_{i,j})h_{\beta_{i,j}}(z_{i,j})\right] \\
&= \mathbb{1}[\alpha=\beta]
\end{aligned}
$$

□

We now provide a proof for obtaining a closed form for the Hermite coefficients.

Lemma 45 *Let $(x,y)\sim D$ for $D=\mathcal{N}\left(0,\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}\right)$. Then, we have*

$$
\mathop{\mathbb{E}}_{(x,y)\sim D}\left[h_k(x)\cdot h_\ell\left(\frac{y}{\sigma_y}\right)\right]=\sqrt{\frac{k!}{\ell!}}\cdot\frac{1}{((k-\ell)/2)!\cdot 2^{(k-\ell)/2}}\cdot(\sigma_x^2-1)^{(k-\ell)/2}\cdot\left(\frac{\sigma_{xy}}{\sigma_y}\right)^{\ell}.
$$

for $k\geqslant\ell$ and $k-\ell$ even and 0 otherwise.

Proof. We note that a direct computation using Wick's probability theorem Wick (1950), also known as Isserlis' celebrated theorem for Gaussian random variables Isserlis (1918) is in principle possible, but the resulting calculations appear tedious. To circumvent this we will try to obtain the above closed form expression for the expectations as coefficients of certain terms in an infinite sum. We begin by utilizing the exponential generating function for Hermite polynomials Roman (1984) (See also the Wikipedia entry Wikipedia contributors (2025)). We have for every $t\in\mathbb{R}$ that

$$
e^{xt-t^2/2}=\sum_{k=0}^{\infty}h_k(x)\frac{t^k}{\sqrt{k!}}
$$

For $\tilde{y}:=y/\sigma_y$, we similarly have for every $u\in\mathbb{R}$ that

$$
e^{\tilde{y}u-u^2/2}=\sum_{\ell=0}^{\infty}h_\ell(\tilde{y})\frac{u^\ell}{\sqrt{\ell!}}
$$

Multiplying both equations we obtain

$$\begin{aligned}\exp\bigl(xt - t^2/2 + \tilde{y}u - u^2/2\bigr) &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\left(h_k(x)\frac{t^k}{\sqrt{k!}}\right)\cdot\left(h_\ell(\tilde{y})\frac{u^\ell}{\sqrt{\ell!}}\right)\\ &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}(h_k(x)h_\ell(\tilde{y}))\frac{t^k}{\sqrt{k!}}\frac{u^\ell}{\sqrt{\ell!}}\end{aligned}$$

We will take the expectation on both sides. Now computing the expectation on the LHS we have that

$$\begin{aligned}\mathbb{E}\bigl[\exp\bigl(xt - t^2/2 + \tilde{y}u - u^2/2\bigr)\bigr] &= \mathbb{E}[\exp(-t^2/2 - u^2/2)\cdot\exp(xt + \tilde{y}u)]\\ &= e^{-\frac{1}{2}\left(t^2+u^2\right)}\cdot\mathbb{E}\left[e^{(tx+u\tilde{y})}\right]\end{aligned}$$

Using the following fact about Gaussian distributions

$$\mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\Sigma)}\left[e^{\langle t,X\rangle}\right] = \exp\left(\frac{t^T\Sigma t}{2}\right)$$

allows us to compute a closed form for the expectation. Therefore by a direct computation we have the exponent being

$$[t,u]^T\begin{bmatrix}\sigma_x^2 & \sigma_{x\tilde{y}}\\ \sigma_{x\tilde{y}} & \sigma_{\tilde{y}}^2\end{bmatrix}\begin{bmatrix}t\\ u\end{bmatrix} = \sigma_x^2t^2 + 2tu\sigma_{x\tilde{y}} + u^2\sigma_{\tilde{y}}^2$$

Therefore we have that the LHS evaluates after taking expectations to the following.

$$\begin{aligned}\mathbb{E}\bigl[\exp\bigl(xt - t^2/2 + \tilde{y}u - u^2/2\bigr)\bigr] &= \exp\left(t^2\cdot\frac{\sigma_x^2-1}{2} + u^2\cdot\frac{\sigma_{\tilde{y}}^2-1}{2} + tu\sigma_{x\tilde{y}}\right)\\ &= \exp\left(t^2\cdot\frac{\sigma_x^2-1}{2} + tu\sigma_{x\tilde{y}}\right)\\ &= \exp\left(t^2\cdot\frac{\sigma_x^2-1}{2}\right)\exp(tu\sigma_{x\tilde{y}})\end{aligned}$$

We now observe that we need to read off the coefficient for t^ku^ℓ by expanding out the above sum to get the expectation of the product of the Hermite polynomials. By using the Taylor series expansion for e^x we have

$$\sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\mathbb{E}[h_k(x)h_\ell(\tilde{y})]\cdot\frac{t^k}{\sqrt{k!}}\frac{u^\ell}{\sqrt{\ell!}} = \left(\sum_{i=0}^{\infty}\frac{\left(\frac{t^2}{2}\right)^i\left(\sigma_x^2-1\right)^i}{i!}\right)\cdot\left(\sum_{j=0}^{\infty}\frac{(tu\sigma_{x\tilde{y}})^j}{j!}\right)$$

By equating terms on both sides, we require $j=\ell$. Since $k = 2i + j = 2i + \ell$, we require that $i = (k-\ell)/2$. Now since $i \geqslant 0$ we also need $k \geqslant \ell$ and $k-\ell$ even as well for the coefficients to be non-zero. Finally the result follows by picking the coefficient of t^ku^ℓ on the LHS. □

We now provide the proof of our bound on the combinatorial coefficients.

Lemma 47 *For a fixed choice of $p, m, \{d_i\}_{i\in S}$, we have*

$$\prod_{i\in S}(C_{k_i,\ell_i})^2 = \prod_{i\in S}\left(\sqrt{\frac{k_i!}{\ell_i!}}\cdot\frac{1}{((k_i-\ell_i)/2)!\cdot 2^{(k_i-\ell_i)/2}}\right)^2 \leqslant 2^p$$

Proof. We begin by defining

$$T_i \coloneqq \sqrt{\frac{k_i!}{\ell_i!}}\cdot\frac{1}{((k_i-\ell_i)/2)!\cdot 2^{(k_i-\ell_i)/2}}$$

Taking natural logarithm we have

$$\log T_i = \frac{1}{2}\log k_i! - \frac{1}{2}\log \ell_i! - \log((k_i-\ell_i)/2)! - (k_i-\ell_i)/2\cdot\log 2$$

Using Stirling's approximation for d_i, ℓ_i, k_i

$$\begin{aligned}\log T_i &\approx \frac{1}{2}(k_i\log k_i - k_i) - \frac{1}{2}(\ell_i\log\ell_i - \ell_i)\\ &\quad - \left(\frac{1}{2}(k_i-\ell_i)\log(k_i-\ell_i)/2 - (k_i-\ell_i)/2\right)\\ &\quad - (k_i-\ell_i)/2\cdot\log 2\\ &= \frac{1}{2}(k_i\log k_i - \ell_i\log\ell_i) - \frac{1}{2}(k_i-\ell_i)\log(k_i-\ell_i)\end{aligned}$$

We express the above equation in terms of k_i and set the derivative with respect to k_i to zero to find the maximum. Recall that since d_i is fixed, it suffices to maximize the terms corresponding to the samples individually. Since $\ell_i = d_i - k_i$ its derivative with respect to k_i is -1. Using this we have

$$\begin{aligned}&\log k_i + 1 - (\log\ell_i + 1)(-1) - (\log(k_i-\ell_i)+1)(1-(-1)) = 0 \iff\\ &\log k_i\ell_i + 2 - 2\log(k_i-\ell_i) - 2 = 0 \iff\\ &\log\sqrt{k_i\ell_i} = \log(k_i-\ell_i)\end{aligned}$$

This gives the following condition.

$$(k_i\ell_i) = (k_i-\ell_i)^2$$

Expanding the RHS and dividing both sides by ℓ_i^2 we obtain the following quadratic equation

$$\left(\frac{k_i}{\ell_i}\right)^2 - 3\frac{k_i}{\ell_i} + 1 = 0$$

This has solutions

$$\frac{k_i}{\ell_i} = \frac{3\pm\sqrt{5}}{2}$$

Since $k_i \geqslant \ell_i$, we pick the solution for $\frac{k_i}{\ell_i} = \frac{3+\sqrt{5}}{2}$. To show that this is a global maxima, we also consider the second derivative test by taking the derivative of

$$\log k_i\ell_i - 2\log(k_i-\ell_i) = \log k_i + \log\ell_i - 2\log(k_i-\ell_i)$$

Taking derivative with respect to k_i we have that

$$\frac{1}{k_i} - \frac{1}{\ell_i} - \frac{2}{k_i - \ell_i} \cdot 2 = \frac{1}{k_i} - \frac{1}{\ell_i} - \frac{4}{k_i - \ell_i}$$

which is always negative since $k_i \geqslant \ell_i$. Substituting back in the original expression we obtain that

$$\begin{aligned}
\log T_i &\approx \frac{1}{2}(k_i \log k_i - \ell_i \log \ell_i) - \left(\frac{1}{2}(k_i - \ell_i)\log(k_i - \ell_i)\right) \\
&= \frac{1}{2}(k_i \log k_i - \ell_i \log \ell_i) - \frac{(k_i - \ell_i)}{2} \log \sqrt{k_i \ell_i} \\
&= \frac{1}{2}(k_i \log k_i - \ell_i \log \ell_i) - \frac{(k_i - \ell_i)}{4} \log k_i - \frac{(k_i - \ell_i)}{4} \log \ell_i \\
&= \left(\frac{k_i + \ell_i}{4}\right) \log k_i - \left(\frac{\ell_i}{2} + \frac{k_i - \ell_i}{4}\right) \log \ell_i \\
&= \frac{d_i}{4} \log \frac{k_i}{\ell_i} = \frac{d_i}{4} \log \frac{3 + \sqrt{5}}{2} = \frac{d_i}{4} \log \left(\frac{1 + \sqrt{5}}{2}\right)^2 = \frac{d_i}{2} \log \frac{1 + \sqrt{5}}{2} \leqslant \frac{d_i}{2} \log 2
\end{aligned}$$

Therefore, after taking product over the m samples for a fixed set of d_i, we have that

$$\prod_{i \in S} T_i^2 \leqslant \prod_{i \in S} 2^{d_i} = 2^p$$

□

D.4. Reduction

In this section we will state our reduction. Our reduction is slightly involved as we have an *asymmetric* robust regression instance in the null distribution and the alternative distribution. This is because the label noise in both cases are different, and in the null distribution it could be arbitrarily large. We now define the following more specific problem which we will utilize for our reduction.

Problem 57 (Robust Linear Testing with a Fixed Adversary) *Given corruption rate $\varepsilon \in (0, 1/2)$, $\kappa \geqslant 1$, signal strength $\alpha \in \mathbb{R}_+$, sample size $n \in \mathbb{N}$, dimension $d \in \mathbb{N}$, define $\sigma_y^2 \coloneqq \alpha^2 + 1, \delta \coloneqq \alpha\sqrt{\kappa}$ and consider the following hypothesis testing problem with input samples $\{z_i\}_{i=1}^n \in \mathbb{R}^{n \times (d+1)}$.*

1. *H_0: Null $\mathcal{Q}$: Let $X \sim \mathcal{N}(0, I_d)$ and $Y \sim \mathcal{N}(0, \sigma_y^2)$ be independent. Define $Z \coloneqq (X, Y) \in \mathbb{R}^{d+1}$. Then*

$$z_1, \dots, z_n \overset{\text{iid}}{\sim} \mathcal{Q} \overset{\text{def}}{=} \mathcal{N}\big(0, \operatorname{diag}\big(I_d, \sigma_y^2\big)\big).$$

2. *H_1: Alternative $\mathcal{P}$: $v \sim \operatorname{Unif}(\mathcal{S}^{d-1})$. Conditioned on v,*

$$z_1, \dots, z_n \overset{\text{iid}}{\sim} \mathcal{P} \overset{\text{def}}{=} (1 - \varepsilon) \cdot \mathcal{N}(0, \Sigma_1) + \varepsilon \cdot \mathcal{N}(0, \Sigma_2)$$

where

$$\Sigma_1 = \begin{bmatrix} I_d - (1 - 1/\kappa) v v^T & \frac{\delta}{\kappa} v \\ \frac{\delta}{\kappa} v^T & \frac{\delta^2}{\kappa} + 1 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} I_d + \frac{(1-\varepsilon)}{\varepsilon} \cdot (1 - 1/\kappa) vv^T & -\frac{(1-\varepsilon)}{\varepsilon} \frac{\delta}{\kappa} v \\ -\frac{(1-\varepsilon)}{\varepsilon} \frac{\delta}{\kappa} v^T & \frac{\delta^2}{\kappa} + 1 \end{bmatrix}.$$

We observe that the $\mathcal{N}(0, \Sigma_1)$ above describes an uncorrupted linear regression model with $\beta = \delta v$, $\Sigma = I - (1 - 1/\kappa)vv^T$ and $\sigma^2 = 1$.

Conjecture 58 *For any $\alpha > 0$, $\varepsilon\kappa \geqslant C$ for some constant $C > 0$ sufficiently large, Problem 57 is computationally hard for efficient algorithms unless $n = \tilde{\Omega}(\min\{d\varepsilon^2\kappa^2, \varepsilon^2 d^2\})$.*

Conjecture 58 is supported by our main low-degree lower bound in Theorem 41. In the above instance we set the label noise variance to 1 for simplicity. We now provide a reduction from Problem 57 to robust regression as follows.

D.4.1. THE REGRESSION ALGORITHM

The estimation algorithm $\mathcal{A}$ takes as input (i) ε-corrupted samples from the linear model $y = \langle X, \beta \rangle + \eta$ for $X \sim \mathcal{N}(0, \Sigma)$, $\eta \sim \mathcal{N}(0, \sigma^2)$ is independent noise and $\varepsilon \in (0, 1/2)$ and (ii) a parameter α such that whenever $\|\Sigma^{1/2}\beta\| = \alpha$ and $0 < \sigma^2 \leqslant 1$, the algorithm outputs an estimator $\widehat{\beta}$ satisfying

$$\left\|\Sigma^{1/2}\left(\widehat{\beta} - \beta\right)\right\| \leqslant 0.1\alpha$$

with probability $1 - o(1)$, using n samples and running time T.

We note here that the choice of $0 < \sigma^2 \leqslant 1$ is for simplicity and one can indeed let the error of the regression algorithm scale with σ more explicitly. We will assume $\alpha = \Omega(1)$ for our reduction. Information-theoretically, one requires $\alpha = \Omega(\sigma\varepsilon)$, which in our case reduces to $\Omega(\varepsilon)$. In this regard, our reduction requires (slightly) stronger assumption.

As a first step, we will prove that an estimator achieving the above error guarantee has good correlation with the underlying unknown vector in our hard instance.

Lemma 59 (Small prediction error implies correlation) *When $\mathcal{A}$ is run on samples from the alternative distribution $\mathcal{P}$ in Problem 57, using the first d coordinates as features and the last coordinate as the label, the estimator $\widehat{\beta}$ satisfies*

$$|\langle \widehat{\beta}, v \rangle| \geqslant 0.9\delta \qquad \text{and} \qquad \|\widehat{\beta}\| \leqslant 1.1\delta$$

with probability $1 - o(1)$.

Proof. Under the alternative, the regression vector is $\beta = \delta v$, where v is a unit vector and

$$\delta = \alpha\sqrt{\kappa}.$$

Moreover, since

$$\Sigma = I_d - (1 - 1/\kappa)vv^\top,$$

we have

$$\Sigma v = \frac{1}{\kappa} v \qquad \text{and hence} \qquad \Sigma^{1/2} v = \frac{1}{\sqrt{\kappa}} v.$$

By the guarantee of $\mathcal{A}$, with probability $1 - o(1)$,

$$\|\Sigma^{1/2}(\widehat{\beta} - \beta)\| \leqslant 0.1\alpha.$$

On this event,

$$\begin{aligned}\left|\left\langle \widehat{\beta} - \beta,\, v \right\rangle\right| &= \left|\left\langle \Sigma^{1/2}(\widehat{\beta} - \beta),\, \Sigma^{-1/2} v \right\rangle\right| \\ &\leqslant \|\Sigma^{1/2}(\widehat{\beta} - \beta)\| \cdot \|\Sigma^{-1/2} v\| \\ &= \|\Sigma^{1/2}(\widehat{\beta} - \beta)\| \cdot \sqrt{\kappa} \\ &\leqslant 0.1\alpha\sqrt{\kappa} = 0.1\delta.\end{aligned}$$

Using $\beta = \delta v$, it follows that

$$\begin{aligned}|\langle \widehat{\beta}, v\rangle| &= |\langle \beta, v\rangle + \langle \widehat{\beta} - \beta, v\rangle| \\ &\geqslant |\langle \beta, v\rangle| - |\langle \widehat{\beta} - \beta, v\rangle| \\ &\geqslant \delta - 0.1\delta = 0.9\delta.\end{aligned}$$

Now, for the norm bound, we similarly have

$$\begin{aligned}\|\widehat{\beta}\| &\leqslant \|\widehat{\beta} - \beta\| + \|\beta\| \\ &\leqslant \|\Sigma^{-1/2}\|\, \|\Sigma^{1/2}(\widehat{\beta} - \beta)\| + \delta.\end{aligned}$$

Since the smallest eigenvalue of Σ is $1/\kappa$, we have $\|\Sigma^{-1/2}\| = \sqrt{\kappa}$. Therefore,

$$\begin{aligned}\|\widehat{\beta}\| &\leqslant \sqrt{\kappa} \cdot 0.1\alpha + \delta \\ &= 0.1\delta + \delta \\ &= 1.1\delta.\end{aligned}$$

This proves the claim. □

Therefore, under the alternative distribution, the algorithm $\mathcal{A}$ outputs an estimator that is well correlated with the hidden direction v. We now use this observation to formalize the reduction. The ideas underlying this reduction are standard and closely related to those used in the literature; see, for example, Brennan et al. (2018); Brennan and Bresler (2020).

Lemma 60 (Testing using an estimation algorithm) *There is an algorithm $\mathcal{B}$ that*

1. *takes $2n$ samples from Problem 57,*
2. *runs in time $T + \mathrm{poly}(n, d)$, and*
3. *distinguishes the null and alternative in Problem 57 with probability $1 - o(1)$.*

Before proving the lemma, we note that the testing problem itself requires $n = \Omega(d)$ samples information-theoretically. In the regime $\varepsilon \gg 1/\sqrt{d}$, the sample size required by the tester constructed below will be much smaller than d, so this information-theoretic lower bound does not obstruct the reduction.

Proof. The algorithm $\mathcal{B}$ proceeds as follows. Given samples

$$z_1, z_2, \dots, z_{2n}$$

from Problem 57, split them into two halves. Using the first half, namely $z_1, \dots, z_n$, run the regression algorithm $\mathcal{A}$ with the first d coordinates of each sample as the feature vector and the last coordinate as the response. Let $\widehat{\beta}$ denote the output. Define

$$\widehat{v} := \begin{cases} \widehat{\beta}/\|\widehat{\beta}\|, & \text{if } \widehat{\beta} \neq 0, \\ \hat{e}, & \text{if } \widehat{\beta} = 0, \end{cases}$$

where $\hat{e}$ is any fixed unit vector. By construction, $\widehat{v}$ depends only on the first half of the samples, and is therefore independent of the second half. We then use the remaining samples $z_{n+1}, \dots, z_{2n}$ to compute a test statistic, with the choice depending on the value of κ. Writing

$$X_{i+n} := z_{i+n,1:d} \qquad \text{and} \qquad y_{i+n} := z_{i+n,d+1},$$

define

$$f(z) := \begin{cases} \dfrac{1}{\alpha^2\sigma_y^2} \displaystyle\sum_{i=1}^{n} \langle X_{i+n}, \widehat{v}\rangle^2 \big(y_{i+n}^2 - \sigma_y^2\big), & \text{if } \kappa \leqslant \sqrt{d}, \\ \displaystyle\sum_{i=1}^{n} \Big(\langle X_{i+n}, \widehat{v}\rangle^4 - 3\Big), & \text{if } \kappa > \sqrt{d}. \end{cases}$$

The sample split ensures that the randomness in $\widehat{v}$ is independent of the samples used to compute the test statistic. We analyze the two regimes separately.

Small Condition Number

In this regime we assume $\kappa \leqslant \sqrt{d}$ and consider the statistic

$$f(z) = \frac{1}{\alpha^2\sigma_y^2} \sum_{i=1}^{n} \langle X_{i+n}, \widehat{v}\rangle^2 \big(y_{i+n}^2 - \sigma_y^2\big).$$

Since $\widehat{v}$ is computed from the first half of the samples, it is independent of the second half. Thus, throughout the argument below, we will condition on $\widehat{v}$.

NULL DISTRIBUTION

Under the null, conditioned on $\widehat{v}$, we have

$$X_{i+n} \sim \mathcal{N}(0, I_d), \qquad y_{i+n} \sim \mathcal{N}(0, \sigma_y^2),$$

and these are independent. Therefore,

$$\begin{aligned} \mathbb{E}[f(z) \mid \widehat{v}] &= \frac{1}{\alpha^2\sigma_y^2} \sum_{i=1}^{n} \mathbb{E}\big[\langle X_{i+n}, \widehat{v}\rangle^2 (y_{i+n}^2 - \sigma_y^2) \mid \widehat{v}\big] \\ &= \frac{1}{\alpha^2\sigma_y^2} \sum_{i=1}^{n} \mathbb{E}\big[\langle X_{i+n}, \widehat{v}\rangle^2 \mid \widehat{v}\big]\, \mathbb{E}\big[y_{i+n}^2 - \sigma_y^2\big] \end{aligned}$$

$$= 0.$$

For the variance, the summands are independent and centered, so

$$\mathbb{V}(f(z) \mid \widehat{v}) = n\,\mathbb{V}\left(\frac{\langle X, \widehat{v}\rangle^2 (y^2 - \sigma_y^2)}{\alpha^2 \sigma_y^2} \;\middle|\; \widehat{v}\right),$$

where $X \sim N(0, I_d)$ and $y \sim N(0, \sigma_y^2)$ are independent. Hence

$$\begin{aligned}\mathbb{V}(f(z) \mid \widehat{v}) &= \frac{n}{\alpha^4 \sigma_y^4}\,\mathbb{E}[\langle X, \widehat{v}\rangle^4 \mid \widehat{v}]\;\mathbb{E}[(y^2 - \sigma_y^2)^2] \\ &= \frac{n}{\alpha^4 \sigma_y^4} \cdot 3 \cdot 2\sigma_y^4 \\ &= \frac{6n}{\alpha^4}.\end{aligned}$$

Thus,

$$\mathbb{V}(f(z) \mid \widehat{v}) = O\Big(\frac{n}{\alpha^4}\Big).$$

ALTERNATIVE DISTRIBUTION

We first compute $\mathbb{E}[XX^\top y^2]$ under the alternative.

Fact 61 *If (X, y) is jointly Gaussian with mean zero and block covariance*

$$\begin{pmatrix} \Sigma & c \\ c^\top & \sigma_y^2 \end{pmatrix},$$

then

$$\mathbb{E}[XX^\top y^2] = \sigma_y^2 \Sigma + 2cc^\top.$$

Under the alternative, the first mixture component has

$$\Sigma_{x,1} = I_d - (1 - 1/\kappa) vv^\top, \qquad c_1 = \frac{\delta}{\kappa} v,$$

while the second has

$$\Sigma_{x,2} = I_d + \frac{1-\varepsilon}{\varepsilon}(1 - 1/\kappa) vv^\top, \qquad c_2 = -\frac{1-\varepsilon}{\varepsilon}\frac{\delta}{\kappa} v.$$

Applying Fact 61 to each component and averaging over the mixture gives

$$\mathbb{E}[XX^\top y^2] = (1-\varepsilon)\big(\sigma_y^2 \Sigma_{x,1} + 2c_1 c_1^\top\big) + \varepsilon\big(\sigma_y^2 \Sigma_{x,2} + 2c_2 c_2^\top\big).$$

Since

$$(1-\varepsilon)\Sigma_{x,1} + \varepsilon \Sigma_{x,2} = I_d,$$

the covariance contribution is $\sigma_y^2 I_d$. Also,

$$(1-\varepsilon) c_1 c_1^\top + \varepsilon c_2 c_2^\top = (1-\varepsilon)\frac{\delta^2}{\kappa^2} vv^\top + \varepsilon\left(\frac{1-\varepsilon}{\varepsilon}\right)^2 \frac{\delta^2}{\kappa^2} vv^\top$$

$$= \frac{1-\varepsilon}{\varepsilon}\frac{\delta^2}{\kappa^2} vv^\top$$
$$= \frac{1-\varepsilon}{\varepsilon}\frac{\alpha^2}{\kappa} vv^\top.$$

Therefore,

$$\mathbb{E}[XX^\top y^2] = \sigma_y^2 I_d + 2\frac{1-\varepsilon}{\varepsilon}\frac{\alpha^2}{\kappa} vv^\top.$$

Using also $\mathbb{E}[XX^\top] = I_d$, we obtain

$$\begin{aligned}\mathbb{E}[f(z) \mid \widehat{v}] &= \frac{n}{\alpha^2\sigma_y^2}\,\mathbb{E}\big[\langle X, \widehat{v}\rangle^2 (y^2 - \sigma_y^2) \mid \widehat{v}\big] \\ &= \frac{n}{\alpha^2\sigma_y^2}\widehat{v}^\top\big(\mathbb{E}[XX^\top y^2] - \sigma_y^2 I_d\big)\widehat{v} \\ &= \frac{2n(1-\varepsilon)}{\varepsilon\kappa\sigma_y^2}\langle v, \widehat{v}\rangle^2.\end{aligned}$$

By Lemma 59,

$$|\langle \widehat{\beta}, v\rangle| \geqslant 0.9\delta \qquad \text{and} \qquad \|\widehat{\beta}\| \leqslant 1.1\delta,$$

and hence

$$\langle v, \widehat{v}\rangle^2 = \frac{\langle \widehat{\beta}, v\rangle^2}{\|\widehat{\beta}\|^2} \geqslant \left(\frac{0.9}{1.1}\right)^2 > \frac{1}{4}.$$

Since $\varepsilon < 1/2$, we also have $1 - \varepsilon \geqslant 1/2$. Therefore,

$$\mathbb{E}[f(z) \mid \widehat{v}] \geqslant \frac{n}{4\varepsilon\kappa\sigma_y^2} = \frac{n}{4\varepsilon\kappa(\alpha^2+1)}.$$

We now bound the variance. Writing

$$W := \frac{\langle X, \widehat{v}\rangle^2 (y^2 - \sigma_y^2)}{\alpha^2\sigma_y^2},$$

we have

$$\mathbb{V}(f(z) \mid \widehat{v}) = n\,\mathbb{V}(W \mid \widehat{v}) \leqslant n\,\mathbb{E}[W^2 \mid \widehat{v}].$$

Thus

$$\mathbb{V}(f(z) \mid \widehat{v}) \leqslant \frac{n}{\alpha^4\sigma_y^4}\,\mathbb{E}\big[\langle X, \widehat{v}\rangle^4 (y^2 - \sigma_y^2)^2 \mid \widehat{v}\big].$$

By Cauchy–Schwarz,

$$\mathbb{E}\big[\langle X, \widehat{v}\rangle^4 (y^2 - \sigma_y^2)^2 \mid \widehat{v}\big] \leqslant \sqrt{\mathbb{E}[\langle X, \widehat{v}\rangle^8 \mid \widehat{v}]}\,\sqrt{\mathbb{E}[(y^2 - \sigma_y^2)^4]}.$$

Since $y \sim N(0, \sigma_y^2)$ in each mixture component,

$$\mathbb{E}[(y^2 - \sigma_y^2)^4] = O(\sigma_y^8).$$

Moreover, under the first mixture component,

$$\langle X, \widehat{v}\rangle \sim N(0, s_1^2), \qquad s_1^2 = 1 - (1 - 1/\kappa)\langle \widehat{v}, v\rangle^2 \leqslant 1,$$

while under the second,

$$\langle X, \widehat{v}\rangle \sim N(0, s_2^2), \qquad s_2^2 = 1 + \frac{1-\varepsilon}{\varepsilon}(1-1/\kappa)\langle \widehat{v}, v\rangle^2 \leqslant \frac{1}{\varepsilon}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\langle X, \widehat{v}\rangle^8 \mid \widehat{v}] &= 105\big((1-\varepsilon)s_1^8 + \varepsilon s_2^8\big)\\ &= O(\varepsilon^{-3}).\end{aligned}$$

Combining the above bounds yields

$$\mathbb{V}(f(z) \mid \widehat{v}) \leqslant O\Big(\frac{n}{\alpha^4\varepsilon^{3/2}}\Big).$$

We now define the test to accept the alternative iff

$$f(z) \geqslant T, \qquad T := \frac{n}{8\varepsilon\kappa(\alpha^2+1)}.$$

Under the null, Chebyshev's inequality gives

$$\mathop{\mathbb{P}}_{H_0}\big(|f(z)| > T \mid \widehat{v}\big) \leqslant \frac{\mathbb{V}(f(z) \mid \widehat{v})}{T^2} = O\bigg(\frac{\varepsilon^2\kappa^2(\alpha^2+1)^2}{n\alpha^4}\bigg).$$

Under the alternative, since $\mathbb{E}[f(z) \mid \widehat{v}] \geqslant 2T$, another application of Chebyshev gives

$$\mathop{\mathbb{P}}_{H_1}\big(f(z) \leqslant T \mid \widehat{v}\big) \leqslant \frac{\mathbb{V}(f(z) \mid \widehat{v})}{T^2} = O\bigg(\frac{\varepsilon^{1/2}\kappa^2(\alpha^2+1)^2}{n\alpha^4}\bigg).$$

Since $\alpha = \Omega(1)$, we have

$$\frac{(\alpha^2+1)^2}{\alpha^4} = O(1),$$

and hence

$$\mathop{\mathbb{P}}_{H_0}\big(|f(z)| > T \mid \widehat{v}\big) = O\bigg(\frac{\varepsilon^2\kappa^2}{n}\bigg), \qquad \mathop{\mathbb{P}}_{H_1}\big(f(z) \leqslant T \mid \widehat{v}\big) = O\bigg(\frac{\varepsilon^{1/2}\kappa^2}{n}\bigg).$$

Therefore, it suffices to take

$$n \gg \varepsilon^{1/2}\kappa^2$$

for this to yield a valid distinguisher. Since $n \gtrsim d$ samples is required for the regression algorithm even information-theoretically, and $\varepsilon^{1/2}\kappa^2 \lesssim d$, we are done.

Large condition number

We now consider the regime $\kappa > \sqrt{d}$, in which the test statistic is

$$f(z) = \sum_{i=1}^{n}\Big(\langle X_{i+n}, \widehat{v}\rangle^4 - 3\Big).$$

As before, we condition on $\widehat{v}$, which is independent of the second half of the samples.

NULL DISTRIBUTION

Under the null, conditioned on $\widehat{v}$, each

$$\langle X_{i+n}, \widehat{v}\rangle \sim N(0,1).$$

Since the fourth moment of a standard Gaussian is 3, we have

$$\mathbb{E}[f(z) \mid \widehat{v}] = 0.$$

Moreover, the summands are independent, so

$$\begin{aligned}\mathbb{V}(f(z) \mid \widehat{v}) &= n\ \mathbb{V}\big(\langle X, \widehat{v}\rangle^4 - 3 \mid \widehat{v}\big) \\ &\leqslant n\ \mathbb{E}\big[(\langle X, \widehat{v}\rangle^4 - 3)^2 \mid \widehat{v}\big] \\ &= n\,(\mathbb{E}[U^8] - 9),\end{aligned}$$

where $U \sim N(0,1)$. Since $\mathbb{E}[U^8] = 105$, it follows that

$$\mathbb{V}(f(z) \mid \widehat{v}) = 96n = O(n).$$

ALTERNATIVE DISTRIBUTION

Let

$$t^2 := \langle \widehat{v}, v\rangle^2.$$

Under the first mixture component,

$$\langle X, \widehat{v}\rangle \sim N(0, s_1^2), \qquad s_1^2 = 1 - a, \qquad a := (1 - 1/\kappa)t^2,$$

while under the second,

$$\langle X, \widehat{v}\rangle \sim N(0, s_2^2), \qquad s_2^2 = 1 + b, \qquad b := \frac{1-\varepsilon}{\varepsilon}(1 - 1/\kappa)t^2.$$

Therefore,

$$\mathbb{E}[\langle X, \widehat{v}\rangle^4 - 3 \mid \widehat{v}] = 3\big((1-\varepsilon)(1-a)^2 + \varepsilon(1+b)^2 - 1\big).$$

Expanding the right-hand side and using

$$-(1-\varepsilon)a + \varepsilon b = 0,$$

we obtain

$$\begin{aligned}(1-\varepsilon)(1-a)^2 + \varepsilon(1+b)^2 - 1 &= (1-\varepsilon)a^2 + \varepsilon b^2 \\ &= \frac{1-\varepsilon}{\varepsilon}a^2.\end{aligned}$$

Hence,

$$\mathbb{E}[\langle X, \widehat{v}\rangle^4 - 3 \mid \widehat{v}] = 3\frac{1-\varepsilon}{\varepsilon}(1 - 1/\kappa)^2 t^4.$$

By Lemma 59 once again we have,

$$|\langle\widehat{\beta}, v\rangle| \geqslant 0.9\delta \qquad \text{and} \qquad \|\widehat{\beta}\| \leqslant 1.1\delta,$$

and therefore

$$t^2 = \langle\widehat{v}, v\rangle^2 \geqslant \left(\frac{0.9}{1.1}\right)^2 > \frac{1}{4}, \qquad \text{so} \qquad t^4 \geqslant \frac{1}{16}.$$

Since $\kappa > \sqrt{d}$, we have $(1 - 1/\kappa)^2 = \Theta(1)$ in the regime of interest, and thus

$$\mathbb{E}[\langle X, \widehat{v}\rangle^4 - 3 \mid \widehat{v}] = \Omega(1/\varepsilon).$$

Summing over the n independent samples yields

$$\mathbb{E}[f(z) \mid \widehat{v}] = \Omega(n/\varepsilon).$$

For the variance, we use the same eighth-moment bound as in the small-condition-number case:

$$\mathbb{E}[\langle X, \widehat{v}\rangle^8 \mid \widehat{v}] = O(\varepsilon^{-3}).$$

Therefore,

$$\begin{aligned}\mathbb{V}(f(z) \mid \widehat{v}) &= n\ \mathbb{V}\big(\langle X, \widehat{v}\rangle^4 - 3 \mid \widehat{v}\big) \\ &\leqslant n\ \mathbb{E}[\langle X, \widehat{v}\rangle^8 \mid \widehat{v}] \\ &= O(n/\varepsilon^3).\end{aligned}$$

We now choose a threshold

$$T := c\,\frac{n}{\varepsilon}$$

for a sufficiently small absolute constant $c > 0$. Under the null, Chebyshev's inequality gives

$$\mathop{\mathbb{P}}_{H_0}\big(|f(z)| > T \mid \widehat{v}\big) \leqslant \frac{\mathbb{V}(f(z) \mid \widehat{v})}{T^2} = O\left(\frac{\varepsilon^2}{n}\right).$$

Under the alternative, since $\mathbb{E}[f(z) \mid \widehat{v}] \geqslant 2T$ for a suitable choice of c, another application of Chebyshev's inequality gives

$$\mathop{\mathbb{P}}_{H_1}\big(f(z) \leqslant T \mid \widehat{v}\big) \leqslant \frac{\mathbb{V}(f(z) \mid \widehat{v})}{T^2} = O\left(\frac{1}{\varepsilon n}\right).$$

Hence it suffices to take

$$n \gg \frac{1}{\varepsilon}$$

for this to yield a valid distinguisher and this is much fewer than what we require even information-theoretically for reasonable parameter choices of $\varepsilon \gg \frac{1}{d}$ in the high-dimensional setting. Combining this with the analysis in the regime $\kappa \leqslant \sqrt{d}$ completes the proof. □

Corollary 62 (Hardness of estimation) *Assuming Conjecture 58, for signal strength $\alpha = \Omega(1)$ it is computationally hard for efficient algorithms to output an estimator $\widehat{\beta}$ satisfying*

$$\left\|\Sigma^{1/2}(\widehat{\beta} - \beta)\right\| \leqslant 0.1\alpha$$

using

$$n = o\big(\min\{d\varepsilon^2\kappa^2,\ \varepsilon^2 d^2\}\big)$$

samples, up to $\mathrm{poly}(\log d)$ *factors.*

In particular, this suggests that even achieving constant-factor relative error in prediction norm may be computationally hard below the above sample complexity.

Appendix E. Consequences for Private Regression

We now state the consequences of Corollary 62 for differentially private regression using the relationship between privacy and robustness. This relationship has been investigated intensively, see for e.g., Dwork and Lei (2009); Georgiev and Hopkins (2022). Throughout this section we closely follow Diakonikolas et al. (2025b)'s analogous argument for the related covariance-aware mean estimation problem.

Let $n_{\text{private, eff}}(\alpha, \gamma, \varepsilon, \delta)$ denote the sample complexity of the best efficient (ε, δ)-DP algorithm that when the input is a set of n i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from the linear model $Y = \langle X, \beta\rangle + \zeta$ for $X \sim \mathcal{N}(0, \Sigma)$ for unknown Σ and $\zeta \sim \mathcal{N}(0, 1)$ independent of X outputs with probability $1 - \gamma$ an estimate $\hat{\beta}$ satisfying $\|\Sigma^{1/2}(\hat{\beta} - \beta)\| \leqslant \alpha$.

Let $n_{\text{robust, eff}}(\alpha, \gamma, \eta)$ denote the sample complexity of the best efficient robust algorithm that takes as input η corrupted set of n samples $(X_i, Y_i)_{i=1}^n$ from the linear model $Y = \langle X, \beta\rangle + \zeta$ for $X \sim \mathcal{N}(0, \Sigma)$ for unknown Σ and $\zeta \sim \mathcal{N}(0, 1)$ independent of X and with probability $1 - \gamma$ outputs $\hat{\beta}$ such that $\|\Sigma^{1/2}(\hat{\beta} - \beta)\| \leqslant \alpha$.

Conjecture 63 *For any $\alpha \gtrsim \eta \log 1/\eta$ and $\eta\kappa \geqslant C$ for a sufficiently large constant $C > 0$,*

$$n_{\text{robust, eff}}(\alpha, \gamma, \eta) \gg \min\big\{d\eta^2\kappa^2, \eta^2 d^2\big\}$$

where κ is the condition number of Σ.

Corollary 62 provides evidence towards Conjecture 63.

Proposition 64 *Under Conjecture 63, for any $\alpha \gtrsim \eta \log 1/\eta$ and $\kappa \geqslant \sqrt{d}$ we have*

$$n_{\text{private, eff}}(\alpha, \gamma, \varepsilon, \delta) \gg \max_{t \in (0,1/2)} \min\left(d^{2-2t}, d^t \cdot \frac{\log(1/\gamma)}{\varepsilon}, d^t \cdot \frac{\log(1/\delta)}{\varepsilon}\right)$$

We now provide context behind Proposition 64 and then conclude this section with a formal proof. Let $n_{\text{private}}(\alpha, \gamma, \varepsilon, \delta)$ be the *information-theoretic* sample complexity for Private Regression using an (ε, δ)-DP algorithm to obtain error α with probability $1 - \gamma$. Specializing to the setting $\alpha = 1$ and $\varepsilon = 1$, we have from (Liu et al., 2022, Corollary 4.16) that

$$n_{\text{private}}(1, \gamma, 1, \delta) \lesssim d + \log(1/\delta) + \log(1/\gamma)$$

Proposition 64 suggests that achieving the above sample complexity might be difficult for efficient algorithms since for $\kappa \geqslant \sqrt{d}$ we have that

$$n_{\text{private, eff}}(1,\gamma,1,\delta) \gg \max_{t\in(0,1/2)} \min\bigl(d^{2-2t}, d^t\cdot\log(1/\gamma), d^t\cdot\log(1/\delta)\bigr)$$

The current state-of-the-art efficient algorithm for this problem due to Anderson et al. (2025) uses

$$n \lesssim d^2 + d + \log(1/\delta) + \log(1/\gamma)$$

samples. While this algorithm matches the information-theoretic dependency on the privacy parameters, it uses $\Omega(d^2)$ samples. In contrast, the best known algorithm for this problem in the sub-quadratic sample regime due to Brown et al. (2024) uses

$$n \lesssim d + d\sqrt{\log(1/\delta)} + d(\log(1/\delta))^2$$

samples[9]. Proposition 64 gives evidence (in the regime $\kappa \geqslant \sqrt{d}$) that algorithms using $d^{2-\Omega(1)}$ samples need to incur a polynomial factor of d in front of $\log(1/\gamma)$ or $\log(1/\delta)$. In other words, such algorithms cannot decouple the dimension d and the logarithmic terms.

Proof. From (Georgiev and Hopkins, 2022, Theorem 3.1), we have that private estimators that succeed with *high probability* are also robust to a large fraction of corruptions. More precisely, suppose an estimation algorithm $\mathcal{A}$ is such that

1. $\mathcal{A}$ is (ε,δ)-DP and
2. whenever the input is a set of n iid samples $(X_i,y_i)_{i=1}^n$ from the linear model $y=\langle X,\beta\rangle+\zeta$ for $X\sim\mathcal{N}(0,\Sigma)$ for unknown Σ and $\zeta\sim\mathcal{N}(0,1)$ independent of X, $\mathcal{A}$ outputs with probability $1-\gamma$ an estimate $\hat\beta$ satisfying $\|\Sigma^{1/2}(\hat\beta-\beta)\|\leqslant\alpha$.

Then, $\mathcal{A}$ also has the following robustness guarantee: Given as input an η-corrupted set of samples from the above linear model where

$$\eta = \Theta\left(\min\left(\frac{\log(1/\gamma)}{\varepsilon n}, \frac{\log(1/\delta)}{\varepsilon n+\log n}\right)\right)$$

$\mathcal{A}$ outputs $\hat\beta$ such that $\|\Sigma^{1/2}(\hat\beta-\beta)\|_2\leqslant\alpha$ with probability $1-\gamma^{\Omega(1)}$. We will now provide a proof by contradiction. Suppose $\mathcal{A}$ is an efficient (ε,δ)-DP algorithm that uses n_s samples which is fewer than the claimed samples for a certain t. Then we have

$$n_s \leqslant d^t\cdot\frac{\log(1/\gamma)}{\varepsilon} \text{ and } n_s\leqslant d^t\cdot\frac{\log(1/\delta)}{\varepsilon}.$$

This implies that $\mathcal{A}$ is also robust towards η_s-corruption for

$$\eta_s = \Theta\left(\min\left(\frac{\log(1/\gamma)}{\varepsilon n_s}, \frac{\log(1/\delta)}{\varepsilon n_s+\log n_s}\right)\right) \geqslant d^{-t}.$$

However Conjecture 63 implies that such an algorithm requires

$$n_{\text{robust, eff}}(\alpha,\gamma,\eta_s) \gg d^{2-2t}$$

samples leading to a contradiction. □

9. Both algorithms work for any $\kappa\geqslant 1$.

Appendix F. Covariance Estimation to Regression

In this section we will provide a formal approach to solving robust regression problems given access to a robust covariance estimation algorithm that can estimate the underlying covariance to good accuracy in spectral norm. Specifically, consider the following algorithm $\mathcal{A}$. $\mathcal{A}$ takes as input $\varepsilon-$corrupted samples $\{X_1, X_2, \ldots, X_n\} \sim N(0, \Sigma)$ and outputs $\hat{\Sigma}$ such that

$$0.9\Sigma \preceq \hat{\Sigma} \preceq 1.1\Sigma$$

Such a matrix can be efficiently computed Kothari et al. (2018); Diakonikolas et al. (2025b) using roughly $O(\varepsilon^2 d^2)$ samples. We can draw $2n$ samples from our regression model and use the first n samples to obtain the preconditioner and use the remaining n samples for regression. Abusing notation and reusing the indices, we apply the following preconditioner to the datapoints $\{X_1, X_2, \ldots, X_n\}$. We obtain new samples by performing the following linear operation $X_i' \coloneqq \hat{\Sigma}^{-1/2} X_i$. We then consider the new regression instance

$$y_i = \langle X_i', \hat{\Sigma}^{1/2}\beta \rangle + \eta$$

We observe that this is the same as the model

$$y_i = \langle X_i, \beta \rangle + \eta$$

We now run a robust regression algorithm e.g. Theorem 2 to obtain an estimate $\widetilde{\beta}$ such that $\|\widetilde{\beta} - \hat{\Sigma}^{1/2}\beta\| \lesssim \sigma\sqrt{\varepsilon}$ on this *well-conditioned* instance. We finally output $\hat{\beta} = \hat{\Sigma}^{-1/2}\widetilde{\beta}$. Observe that we have for all $u \in \mathbb{R}^d$ that $u^T\Sigma u \leqslant 2 \cdot u^T\hat{\Sigma}u$. In particular for $u = \hat{\beta} - \beta$, it suffices to show that $u^T\hat{\Sigma}u$ is small. To show this, we observe that

$$(\hat{\beta} - \beta)^T\hat{\Sigma}(\hat{\beta} - \beta) = \|\hat{\Sigma}^{1/2}(\hat{\beta} - \beta)\|^2 = \|\hat{\Sigma}^{1/2}(\hat{\Sigma}^{-1/2}\widetilde{\beta} - \beta)\|^2 = \|\widetilde{\beta} - \hat{\Sigma}^{1/2}\beta\|^2 \lesssim \sigma^2\varepsilon$$

Therefore we have that $\|\hat{\beta} - \beta\|_\Sigma \lesssim \sigma\sqrt{\varepsilon}$.

Appendix G. Fast covariance-aware mean estimation

In this section we will describe a fast algorithm for covariance-aware mean estimation where given ε-corrupted samples from $\mathcal{N}(\mu, \Sigma)$ the goal is to output an estimator $\widehat{\mu}$ such that $\|\Sigma^{-1/2}(\widehat{\mu} - \mu)\|$ is small assuming Σ is unknown. Assuming that we are given corrupted samples from $\mathcal{N}(\mu, \Sigma)$, for $M \cdot I_d \preceq \Sigma \preceq L \cdot I_d$ with condition number $\kappa = L/M$, we first run a fast robust mean estimation procedure for distributions with bounded covariance (Dong et al., 2019, Theorem 1.1). In our case since $\Sigma \preceq L \cdot I_d$ this gives us an approach that uses $\widetilde{O}(d)$ samples, runs in time $\widetilde{O}(nd)$ and outputs an estimate $\widehat{\mu}$ that satisfies

$$\|\widehat{\mu} - \mu\| \leqslant O\Big(\sqrt{L\varepsilon}\Big).$$

We now pay for the norm conversion penalty. Using the same estimator we observe that we get

$$\|\Sigma^{-1/2}(\widehat{\mu} - \mu)\| \leqslant \|\Sigma^{-1/2}\| \cdot \|\widehat{\mu} - \mu\| \leqslant \frac{1}{\sqrt{M}} \cdot \sqrt{L\varepsilon} \leqslant O(\sqrt{\varepsilon\kappa}).$$

Observe that the above error can grow arbitrarily with the condition number.

Appendix H. Code for Verification

The relevant Mathematica and SymPy code for verification is available at the following repository:
`https://github.com/sdeepaknarayanan/RegressionCOLT26`.