Linear Regression under Missing or Corrupted Coordinates

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin-Madison

Jelena Diakonikolas JELENA@CS.WISC.EDU
University of Wisconsin-Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Jasper C. H. Lee JASPERLEE@UCDAVIS.EDU
University of California, Davis

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin-Madison

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study multivariate linear regression under Gaussian covariates in two settings, where data may be erased or corrupted by an adversary under a coordinate-wise budget. In the incomplete data setting, an adversary may inspect the dataset and delete entries in up to an η-fraction of samples per coordinate—a strong form of the Missing Not At Random model. In the corrupted data setting, the adversary instead replaces values arbitrarily, and the corruption locations are unknown to the learner. Despite substantial work on missing data, linear regression under such adversarial missingness remains poorly understood, even information-theoretically. Unlike the clean setting, where estimation error vanishes with more samples, here the optimal error remains a positive function of the problem parameters. Our main contribution is to characterize this error up to constant factors across essentially the entire parameter range. Specifically, we establish novel information-theoretic lower bounds on the achievable error that match the error of (computationally efficient) algorithms. A key implication is that, perhaps surprisingly, the optimal error in the missing data setting matches the error in the corruption setting—so knowing the corruption locations offers no general advantage.

Keywords: High-dimensional Statistics, Linear Regression, Robust Statistics, Missing Data

1. Introduction

We study regression tasks in settings that deviate from the ideal i.i.d. assumption, motivated by the prevalence of imperfect or incomplete data in modern machine learning. Missing values can arise from sensor failures, survey nonresponse, or data aggregation processes such as crowdsourcing (Vuurens et al., 2011) and peer grading (Piech et al., 2013; Kulkarni et al., 2013). Moreover, datasets may include corrupted entries, which may stem from data poisoning attacks (Barreno et al., 2010; Biggio et al., 2012; Steinhardt et al., 2017; Tran et al., 2018; Hayase et al., 2021), out-of-distribution examples (Yang et al., 2024), or biological anomalies (Rosenberg et al., 2002; Paschou et al., 2010; Li et al., 2008). These two types of impurities (incompleteness and corruption) have given rise to two largely distinct lines of research: *learning with missing data* and *robust statistics*.

Incomplete data Rubin's seminal work (Rubin, 1976) classifies missing data mechanisms based on their dependence on observed values: if missingness is independent of the data, it is Missing Completely At Random (MCAR); if it depends only on observed values, it is Missing At Random

(MAR); otherwise, it is Missing Not At Random (MNAR). A common approach to handling missing data is *imputation* of missing entries (Josse et al., 2024; Morvan et al., 2021; Ayme et al., 2023), though not all methods rely on it. Early work in the area focused on parameter estimation (Little, 1992, 1993; Loh and Wainwright, 2012; Rosenbaum and Tsybakov, 2010), whereas more recent research has shifted toward prediction (Morvan et al., 2020a,b; Josse et al., 2024), a more challenging task since missing features during test time can prevent model evaluation.

Corrupted data The robust statistics, initiated in the 1960s (Tukey, 1960; Huber, 1992), focuses on parameter estimation from fully observed data with a small fraction of arbitrary corruptions. A recent resurgence in this area (Lai et al., 2016; Diakonikolas et al., 2019a) has extended classical one-dimensional results to high dimensions while preserving sample and computational efficiency (see Diakonikolas and Kane (2023)). While this literature can also handle missing data by naïvely imputing erased locations with arbitrary values, this simple strategy ignores the fact that the erasure locations are known and can potentially be leveraged by a more sophisticated algorithm. Additional related work on missing data and robust statistics is given in Appendix A.5.

In this work, we study the relationship between these two types of data impurities in the context of parameter estimation in Gaussian linear regression.

Definition 1 (Linear regression) *A labeled example (X, y) follows the linear regression model with regressor β and additive noise level σ if $X \sim \mathcal{N}(0, I)$ and $y = \beta^\top X + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$.*

We model data impurities via a coordinate-wise adversary: for each coordinate[1], at most an η-fraction of samples may be modified by erasure or replacement. Unlike much of the robust statistics literature—which treats each sample as fully clean or corrupted—partial corruptions are common, especially when coordinates correspond to different sensors or measurement methods (Troyanskaya et al., 2001). Definition 2 is a strong MNAR variant, as missing-entry locations are chosen adversarially based on the dataset.

Definition 2 (Coordinate-wise incompleteness) *We define the coordinate-wise incomplete version $\widetilde{S}$ of a set $S = \{(X^{(i)}, y^{(i)})\}_{i\in[n]}$ of n labeled examples, where $X^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$, as follows: An adversary inspects S and, for each coordinate $j \in [d]$, is allowed to erase the j-th coordinate of up to ηn of the $X^{(i)}$'s, replacing them with the special symbol $\perp$ to yield $\{\widetilde{X}^{(i)}\}$. Similarly, it can replace up to an η-fraction of the $y^{(i)}$'s with $\perp$ to yield $\{\widetilde{y}^{(i)}\}$. Then, $\widetilde{S} = \{(\widetilde{X}^{(i)}, \widetilde{y}^{(i)})\}_{i\in[n]}$.*

Definition 3 (Coordinate-wise corruption) *As in Definition 2, but instead of erasing coordinates (replacing them with $\perp$), the adversary may substitute arbitrary values. Crucially, the locations of the modified values are not known to the algorithm.*

The most relevant prior work on Definitions 2 and 3 is Liu et al. (2021), which focuses on mean estimation. While linear regression has been studied under various missing data settings (Loh and Wainwright, 2012; Aladin et al., 2020; Cheng et al., 2023), existing work typically assumes much milder missingness than the fully adversarial pattern in Definition 2. To the best of our knowledge, even for the most "textbook" and standard setting of Gaussian covariates used in Definition 2, linear regression under the aforementioned corruption models remains unexplored, and our understanding is limited even from an information-theoretic standpoint. Thus, we ask:

1. The corruption budget is applied per coordinate rather than globally, to prevent cases where β has a single non-zero entry that receives all corruptions.

What is the optimal estimation error achievable with unlimited samples and computational power?

Due to the adversarial nature of Definition 2, the error need not vanish as the number of samples grows. Instead, the best achievable error is a positive function of η, σ, d, β that we seek to characterize. We provide matching upper and lower bounds, with the upper bounds attained by algorithms that are both sample and computationally efficient.

Another question arises from comparing Definitions 2 and 3. When coordinates are missing rather than replaced, their locations are visible and may be leveraged for improved error, thus:

Is the optimal estimation error under missing data smaller than the error under replaced data?

Prior algorithms for related problems with missing data (Liu et al., 2021; Huber and Ronchetti, 2009), like mean estimation under Definition 2 and low-rank structure (Liu et al., 2021), rely heavily on knowing corruption locations and only handle special cases under replacement corruption. Surprisingly, for linear regression, the answer to the above question is negative.

1.1. Our Results

We now present the upper and lower bounds that address the two stated questions. The upper bounds follow from existing computationally-efficient estimators, while the lower bounds are the main contributions of this work.

We start with three estimators, all applicable to the stronger corruption model of Definition 3 (i.e., replaced data) and efficient in runtime and sample complexity. The first, $\mathcal{A}_1$, is simply using an estimator from existing robust statistics literature, designed for the setting where ε-samples are (entirely) corrupted with $\varepsilon = \eta d$. The second, $\mathcal{A}_2$, is based on the fact that under Definition 1, $\mathbb{E}[yX] = \beta$, thus one can estimate each coordinate of β using existing 1D robust Gaussian mean estimators tolerant to 2η corruption (like trimmed mean or median). The third, $\mathcal{A}_3$, is the trivial algorithm returning always the zero vector. See Appendix A for details on these algorithms. As a note, $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ are known to generalize naturally to other covariate distributions beyond Gaussians, but for concreteness and for clear comparison with our lower bounds, we focus on the most basic setting stated below.

Fact 4 (Upper bounds on estimation error) *There are three polynomial-time algorithms $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ that take as input $n = O(\mathrm{poly}(d/\eta))$ samples from the d-dimensional linear regression model with σ-additive noise (Definition 1), with η-fraction of coordinate-wise corruptions according to Definition 3, and output $\widehat{\beta} \in \mathbb{R}^d$ that satisfies the following guarantees with probability at least 0.99:*

- *Guarantee for $\mathcal{A}_1$: $\|\widehat{\beta} - \beta\| = O(\eta d \sigma)$ whenever $\eta d < 0.49$.*
- *Guarantee for $\mathcal{A}_2$: $\|\widehat{\beta} - \beta\| = O(\eta\sqrt{d}\sqrt{\|\beta\|^2 + \sigma^2})$, whenever $\eta < 0.24$.*
- *Guarantee for $\mathcal{A}_3$: $\|\widehat{\beta} - \beta\| \le \|\beta\|$.*

Depending on η, σ, and $\|\beta\|$, a different algorithm among $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ achieves the lowest error. Tables 2 to 4 in Appendix A present this by partitioning the parameter space into regimes based on which algorithm's error is best (Appendix A.2 has the explicit construction of the partition). In Appendix A.3, we show that $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ can be unified into a single algorithm that adaptively switches between the three strategies to consistently get the best error (up to constant factors).

We now present the lower bounds (the main results of this paper) in a combined statement in Theorem 5. Since some parameter regimes in Theorem 5 overlap, we show only the strongest lower bound in Table 1. A more expanded version is given in Tables 2 to 4 which might be more

convenient to look at, but due to space constraints are moved to Appendix A. Theorem 5 provides information-theoretic lower bounds, applying to any estimator, independent of runtime and for any sufficiently large sample size.[2]

Theorem 5 (Lower bounds on estimation error; combined statement) *The following holds for every $d \in \mathbb{Z}_+$, $\eta \in [0,1], \sigma \in \mathbb{R}_+$, $b \in \mathbb{R}_+$, $c \in (0,1)$, and any algorithm $\mathcal{A}$ that takes as input η, σ, b as well as n labeled examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ with $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$ and outputs a vector in $\mathbb{R}^d$: There exists a $\beta \in \mathbb{R}^d$ with $\|\beta\| = b$ such that running $\mathcal{A}$ to solve the associated linear regression problem (Definition 1) in the coordinate-wise corruption model with missing data (Definition 2), the output $\widehat{\beta}$ satisfies the following with probability at least $\frac{1}{2} - \frac{d+1}{2} e^{-\Omega(c^2 \eta n)}$ over the randomness of the clean samples:*

(a) (Small-β regime; cf. Theorem 22) If $\|\beta\| \le \sigma$, $\mathcal{A}$ has error $\Omega(\min(\|\beta\|, \eta\sqrt{d}\sigma))$.

(b) (Large-η regime; cf. Theorem 23) If $\frac{7}{\sqrt{d}} \le \eta \le 1$, $\mathcal{A}$ has error $\Omega(\|\beta\|)$.

(c) (Medium-η regime; cf. Theorem 25) If $\frac{(2+c)}{d} \le \eta \le \frac{7}{\sqrt{d}}$, $\mathcal{A}$ has error $\Omega(c\eta\sqrt{d}\|\beta\|)$.

(d) (Small-η regime; cf. Theorem 27) If $0 \le \eta \le \frac{1}{d}$, $\mathcal{A}$ has error $\Omega(\min(\eta d\sigma, \eta\sqrt{d}\|\beta\|))$.

Parameter regime	Best lower bound from Theorem 5		Best upper bound from Fact 4	
$\eta \in (0, \frac{0.49}{d}), \quad \|\beta\| \in [\sqrt{d}\sigma, +\infty)$	(part (d))	$\Omega(\eta d\sigma)$	(Alg. $\mathcal{A}_1$)	$O(\eta d\sigma)$
$\eta \in [\frac{0.49}{d}, \frac{7}{\sqrt{d}}), \|\beta\| \in [\sigma, +\infty)$ $\eta \in (0, \frac{0.49}{d}), \quad \|\beta\| \in [\sigma, \sqrt{d}\sigma)$	(part (c))* (part (d))	$\Omega(\eta\sqrt{d}\|\beta\|)$	(Alg. $\mathcal{A}_2$)	$O(\eta\sqrt{d}\|\beta\|))$
$\eta \in [\frac{0.49}{d}, \frac{7}{\sqrt{d}}), \|\beta\| \in [\eta\sqrt{d}\sigma, \sigma)$ $\eta \in (0, \frac{0.49}{d}), \quad \|\beta\| \in [\eta\sqrt{d}\sigma, \sigma)$	(part (a)) (part (a))	$\Omega(\eta\sqrt{d}\sigma)$	(Alg. $\mathcal{A}_2$)	$O(\eta\sqrt{d}\sigma)$
$\eta \in [\frac{7}{\sqrt{d}}, 1], \quad \|\beta\| \in [0, +\infty)$ $\eta \in (0, \frac{7}{\sqrt{d}}), \quad \|\beta\| \in [0, \eta\sqrt{d}\sigma)$	(part (b)) (part (a))	$\Omega(\|\beta\|)$	(Alg. $\mathcal{A}_3$)	$O(\|\beta\|)$

Table 1: Comparison of error bounds across the different possible parameter regimes. *Theorem 5(c) requires $\eta \in [\frac{2+c}{d}, \frac{7}{\sqrt{d}}]$ where $c > 0$ can be any arbitrarily small absolute constant.

A few comments are in order regarding the theorem statement. First, the failure probability in the theorem statement is close to a half (instead of one) because the lower bound is obtained from a reduction to a hypothesis testing problem that asks for distinguishing between two parameter vectors β, β' (see Section 2 for more details), thus a trivial algorithm that outputs each of β, β', with probability $1/2$ succeeds with the same probability. Second, parts (b) and (c) hold under arbitrary additive noise, not necessarily Gaussian. This is because the conclusion as stated in the theorem holds in the noiseless case ($\sigma = 0$), thus by a straightforward reduction this implies hardness under any noise distribution: given an estimator for the noisy model, one could add the same noise to noiseless hard instance and then apply the estimator. Finally, in terms of the covariate distribution,

2. Although there is a dependence on n in the theorem statement, it stems from technicalities (our construction deletes an η-fraction in expectation rather than exactly (as in Definition 2) and this influence vanishes as n grows, so the bounds remain valid even with infinite samples. Moreover, unless $n \gg d$, it is folklore that estimating β to constant error is impossible even with clean data.

Theorem 5 shows that the problem is hard even for the most restricted and basic setting of Gaussian covariates; the hardness trivially extends to any family of distributions that contains Gaussians.

We now discuss the conceptual takeaways and how Table 1 addresses our main research questions.

Missing vs. Replaced Data The upper bounds hold under the stronger model of data replacement (Definition 3), while the lower bounds assume the weaker model of missing data (Definition 2). Since the bounds match (up to constants) for essentially the entire range of η, σ, and $\|\beta\|$, both models are equally hard in terms of optimal estimation error. While this may be unsurprising for the robust statistics framework where samples are either entirely corrupted or uncorrupted, our more delicate lower bound constructions show that the fact remains true even under coordinate-wise erasure/corruption.

Corruption Threshold for Non Trivial Estimation In the classical robust statistics model with ε-fraction of (entirely) corrupted samples, estimation becomes impossible at $\varepsilon = 1/2$, since half the samples could come from two different models and there is no way to tell which are the inliers. Theorem 5(b) shows that the analogous transition point for Definitions 2 and 3 is $\eta = \Theta(1/\sqrt{d})$. This is interesting, as there is no equally obvious explanation for this threshold in this model.

$\|\beta\|$-Dependence on Error is Necessary Ignoring the regime $\eta \geq 1/\sqrt{d}$, where meaningful estimation is impossible, our results show that $\|\beta\|$-dependence in the error is unavoidable in general —in stark contrast to the classical robust model (with each sample either fully clean or corrupted), where optimal error is independent of $\|\beta\|$. Here is the intuition: in that setting, a variant of $\mathcal{A}_2$ gives error $O(\varepsilon(\|\beta\| + \sigma))$ (with ε corruption rate). But by shifting labels via $y \leftarrow y - \widehat{\beta}^\top X$, one reduces the regression target to $\beta - \widehat{\beta}$, which has smaller norm, and can recursively apply the algorithm to eliminate the $\|\beta\|$-dependence. This technique fails under Definition 3: even one corrupted coordinate in X corrupts $\widehat{\beta}^\top X$, and the transformation $y \leftarrow y - \widehat{\beta}^\top X$ can amplify the label corruption rate from η to ηd in a single step. Our results show that no workaround exists.

Role of Label Corruptions As can be inferred from the proofs, all of our lower bounds hold even if y is never masked. Thus, the difficulty in estimating β comes only from the missing X-coordinates.

Additional Discussion Theorem 5(b) may extend to other activation functions. Since Theorem 5(b) and (c) hold for $\sigma = 0$, they imply that for any estimator, the mean squared prediction error can be bounded below for a broad class of GLMs, such as a class of GLMs with non-decreasing Lipschitz activations, including ReLUs in particular. See Appendix A.4 for the relevant discussion.

1.2. Overview of Techniques

The general approach for our lower bounds is the following. We first reduce estimation to a hypothesis testing problem. If no algorithm can distinguish between the regressor being β vs. β', then no estimator can achieve error less than $\|\beta - \beta'\|/2$. To show testing hardness, it suffices to construct a coupling between the distributions of the labeled examples in Definition 1 with regressors β and β' and with small coordinate-wise disagreements, that is, a pair of jointly distributed labeled examples $(X, y), (X', y')$ such that $\mathbb{P}[X_i \neq X'_i] \leq \eta$ for all $i \in [d]$ and $\mathbb{P}[y \neq y'] \leq \eta$. This is because an adversary can then delete differing coordinates, making the datasets indistinguishable (cf. Lemma 9). This leaves us with the (highly) non-trivial task of constructing that coupling, outlined below.

Basic construction The high-level idea of the coupling construction boils down to two main steps: (1) induce the same y-marginal distribution for some sufficiently distinct β and β', and (2) for each value of y, construct a masking scheme to make $X \mid y$ and $X' \mid y$ indistinguishable. Recall that, since X (and X') and the linear regression noise are all Gaussian, the conditional distributions $X \mid y$ and $X' \mid y'$ are also Gaussian; denote them by $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$. In all considered regimes of η, our choices of β and β' are such that the covariances Σ and Σ' are the

same. Then, to obtain a masking scheme that makes the two conditional distributions the same *while fully-leveraging the coordinate-wise masking capabilities of the adversary*, in Lemma 10, we construct a *sequence* of Gaussians $N^{(0)}, N^{(1)}, N^{(2)}, \dots, N^{(d)}$ with the same covariance matrix, which interpolates between $X \mid y$ and $X' \mid y$. Namely, $N^{(0)} = X \mid y$ and $N^{(d)} = X' \mid y$, and each consecutive pair $N^{(i)}, N^{(i+1)}$, $i \in \{0, \dots, d-1\}$, has means that differ over a single coordinate. We then construct a masking scheme for the pair $N^{(i)}$ and $N^{(i+1)}$ using a *maximal coupling*[3] that masks only the i^{th} coordinate, with probability $D_{\mathrm{TV}}(N^{(i)}, N^{(i+1)})$ (cf. Claim 11), and that makes the pair of Gaussians indistinguishable.

Case $\|\beta\| \le \sigma$ The above basic construction is sufficient for $\|\beta\| \le \sigma$ (Theorem 5(a)). In this case, we choose $\beta = (r, (b/\sqrt{d})\mathbb{1}_{d-1})$ and $\beta' = -\beta$, where $b \le \eta\sqrt{d}\sigma$, $\mathbb{1}_\ell$ denotes the dimension-ℓ vector of all ones, and r is chosen to set the norms of β, β' to the target value.

Handling the case $\sigma/\|\beta\| \to 0$ **and** $\eta \gg 1/\sqrt{d}$[4] The basic construction described above is only meaningful (both in terms of the error and the proof itself) when $0 < \|\beta\| \le \sigma$. As $\sigma/\|\beta\| \to 0$, the conditional distributions $X \mid y$ and $X' \mid y$ approach Gaussians with singular covariances, making the TV-distances between the conditionals equal to one. Thus, the distributions cannot be made indistinguishable by deleting only an $\eta < 1$ fraction of coordinates.

Our strategy in this case is to effectively reduce the data dimension by one and treat one of the coordinates as noise. For this discussion, fix $\|\beta\| = \sqrt{d}$, let $\beta = \mathbb{1}_d$ and $\beta' = -\beta$; the same ideas transfer to other values of $\|\beta\|$. The core task again is to couple $X \mid y$ with $X' \mid y'$. Now, given the particular form of β and β' and that $\beta = -\beta'$, the core task can be re-phrased as coupling $X \sim \mathcal{N}(0, I)$ conditioned on $\sum_i X_i = t$ with X' conditioned on $\sum_i X'_i = t'$ where $t = -t'$. Since there is no noise ($\sigma = 0$), the key idea is to treat the last coordinate as "noise": we couple the first $d-1$ coordinates using the basic coupling method, which gives $\mathbb{P}[X_i \ne X'_i] \le |t - t'|/d$ for $i \in [d-1]$. We then set the final coordinate to ensure the total sums are t and t'. This makes the last coordinate differ with probability 1 between the two scenarios, but randomizing the coordinate order leads to coupled X, X' where each coordinate differs with probability $(|t - t'| + 1)/d$. This idea and construction is formally stated in Lemma 13 (which gives a construction for general t and t' instead of only $t = -t'$). Finally, since in expectation $|t - t'| = 2|t| = 2|y|$ is $O(\sqrt{d})$, this yields a bound of $\Omega(\|\beta\|)$ as long as $\eta \gg 1/\sqrt{d}$, leading to Theorem 5(b).

Handling $1/d \ll \eta \ll 1/\sqrt{d}$ When $1/d \ll \eta \ll 1/\sqrt{d}$, the approach described so far fails because $\mathbb{E}[|t - t'|]$ is too large, making the probability of coordinate-disagreement larger than the assumed η upper bound. Instead, we construct a new hard instance with β and β' proportional to $(\varepsilon\mathbb{1}_{d/2}, \mathbb{1}_{d/2})$ and $(-\varepsilon\mathbb{1}_{d/2}, \mathbb{1}_{d/2})$ respectively. The coupling proceeds as follows: sample the labels $y = y'$ from the distribution $\mathcal{N}(0, (1+\varepsilon^2)/d))$, which is the distribution of the labels. Now consider coupling $t := \sum_{i=1}^{d/2} X_i$ and $t' := \sum_{i=1}^{d/2} X'_i$, which are the sums of the first half of the coordinates under each hypothesis. From calculations, the marginals t and t' are both univariate Gaussians with the same variance, with difference in mean equal to $\Theta(\varepsilon y)$. We thus sample t and t' in a coupled way where $t - t' = \Theta(\varepsilon y)$ always (this is possible as the distribution of t is the same as the distribution of t' shifted by $\Theta(\varepsilon y)$). Now it remains to couple $X \mid y$ and $X' \mid y'$, which we do by *first* coupling the first half of the coordinates using the coupling between t and t', then coupling the second half of the coordinates based on the values of y, t and y', t'. More concretely, we use Lemma 13 to couple $(X_1, \dots, X_{d/2}) \mid t$ and $(X'_1, \dots, X'_{d/2}) \mid t'$. Then, we use Lemma 13 again to couple

3. There is always a coupling (called *maximal*) between P, Q with $\mathbb{P}[X \ne X'] = D_{\mathrm{TV}}(P, Q)$ (Fact 7).
4. Henceforth, the notation $a \gg b$ indicates that where exists some sufficiently large constant $C > 0$ such that $a \ge Cb$, and $\ll$ is defined similarly. For the discussion in this section, $C \in [1, 4]$.

$(X_{d/2+1}, \ldots, X_d)$ and $(X'_{d/2+1}, \ldots, X'_d)$ so their sums are $y - \varepsilon t$ and $y + \varepsilon t'$, respectively. Going through the calculations, this strategy yields an expected number of disagreeing coordinates bounded by $\Theta(1 + \varepsilon\sqrt{d})$, which is at most ηd as long as $\varepsilon \ll \eta\sqrt{d}$ (recall that we assume $\eta \gg 1/d$ in this paragraph). This translates to an estimation error lower bound of $\Omega(\eta\sqrt{d}\|\beta\|)$, proving Theorem 5(c).

The case of small η The remaining regime is $\eta \ll 1/d$ (Theorem 5(d)). The (already complicated) construction in the last paragraph does not handle that case, given that the expected number of disagreeing coordinates there is $\Omega(1) \geq 1/d \gg \eta$. For this final regime, we use a similar hard hypothesis test instance, showing the impossibility of distinguishing between $\beta = (B\mathbb{1}_{d/2}, E\mathbb{1}_{d/2})/\sqrt{d/2}$ and $\beta' = (B\mathbb{1}_{d/2}, -E\mathbb{1}_{d/2})/\sqrt{d/2}$ for appropriately chosen parameters B and E. The main technical innovation here is a deconstruction of label values into the contribution from the first and second half of the coordinates and a more refined way to sample them in the coupling construction. This is sketched in Section 4 and detailed in Appendix D.

1.3. Preliminaries

We include only the essential preliminaries here; the full version appears in Appendix B. For X, Y with pdfs P_X and P_Y, we write $D_{\mathrm{TV}}(X, Y)$ or $D_{\mathrm{TV}}(P_X, P_Y)$ for the total variation distance, defined as $\frac{1}{2}\int |P_X(u) - P_Y(u)|\mathrm{d}u$. We write $D_{\mathrm{KL}}(X \,\|\, Y)$ or $D_{\mathrm{KL}}(P_X \,\|\, P_Y)$ for the KL divergence, defined as: $D_{\mathrm{KL}}(P_X \,\|\, P_Y) = \int_x P_X(z)\log(P_X(z)/P_Y(z))\mathrm{d}z$. For a vector x, $\|x\|$ denotes its Euclidean norm. Let I_d be the $d \times d$ identity matrix, and $\mathbb{1}_d$ the all-ones vector in $\mathbb{R}^d$. We use $\top$ to denote matrix transposes. We write $a \lesssim b$ (or $O(b)$) to mean $a \leq Cb$ for an absolute constant $C > 0$, independent of the parameters involved (similarly for $\Omega(\cdot)$). We write $\widetilde{O}$ and $\widetilde{\Omega}$ to hide polylog factors.

Fact 6 (see, e.g., Section 8.1.3. in Petersen and Pedersen (2008)) *If $X \sim \mathcal{N}(\mu, I)$ is a Gaussian vector in $\mathbb{R}^d$, $\xi \sim \mathcal{N}(0, \sigma^2)$ is a univariate Gaussian, and $u \in \mathbb{R}^d$ is a fixed vector, then the distribution of X conditioned on $u^\top X + \xi = r$ is the Gaussian $\mathcal{N}\left(ru/(\|u\|^2 + \sigma^2),\ I - uu^\top/(\|u\|^2 + \sigma^2)\right)$.*

If P and Q are distributions over $\mathcal{X}$, a coupling Π between P and Q is any distribution over $\mathcal{X} \times \mathcal{X}$ such that under $(X, Y) \sim \Pi$, the marginals are $X \sim P$ and $Y \sim Q$.

Fact 7 (Maximal coupling (see, e.g., Roch (2024))) *Let P and Q be distributions. There exists a coupling Π between P and Q such that $\mathbb{P}_{(X,Y)\sim\Pi}[X \neq Y] = D_{\mathrm{TV}}(P, Q)$. Moreover for distributions on P, Q on $\mathbb{R}$ where P is a shifted version of Q, if μ_P, μ_Q denote their expectations, Π additionally satisfies $\mathbb{E}_{(X,Y)\sim\Pi}[|X - Y|] = |\mu_P - \mu_Q|$.*

Fact 8 (Total variation between univariate Gaussians (see, e.g., Petersen and Pedersen (2008)) *If $D_1 = \mathcal{N}(\mu_1, \sigma^2)$ and $D_2 = \mathcal{N}(\mu_1, \sigma^2)$ then $D_{\mathrm{TV}}(D_1, D_2) \leq (1/\sqrt{2})|\mu_1 - \mu_2|/\sigma$.*

2. Warm-up: Coupling via Hybrid Argument and Proof of Theorem 5(a)

We show that our goal reduces to finding a coupling with bounded coordinate-wise disagreements. We also present a method (the "hybrid argument") for building such couplings. This alone suffices for Theorem 5(a), but stronger results require a modification that we give in Section 3.

Estimation Hardness via Hypothesis Testing To show linear regression hardness with error $< \varepsilon$, it suffices to show an indistinguishability result for vectors $\beta^{(0)}, \beta^{(1)}$ with $\|\beta^{(0)} - \beta^{(1)}\| > 2\varepsilon$:

- (Null Hypothesis) The data follows the model in Definitions 1 and 2 with $\beta = \beta^{(0)}$.
- (Alternative Hypothesis) The data follows the model in Definitions 1 and 2 with $\beta = \beta^{(1)}$.

This is sufficient because if an estimator $\widehat{\beta}$ exists with error less than ε, the test that rejects the null hypothesis when $\|\widehat{\beta} - \beta^{(0)}\| > \|\widehat{\beta} - \beta^{(1)}\|$ will succeed, as shown by a simple triangle inequality.

Couplings with Small Coordinate-Wise Disagreements We now argue that constructing a suitable coupling suffices for proving hardness of the testing problem. One can think of data generation as follows: a pair of two samples $((X, y), (X', y'))$ is drawn from a coupling Π between the distributions of Definition 1 with $\beta = \beta^{(0)}$ and $\beta = \beta^{(1)}$. Depending on the hypothesis, either (X, y) or (X', y') is added to the dataset, after which the adversary may modify it (Definition 2). If $\mathbb{E}[\mathbb{1}(X_i \neq X_i')] \leq \eta$ for all $i \in [d]$, then by deleting disagreeing coordinates, the adversary can make the dataset look identical under both hypotheses, making testing impossible. Lemma 9 formalizes this. This lemma actually uses a slightly relaxed coordinate-wise disagreement condition: $\mathbb{E}[\sum_i \mathbb{1}(X_i \neq X_i)] \leq \eta d$. At first glance, this may appear to correspond to a slightly different definition of the contamination model, in which the adversary is allowed to corrupt ηd coordinates *per sample* in expectation. However, Definition 2 instead imposes a budget of ηn corruptions *per coordinate*. We can overcome this issue by assuming that the distribution of (X, y) is invariant to permutations of the coordinates of X, because this would imply that after a random permutation each coordinate is deleted with equal probability (at most η). Moreover, because the adversary's budget in Definition 2 is deterministic (rather than in expectation), there is a small probability—dictated by the Chernoff–Hoeffding bound—that our randomized adversary does not satisfy Definition 2. In this case, we may simply assume that the testing problem is solvable. This completes the proof sketch of Lemma 9. Finally, we note that we will actually require the permutation invariance condition in the statement of Lemma 9 to hold separately over the first and second halves of the coordinates, rather than over the entire vector. This modification ensures that the lemma remains applicable in the more involved construction of Theorem 5(d), which treats the two halves of the coordinates independently. The formal proof is given in Appendix C.1.

Lemma 9 *Let D, D' be distributions over labeled examples (X, y) with $X \in \mathbb{R}^d$, $y \in \mathbb{R}$. Assume that for any two permutations $\pi_1, \pi_2 \in S_{d/2}$, the distribution of $(X, y) \sim D$ is the same as that of $(\pi_1(X_1), \pi_2(X_2), y)$, where X_1 and X_2 denote the first and last $d/2$ coordinates of X, respectively. Assume that the same property also holds for D'. Consider the hypothesis testing problem where the null hypothesis is that data are drawn from D under the corruption model of Definition 2, and the alternative hypothesis is that data are drawn from D'. If there exists a coupling Π between D, D' such that for some $c \in (0, 1)$: $\mathbb{P}_{((X,y),(X',y'))\sim\Pi}[y \neq y'] \leq \frac{\eta}{2}(1 - c)$ and*

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] \leq \frac{\eta}{2} d(1 - c), \tag{1}$$

then no test can distinguish D from D' with probability greater than $\frac{1}{2} + \frac{(d+1)}{2} e^{-\Omega(c^2 \eta n)}$.

Hybrid Argument for Coupling Construction We discuss an argument for constructing couplings satisfying Equation (1). Instead of coupling the distributions of Definition 1, we will first show, in

Lemma 10 below, how to construct such couplings between two d-dimensional Gaussians D and D' with same covariance and different means. As we will see in the next subsection, this directly leads to a coupling between the distributions of the linear regression model.

Lemma 10 (Hybrid argument for constructing couplings) *Let $D = \mathcal{N}((\mu_1, \dots, \mu_d), \Sigma)$ and $D' = \mathcal{N}((\mu'_1, \dots, \mu'_d), \Sigma)$. There exists a coupling Π between D and D' such that*

$$\mathop{\mathbb{E}}_{(X,X')\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X'_i)\right] = \sum_{i=0}^{d-1} D_{\mathrm{TV}}\left(Q_i, Q_{i+1}\right),$$

where $Q_i = \mathcal{N}((\mu'_1\mu'_2, \dots, \mu'_i, \mu_{i+1}, \dots, \mu_d), \Sigma)$ is the Gaussian whose mean has the same values as μ' in the first i coordinates and the same as μ elsewhere ($Q_0 = D$ and $Q_d = D'$).

The idea is the following: if the two means differ only in the i-th coordinate, we can adapt the standard maximal coupling (Fact 7) so that any disagreement comes exclusively from the i-th coordinate.

Claim 11 *Consider the two d-dimensional Gaussians $Q = \mathcal{N}((\mu_1, \mu_2, \dots, \mu_d), \Sigma)$ and $Q' = \mathcal{N}((\mu_1, \dots, \mu_{i-1}, \mu'_i, \mu_{i+1}, \dots, \mu_d), \Sigma)$. There is a coupling Π between the distributions Q and Q' such that $\mathbb{P}_{(X,X')\sim\Pi}[X_i \neq X'_i] = D_{\mathrm{TV}}(Q, Q')$ and $\mathbb{P}_{(X,X')\sim\Pi}[X_j \neq X'_j] = 0$ for all $j \neq i$.*

Claim 11 holds because the marginals over $[d]\setminus\{i\}$ are identical under Q and Q', allowing us to match these coordinates, and apply Fact 7 to the distribution of the i-th coordinate conditioned on the rest.

To prove Lemma 10 for means differing in more coordinates, we can consider the distributions Q_i from the statement and couple each consecutive pair using Claim 11. Chaining together the couplings completes the proof of Lemma 10. The formal proofs are in Appendix C.1.

Proof Sketch of Theorem 5(a) We present a proof sketch with routine calculations omitted here; the full proof is in Appendix C.2. We first focus on the regime $\|\beta\| \le \eta\sqrt{d}\sigma/(2\sqrt{2})$. The remaining regime can be handled by an argument that essentially reduces to the $\|\beta\| \le \eta\sqrt{d}\sigma/(2\sqrt{2})$ regime and will be explained at the end. For any $b \in [0, \eta\sqrt{d}\sigma/(2\sqrt{2})]$, we consider the hypothesis testing problem of distinguishing between the regression vectors $\beta^{(0)} = (b/\sqrt{d}, \dots, b/\sqrt{d})$ and $\beta^{(1)} = -\beta^{(0)}$ from n samples from the models in Definitions 1 and 2. Note that $\|\beta^{(0)}\| = \|\beta^{(1)}\| = b$ and $\|\beta^{(0)} - \beta^{(1)}\| = \sqrt{2}b$. By the reduction from estimation to hypothesis testing, showing that no algorithm can solve this testing problem implies that no estimator has Euclidean error smaller than $b/\sqrt{2}$.

To use Lemma 9, we need a coupling $(X, y), (X', y')$ where (X, y) follows Definition 1 with $\beta = \beta^{(0)}$ and (X', y') follows the same model with $\beta = \beta^{(1)}$, satisfying $\mathbb{E}[\sum_{i=1}^{d} \mathbb{1}(X_i = X'_i)] \le \eta d$ and $\mathbb{P}[y \neq y'] \le \eta$. We construct this coupling by: (i) drawing t from $\mathcal{N}(0, \|\beta\|^2 + \sigma^2)$, and setting $y = y' = t$ (the label distribution is the same for both hypotheses), and (ii) drawing X and X' from an appropriate coupling of the conditional distributions given $y = t$. We thus focus on how to do step (ii) with small coordinate-wise disagreements.

Claim 12 *For $t \in \mathbb{R}$, $\sigma > 0$ and $b \in [0, \eta\sqrt{d}\sigma/(2\sqrt{2})]$, let D be the distribution of $X|(y = t)$ under $X \sim \mathcal{N}(0, I_d)$ and $y = (b/\sqrt{d}, \dots, b/\sqrt{d})^\top X + \xi$ for $\xi \sim \mathcal{N}(0, \sigma^2)$. Let D' be defined similarly but using the conditioning $y = -t$. There exists a coupling Π between D and D' such that $\mathbb{E}_{(X,X')\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X'_i)\right] \le \sqrt{2d}|t|b/(\sigma^2 + b^2)$.*

Proof Let $\beta^{(0)} = (b/\sqrt{d}, \dots, b/\sqrt{d})$ and $\beta^{(1)} = -\beta^{(0)}$ as before. By Fact 6, $D^{(i)}$ for $i = 0, 1$ are the Gaussians $\mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$ where

$$\mu^{(i)} := \frac{t}{\sigma^2 + b^2}\beta^{(i)}, \quad \Sigma^{(1)}{=}\Sigma^{(2)}{=}\Sigma := I - \frac{\beta^{(1)}(\beta^{(1)})^\top}{\sigma^2 + b^2}.$$

We now use Lemma 10 to couple $\mathcal{N}(\mu^{(0)}, \Sigma), \mathcal{N}(\mu^{(1)}, \Sigma)$. For each $i \in [d]$ we need to upper bound the TV-distance between the two Gaussians $\mathcal{N}(m^{(i)}, \Sigma), \mathcal{N}(m^{(i+1)}, \Sigma)$, where $m^{(i)} = (\mu_1^{(1)}, \dots, \mu_{i-1}^{(1)}, \mu_i^{(0)}, \mu_{i+1}^{(0)}, \dots, \mu_d^{(0)})$ and $m^{(i+1)} = (\mu_1^{(1)}, \dots, \mu_{i-1}^{(1)}, \mu_i^{(1)}, \mu_{i+1}^{(0)}, \dots, \mu_d^{(0)})$. First, by Pinsker's inequality it suffices to bound $\frac{1}{\sqrt{2}} D_{\mathrm{KL}}(\mathcal{N}(m^{(i)}, \Sigma) \,\|\, \mathcal{N}(m^{(i+1)}, \Sigma))^{1/2}$. The KL-divergence can then be bounded by (see Equation (11)-(16) in Appendix C.2 for the missing steps):

$$\sqrt{D_{\mathrm{KL}}(\mathcal{N}(m^{(i)}, \Sigma) \,\|\, \mathcal{N}(m^{(i+1)}, \Sigma))} \le \sqrt{\tfrac{1}{2}(m^{(i+1)} - m^{(i)})^\top \Sigma^{-1} (m^{(i+1)} - m^{(i)})} \tag{2}$$
$$\le \sqrt{\tfrac{1}{2}(m^{(i+1)} - m^{(i)})^\top (I + \tfrac{\beta^{(1)}\beta^{(1)\top}}{\sigma})(m^{(i+1)} - m^{(i)})} \le \frac{\sqrt{2}|t|}{\sqrt{d}} \frac{b}{\sigma^2 + b^2}.$$

Adding together all the TV-distance terms for $i = 1, \dots, d$ concludes the proof of Claim 12. ■

By turning this into a coupling for the full labeled examples as described in the beginning of this proof sketch, the expected sum of disagreements is bounded by simply taking expectation with respect to $t \sim \mathcal{N}(0, \sigma^2 + b^2)$ above. By using $\mathbb{E}[|t|] \le \sqrt{\sigma^2 + b^2}$ and plugging in $b \le \eta\sqrt{d}\sigma/(2\sqrt{2})$, the expected sum coordinate-wise disagreement becomes at most $\eta d/2$. Applying Lemma 9 completes the proof of the lower bound $\Omega(\|\beta\|)$ in the regime $\|\beta\| \le \eta\sqrt{d}\sigma/(2\sqrt{2})$.

For the remaining regime $\|\beta\| > \eta\sqrt{d}\sigma/(2\sqrt{2})$, consider instead the testing problem in $\mathbb{R}^d$ between $\beta = (r, b/\sqrt{d}, \dots, b/\sqrt{d})$ and $\beta' = (r, -b/\sqrt{d}, \dots, -b/\sqrt{d})$, where $b \le \eta\sqrt{d}\sigma/(2\sqrt{2})$ is as before and r is a tunable parameter allowing $\|\beta\|$ to get any desired value larger than $\eta\sqrt{d}\sigma/(2\sqrt{2})$. The labels can be written as $y = rX_1 + y_0$ and $y' = rX_1' + y_0'$, where $y_0 = (b/\sqrt{d}, \dots, b/\sqrt{d})^\top X_{2:d} + \xi$ and $y_0' = (-b/\sqrt{d}, \dots, -b/\sqrt{d})^\top X_{2:d}' + \xi'$, with $X_{2:d}$ denoting the last d coordinates of X. From earlier, there exists a coupling between $(X_{2:d}, y_0)$ and $(X_{2:d}', y_0')$ with expected disagreements at most $\eta d/2$ per sample which extends trivially to the full (X, y) and (X', y') by adding shared Gaussian noise $\mathcal{N}(0, r^2)$ to the labels. Applying Lemma 9 as before completes the proof. ■

3. Improved Core Coupling Construction

The previous strategy has a key limitation that the result holds only for $\|\beta\| \le \sigma$. When $\sigma = 0$, the matrix Σ in Equation (2) becomes singular, causing the proof to break down and yield no meaningful bound. For some regimes in Theorem 5, we want to overcome this, for example, to establish the lower bound $\Omega(\|\beta\|)$ whenever $\eta \ge 7/\sqrt{d}$, regardless of the value of σ. The key observation is that, because β in Claim 12 points in the all-ones direction, the zero-eigenvalue issue disappears when we restrict our attention to the first $d - 1$ coordinates. We can thus apply the technique to those coordinates and adjust the final one to ensure the total sum is correct. This idea is formalized in Lemma 13, which provides an improved version of Claim 12 for the case $b = \sqrt{d}$ and $\sigma = 0$ (the result extends trivially to any scaling of the vector and any $\sigma > 0$). We will later use Lemma 13 to prove our remaining main results.

Lemma 13 (Improved core coupling) *For any $d \in \mathbb{Z}_+$, $t, t' \in \mathbb{R}$, the following hold. If D denotes the distribution of $(X_1, \dots, X_d) \sim \mathcal{N}(0, I_d)$ conditioned on $\sum_{i=1}^d X_i = t$ and D' the distribution*

of $(X'_1,\dots,X'_d)\sim\mathcal{N}(0,I_d)$ *conditioned on* $\sum_{i=1}^d X'_i=t'$*, then there exists a coupling* Π *between* D,D' *such that* $\mathbb{E}_{(X,X'))\sim\Pi}\left[\sum_{i=1}^d \mathbb{1}\left(X_i\neq X'_i\right)\right]\le 1+|t-t'|$.

Proof Consider generating the pair (X,X') as follows:

- Let Π' be the coupling from Lemma 10 between the first $(d-1)$ coordinates of D and D'. Sample $(X_1,\dots,X_{d-1}),(X'_1,\dots,X'_{d-1})$ from Π'.
- Set the last coordinate as $X_d=t-\sum_{i=1}^{d-1}X_i$ and $X'_d=t'-\sum_{i=1}^{d-1}X'_i$.

By construction, $X\sim D$ and $X'\sim D'$, so the above defines a valid coupling between D and D'. We denote this coupling by Π, and claim that it satisfies the lemma's guarantee. By Lemma 10 and the trivial bound $\mathbb{1}(X_d\neq X'_d)\le 1$ on the d-th coordinate, we have

$$\mathop{\mathbb{E}}_{(X,X')\sim\Pi}\left[\sum_{i=1}^{d}\mathbb{1}(X_i\neq X'_i)\right]\le 1+\sum_{i=0}^{d-2}D_{\mathrm{TV}}\left(Q_i,Q_{i+1}\right), \tag{3}$$

where the Q_i are as defined in Lemma 10. It remains to determine their exact form and bound their total variation distances. Applying Fact 6 with $u=\mathbb{1}_{d-1}$ (the all-ones vector in $\mathbb{R}^{d-1}$) and $\sigma=0$, we find that the first $d-1$ coordinates of D and D' follow $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\mu',\Sigma)$, respectively, where $\mu=\mathbb{1}_{d-1}t/d$, $\mu'=\mathbb{1}_{d-1}t'/d$, and $\Sigma=I_{d-1}-\mathbb{1}_{d-1}\mathbb{1}_{d-1}^\top/d$. Thus, each Q_i is a Gaussian with covariance Σ and mean $\mu^{(i)}$, where $\mu^{(i)}$ equals t'/d on the first i coordinates and t/d on the remaining ones. By Pinsker's inequality, each term $D_{\mathrm{TV}}(Q_i,Q_{i+1})$ in Lemma 10 is

$$\begin{aligned}
&D_{\mathrm{TV}}(Q_i,Q_{i+1})\le\sqrt{\tfrac{1}{2}D_{\mathrm{KL}}(Q_i\,\|\,Q_{i+1})}\le\sqrt{\tfrac{1}{4}(\mu^{(i+1)}-\mu^{(i)})^\top\Sigma^{-1}(\mu^{(i+1)}-\mu^{(i)})} && \text{(Fact 18)}\\
&=\sqrt{\tfrac{1}{4}(\mu^{(i+1)}-\mu^{(i)})^\top\left(I+\mathbb{1}_{d-1}\mathbb{1}_{d-1}^\top\right)(\mu^{(i+1)}-\mu^{(i)})} && \text{(Fact 21)}\\
&\le\sqrt{\tfrac{1}{4}\left(\|\mu^{(i+1)}-\mu^{(i)}\|^2+|(\mu^{(i+1)}-\mu^{(i)})^\top\mathbb{1}_{d-1}|^2\right)}\\
&=\sqrt{\tfrac{1}{4}\left(|t-t'|^2/d^2+|t-t'|^2/d^2\right)}=\frac{1}{\sqrt{2}}\frac{|t-t'|}{d}. && (4)
\end{aligned}$$

Thus, plugging this to Equation (3) we obtain $\mathbb{E}_{(X,X')\sim\Pi}[\sum_{i=1}^d\mathbb{1}(X_i\neq X'_i)]\le 1+|t-t'|$. ■

4. Proofs of parts (b)-(d) of Theorem 5

The proof of Theorem 5(b) is almost the same as the proof of Theorem 5(a): we test between linear regression setups with $\sigma=0$ and $\beta=(1,\dots,1)$ vs. $\beta'=-\beta$. It suffices to use $\sigma=0$ and $\|\beta\|=\sqrt{d}$, since coupling these cases implies a coupling for any scaling $\|\beta\|$ and $\sigma>0$ (see Appendix D for the full proof). Using Lemmata 9 and 13 completes the proof.

The remaining lower bound proofs also use Lemma 13. However, since we no longer aim to prove a lower bound of order $\|\beta\|$ (as the upper bound is smaller), we will not use the hypothesis testing setup with $\beta=(1,\dots,1)$ vs. $\beta'=(-1,\dots,-1)$. Instead, we adjust it as follows. The full proofs for this section are provided in Appendix D.

Proof Sketch of Theorem 5(c) For this lower bound we need a slightly different hypothesis testing problem: Our two hypotheses will use regressors $\beta=s(\varepsilon,\dots,\varepsilon,1,\dots,1)$ and $\beta'=s(-\varepsilon,\dots,-\varepsilon,1,\dots,1)$ respectively, where $s>0$ can be any scaling factor (so that our lower

bound holds for every possible norm of $\|\beta\|$) and $\varepsilon \le 1$ will be chosen shortly. We claim that Lemma 13 implies a coupling Π between the distributions of Definition 1 for β and β' such that $\mathbb{P}_{((X,y)(X',y'))\sim\Pi}[y \neq y'] = 0$ and

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X_i')\right] \le 2 + O(\varepsilon\sqrt{d}). \tag{5}$$

We first complete the theorem proof sketch given Equation (5). Setting $\varepsilon := (\eta - 4/d)\frac{\sqrt{d}}{2C}$ (for C the constant in the big-O), the RHS of Equation (5) becomes $\eta d/2$, making Lemma 9 applicable. By the reduction from estimation to testing, this implies every estimator has error $\Omega(\|\beta - \beta'\|)$, which is $\|\beta - \beta'\| = \varepsilon s\sqrt{d/2} = \frac{\varepsilon}{\sqrt{1+\varepsilon^2}}\|\beta\| \gtrsim \varepsilon\|\beta\| \gtrsim (\eta - 4/d)\sqrt{d}\|\beta\| \gtrsim \eta\sqrt{d}\|\beta\|$, where the last step used $\eta \ge 5/d$. In Appendix we use better constants so only require $\eta \ge (2+o(1))/d$.

We now prove Equation (5). By scaling, it suffices to do this only for $s = 1$ and $\sigma = 0$ (i.e., no additive noise). The coupling is the following: (i) Sample z from the distribution of labels (which is $\mathcal{N}(0, \sigma^2 + \|\beta\|^2) = \mathcal{N}(0, (1+\varepsilon^2)d/2)$ for both hypotheses) and set $y = y' = z$. (ii) Sample t and t' from the distribution of the sum of the first half of the coordinates for the two hypotheses. By using standard facts for Gaussians, it turns out that these two distributions are $\mathcal{N}(\varepsilon z/(1+\varepsilon^2), d/(2(1+\varepsilon^2)), \mathcal{N}(-\varepsilon z/(1+\varepsilon^2), d/(2(1+\varepsilon^2))$. We can also choose to have that $t - t' = 2\varepsilon z/(1+\varepsilon^2)$ always. (iii) Sample the first half of the coordinates using the coupling of Lemma 13 such that they sum to $z - \varepsilon t$. (iv) Sample the second half using Lemma 13 such that they sum to $z + \varepsilon t'$. The expected disagreements from the two applications of Lemma 13 are

$$\mathbb{E}\left[\textstyle\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] \le 2 + O\big(\mathop{\mathbb{E}}_{t,t'}\left[|t-t'|\right] + \varepsilon \mathop{\mathbb{E}}_{t,t'}\left[|t+t'|\right]\big).$$

Using $t - t' = 2\varepsilon z/(1+\varepsilon^2)$ with $z \sim \mathcal{N}(0, (1+\varepsilon^2)d/2)$, $\varepsilon \le 1$ and that t, t' have variance $O(d)$ proves Equation (5). ■

The constant additive term in Equation (5) breaks the proof of Theorem 5(c) when $\eta < 2/d$. To handle this, we improve the coupling below. Roughly speaking, the idea is to break the variables we want to couple into smaller parts and couple the parts separately. When the first parts match, we force the rest to match as well, which (in expectation) reduces the coordinate-wise disagreements.

Proof Sketch of Theorem 5(d) Let $B, E \in (0,1)$ be parameters, let $\mathbb{1}_d \in \mathbb{R}^d$ denote the all-ones vector, and define the problem of distinguishing between regressor $\beta := (\frac{B}{\sqrt{d/2}}\mathbb{1}_{d/2}, \frac{E}{\sqrt{d/2}}\mathbb{1}_{d/2})$ and $\beta' := (\frac{B}{\sqrt{d/2}}\mathbb{1}_{d/2}, \frac{-E}{\sqrt{d/2}}\mathbb{1}_{d/2})$. The claim now is existence of a coupling Π between labeled examples of the linear regression models of Definition 1 with β and β' such that, $\mathbb{P}_{((X,y)(X',y'))\sim\Pi}[y \neq y'] = 0$ and

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X_i')\right] \lesssim \frac{E}{\sigma} + \sqrt{d}\frac{E}{B}. \tag{6}$$

As before, we aim to apply Lemma 9 to conclude that the hypothesis testing problem is hard. To do that, we need to ensure that the RHS above is at most $\eta d/2$. Equivalently $E/\sigma \ll \eta d$ and $\sqrt{d}E/B \ll \eta d$. We thus choose $E := \frac{1}{2C}\min(\eta d\sigma, \eta\sqrt{d}B)$. Noting that $\|\beta - \beta'\| \gtrsim \min(\eta d\sigma, \eta\sqrt{d}\|\beta\|)$ concludes the proof of Theorem 5(d).

It remains to argue Equation (6). The coupling for that is the following. We will use the notation $X = (X_1, X_2)$ to denote the first and second half of coordinates of X and $S_1 = \mathbb{1}_{d/2}^\top X_1/\sqrt{d/2}$,

$S_2 = \mathbb{1}_{d/2}^\top X_2/\sqrt{d/2}$ to denote the scaled sums of these coordinates (and we will use similar primed letters notation for the corresponding parts of X'). The coupling is the following: (i) Draw y from the distribution of labels and let $y' = y$. (ii) Draw (S_2, S_2') from a maximal coupling between the distributions under the null and alternative hypotheses conditioned on y. (iii) If $S_2 = S_2'$: draw X_2 from the null distribution conditioned on the value of S_2 from the previous step, and set $X_2' = X_2$. Draw (S_1, S_1') from a maximal coupling between null and alternative distributions conditioned on (S_2, y). (iii a) If $S_1 = S_1'$ draw X_1 from the null distribution conditioned on S_1 and set $X_1' = X_1$. (iii b) Otherwise use Lemma 13 to jointly sample (X_1, X_1') conditioned on the values of S_1 and S_1'. (iv) If $S_2 \neq S_2'$: couple the conditional distributions S_1 under the null and S_1' under the alternative given (S_2, y) and (S_2', y), respectively. Then use Lemma 13 to jointly sample (X_2, X_2') conditioned on S_2, S_2', and similarly for (X_1, X_1') conditioned on S_1, S_1'.

It is true (and straightforward) that this is indeed a coupling, since the labeled examples are constructed in steps and each step uses the correct conditional distributions under the null/alternative hypotheses, conditioned on all previous steps.

For the disagreement analysis, one must derive the precise form of all the conditional distributions mentioned above and apply Fact 7 and lemma 13 appropriately. This requires some technical work, which we defer to Appendix D. At a high level however, the key point is that the case-splitting in the coupling construction implies that in some cases there are no disagreements, and Lemma 13 is applied only to the remaining cases, which occur with probability strictly less than 1. As a result, the additive 1 term in the disagreement bound from Lemma 13 is scaled by the probability of the corresponding case. Ultimately, this ensures that the expected total number of disagreements is strictly less than 1, which in turn allows our theorem to apply in the regime $\eta < 1/d$.

Acknowledgments

Ilias Diakonikolas was supported by NSF Medium Award CCF-2107079, ONR award number N00014-25-1-2268, and an H.I. Romnes Faculty Fellowship. Jelena Diakonikolas was supported in part by the Air Force Office of Scientific Research under award number FA9550-24-1-0076, by the U.S. Office of Naval Research under contract number N00014-22-1-2348, and by the NSF CAREER Award CCF-2440563. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense. Daniel M. Kane was supported by NSF Medium Award CCF-2107547 and NSF Award CCF-1553288 (CAREER). This work was done while Jasper C. H. Lee was at University of Wisconsin-Madison. At that time he was supported by NSF Medium Award CCF-2107079, NSF AiTF Award CCF-2006206, and a Croucher Fellowship for Postdoctoral Research. Thanasis Pittas was supported by NSF Medium Award CCF-2107079 and NSF Award DMS-2023239 (TRIPODS).

References

K. A. Chandrasekher Aladin, A. el Alaoui, and A. Montanari. Imputation for High-Dimensional Linear Regression, 2020.

A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Minimax rate of consistency for linear models with missing values, 2022.

A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Naive imputation implicitly regularizes high-dimensional linear models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1320–1340. PMLR, 2023.

M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 2010.

D. S. Bernstein. *Scalar, Vector, and Matrix Mathematics: Theory, Facts, and Formulas - Revised and Expanded Edition*. Princeton University Press, 2018. ISBN 978-1-4008-8825-2. doi: 10.1515/9781400888252.

D. Bertsimas, A. Delarue, and J. Pauphilet. Simple imputation rules for prediction with missing data: Theoretical guarantees vs. empirical performance. *Transactions on Machine Learning Research*, 2024.

A. Bhattacharyya, C. Daskalakis, T. Gouleakis, and Y. Wang. Learning High-dimensional Gaussians from Censored Data, 2025.

B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *ICML*, 2012.

C. Cheng, G. Cheng, and J. C. Duchi. Collaboratively Learning Linear Models with Structured Missing Data. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. arXiv, 2023.

Y. Cherapanamjeri, E. Aras, N. Tripuraneni, M. I. Jordan, N. Flammarion, and P. L. Bartlett. Optimal Robust Linear Regression in Nearly Linear Time, 2020.

C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient Statistics, in High Dimensions, from Truncated Samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018. doi: 10.1109/focs.2018.00067.

C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Computationally and Statistically Efficient Truncated Regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 955–960. PMLR, 2019.

C. Daskalakis, D. Rohatgi, and E. Zampetakis. Truncated Linear Regression in High Dimensions. In *Advances in Neural Information Processing Systems*, volume 33, pages 10338–10347. Curran Associates, Inc., 2020.

C. Daskalakis, P. Stefanou, R. Yao, and E. Zampetakis. Efficient Truncated Linear Regression with Unknown Noise Variance. In *Advances in Neural Information Processing Systems*, volume 34, pages 1952–1963. Curran Associates, Inc., 2021.

I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 1 edition, 2023. doi: 10.1017/9781108943161.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 2019a.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient Algorithms and Lower Bounds for Robust Linear Regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2745–2754. SIAM, 2019b. doi: 10.1137/1.9781611975482.170.

I. Diakonikolas, D. Kane, A. Pensia, and T. Pittas. Near-optimal algorithms for gaussians with huber contamination: Mean estimation and linear regression. *Advances in Neural Information Processing Systems*, 36:43384–43422, 2023.

D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009. ISBN 978-0-521-88427-3. doi: 10.1017/CBO9780511581274.

F. Galton. An Examination into the Registered Speeds of American Trotting Horses, with Remarks on Their Value as Hereditary Data. *Proceedings of the Royal Society of London*, 62:310–315, 1897. ISSN 0370-1662.

J. Hayase, W. Kong, R. Somani, and S. Oh. Spectre: defending against backdoor attacks using robust statistics. In *Proc. 38th International Conference on Machine Learning (ICML)*, 2021.

L. Hu and O. Reingold. Robust Mean Estimation on Highly Incomplete Data with Arbitrary Outliers. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1558–1566. PMLR, 2021.

P. J. Huber. Robust estimation of a location parameter. pages 492–518, 1992.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 1 edition, 2009. doi: 10.1002/9780470434697.

J. Josse, J. M. Chen, N. Prost, G. Varoquaux, and E. Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479, 2024. ISSN 1613-9798. doi: 10.1007/s00362-024-01550-4.

A. R. Klivans, P. K. Kothari, and R. Meka. Efficient Algorithms for Outlier-Robust Regression. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1420–1430. PMLR, 2018.

V. Kontonis, C. Tzamos, and M. Zampetakis. Efficient Truncated Statistics with Unknown Truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019. doi: 10.1109/focs.2019.00093.

C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6): 33:1–33:31, 2013. ISSN 1073-0516. doi: 10.1145/2505057.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic Estimation of Mean and Covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016. doi: 10.1109/focs.2016.76.

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 2008.

S. Li, S. Karmalkar, I. Diakonikolas, and J. Diakonikolas. Learning a Single Neuron Robustly to Distributional Shifts and Adversarial Label Noise. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. arXiv, 2024.

R. J. A. Little. Regression With Missing X's: A Review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992. ISSN 0162-1459. doi: 10.2307/2290664.

R. J. A. Little. Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. ISSN 0162-1459. doi: 10.2307/2290705.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019. ISBN 978-0-470-52679-8.

Z. Liu, J. Park, T. Rekatsinas, and C. Tzamos. On Robust Mean Estimation under Coordinate-level Corruption. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6914–6924. PMLR, 2021.

P-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012. ISSN 0090-5364, 2168-8966. doi: 10.1214/12-AOS1018.

K. Mohan and J. Pearl. Graphical Models for Processing Missing Data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.1874961.

M. L. Morvan, N. Prost, J. Josse, E. Scornet, and G. Varoquaux. Linear predictor on linearly-generated data with missing values: Non consistency and solutions. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 3165–3174. PMLR, 2020a.

M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. NeuMiss networks: Differentiable programming for supervised learning with missing values. In *Advances in Neural Information Processing Systems*, volume 33, pages 5980–5990. Curran Associates, Inc., 2020b.

M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What's a good imputation to predict with missing values? In *Advances in Neural Information Processing Systems*, volume 34, pages 11530–11540. Curran Associates, Inc., 2021.

P. Paschou, J. Lewis, A. Javed, and P. Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics*, 2010.

K. Pearson. On the Systematic Fitting of Curves to Observations and Measurements. *Biometrika*, 1 (3):265–303, 1902. ISSN 0006-3444. doi: 10.2307/2331540.

K. Pearson and A. Lee. On the Generalised Probable Error in Multiple Normal Correlation. *Biometrika*, 6(1):59–68, 1908. ISSN 0006-3444. doi: 10.2307/2331556.

A. Pensia, V. Jog, and P.-L. Loh. Robust Regression with Covariate Filtering: Heavy Tails and Adversarial Contamination. *Journal of the American Statistical Association*, pages 1–12, 2024. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2024.2392906.

K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. 2008.

C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller. Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, pages 153–160. International Educational Data Mining Society, 2013.

S. Roch. *Modern Discrete Probability: An Essential Toolkit*. Cambridge University Press, 1 edition, 2024. doi: 10.1017/9781009305129.

M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5), 2010. ISSN 0090-5364. doi: 10.1214/10-AOS793.

N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 2002.

D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976. ISSN 0006-3444. doi: 10.2307/2335739.

T. Sell, T. B. Berrett, and T. I. Cannings. Nonparametric classification with missing data, 2024.

J. Steinhardt, P. W. Koh, and P. Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

G. Tang, R. J. A. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764, 2003. ISSN 0006-3444. doi: 10.1093/biomet/90.4.747.

B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2006. ISBN 978-0-387-32448-7. doi: 10.1007/0-387-37345-4.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008. ISBN 978-0-387-79052-7.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

J. Vuurens, A. P. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, 2011.

P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Robustly Learning a Single Neuron via Sharpness. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 36541–36577. PMLR, 2023.

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized Out-of-Distribution Detection: A Survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-024-02117-4.

Appendix

The structure of the appendix is as follows: In Appendix A, we provide material omitted from Section 1 of the main body, including the detailed description of upper bounds for our learning task, implications for other corruption models, a detailed summary of related work, and open problems. Appendix B records the notation and mathematical background required for our technical results. Finally, Appendix C and Appendix D give the technical proofs omitted from Section 2 and Section 4 respectively.

Appendix A. Omitted Details from Section 1

A.1. Discussion on Fact 4

We provide a more detailed presentation of the three algorithms for linear-regression under coordinate-wise corruptions with the specific references to prior robust statistics literature below.

Fact 4 (Upper bounds on estimation error) *There are three polynomial-time algorithms $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ that take as input $n = O(\mathrm{poly}(d/\eta))$ samples from the d-dimensional linear regression model with σ-additive noise (Definition 1), with η-fraction of coordinate-wise corruptions according to Definition 3, and output $\widehat{\beta} \in \mathbb{R}^d$ that satisfies the following guarantees with probability at least 0.99:*

- *Guarantee for $\mathcal{A}_1$: $\|\widehat{\beta} - \beta\| = O(\eta d \sigma)$ whenever $\eta d < 0.49$.*
- *Guarantee for $\mathcal{A}_2$: $\|\widehat{\beta} - \beta\| = O(\eta \sqrt{d} \sqrt{\|\beta\|^2 + \sigma^2})$, whenever $\eta < 0.24$.*
- *Guarantee for $\mathcal{A}_3$: $\|\widehat{\beta} - \beta\| \leq \|\beta\|$.*

Proof For the first algorithm, one can use any linear regression algorithm developed in the robust statistics literature, with the guarantee that if an ε-fraction of the samples are corrupted (i.e., contain at least one corrupted coordinate), the algorithm's error is $O(\varepsilon\sigma)$. One such polynomial-time algorithm, which uses $\widetilde{O}(d/\varepsilon^2)$ samples and works for any $\varepsilon \in (0, \varepsilon_0)$ for a sufficiently small absolute constant ε_0, is the one from Diakonikolas et al. (2023). Our coordinate-wise contamination model can be reduced to the corruption model for which that algorithm is designed by using $\varepsilon = \eta(d+1)$.

For the second algorithm, observe that for (X, y) drawn from Definition 1, we have $\mathbb{E}[yX] = \beta$. In particular, for each coordinate j, the expectation $\mathbb{E}[yX_j]$ gives an unbiased estimator for the j-th coordinate of β. Under our corruption model, the samples $\{y_i X_i^{(j)}\}_{i=1}^n$ contain at most a 2η fraction of corruptions: up to η from the $\{X_i^{(j)}\}_{i=1}^n$ and up to η from the labels. Moreover, the distribution of the clean samples has sub-exponential tails and variance bounded by $O(\sqrt{\|\beta\|^2 + \sigma^2})$. For such random variables with an $O(\eta)$-fraction of corruptions, it is known that classical estimators such as trimmed mean (cf. Fact 15), yield an error of $O\left(\eta \log(1/\eta) \sqrt{\|\beta\|^2 + \sigma^2}\right)$ using $\widetilde{O}(d/\eta^2)$ samples. Due to the additional Gaussian structure (i.e., the fact that yX is not an arbitrary sub-exponential variable, but the product of Gaussian random variables), the usual $\log(1/\eta)$ factor can be removed with a more refined analysis (Diakonikolas et al., 2018, Theorem 9) (with sample complexity $O(\mathrm{poly}(d/\eta))$).

Finally, the third algorithm is the trivial one that always outputs the zero vector. ■

A.2. Derivation of Regimes in Table 1

With Fact 4 in hand, we discuss how the parameter space can be partitioned based on which algorithm from Fact 4 achieves the best error (up to absolute constant factors). Due to space constraints in Section 1.1, the regimes were summarized there in a single, somewhat condensed table (Table 1). In this section, we present a more detailed version of this partitioning in Tables 2 to 4 and explicitly explain how the different regimes are derived. Table 1 in Section 1 is a combined version of Tables 2 to 4 into a single table.

Regime for η	$0 \le \eta < \frac{0.49}{d}$	$\frac{0.49}{d} \le \eta < \frac{7}{\sqrt{d}}$	$\frac{7}{\sqrt{d}} \le \eta \le 1$
Best upper bound	see Table 4	see Table 3	$O(\|\beta\|)$ ($\mathcal{A}_3$ from Fact 4)
Best lower bound	see Table 4	see Table 3	$\Omega(\|\beta\|)$ (Theorem 5(b))

Table 2: Estimation bounds for η regimes (C denotes a large constant).

Regime for β	$0 \le \|\beta\| < \eta\sqrt{d}\sigma$	$\eta\sqrt{d}\sigma \le \|\beta\| < \sigma$	$\sigma \le \|\beta\| < \infty$
Best upper bound from Fact 4	$O(\|\beta\|)$ (alg. $\mathcal{A}_3$)	$O(\eta\sqrt{d}\sigma)$ (alg. $\mathcal{A}_2$)	$O(\eta\sqrt{d}\|\beta\|)$ (alg. $\mathcal{A}_2$)
Best lower bound from Theorem 5	$\Omega(\|\beta\|)$ (part (a))	$\Omega(\eta\sqrt{d}\sigma)$ (part (a))	$\Omega(\eta\sqrt{d}\|\beta\|)$ (part (c))*

Table 3: Sub-regimes of the $\frac{0.49}{d} \le \eta < \frac{7}{\sqrt{d}}$ case. *Note: Theorem 5(c) is only for $\eta \in [\frac{2+c}{d}, \frac{7}{\sqrt{d}}]$.

Regime for β	$0 \le \|\beta\| < \eta\sqrt{d}\sigma$	$\eta\sqrt{d}\sigma \le \|\beta\| < \sigma$	$\sigma \le \|\beta\| < \sqrt{d}\sigma$	$\sqrt{d}\sigma \le \|\beta\|$
Best upper bound from Fact 4	$O(\|\beta\|)$ (alg. $\mathcal{A}_3$)	$O(\eta\sqrt{d}\sigma)$ (alg. $\mathcal{A}_2$)	$O(\eta\sqrt{d}\|\beta\|)$ (alg. $\mathcal{A}_2$)	$O(\eta d\sigma)$ (alg. $\mathcal{A}_1$)
Best lower bound from Theorem 5	$\Omega(\|\beta\|)$ (part (a))	$\Omega(\eta\sqrt{d}\sigma)$ (part (a))	$\Omega(\eta\sqrt{d}\|\beta\|)$ (part (d))	$\Omega(\eta d\sigma)$ (part (d))

Table 4: Sub-regimes of the $0 \le \eta < \frac{0.49}{d}$ case.

Regime $7/\sqrt{d} \le \eta \le 1$ Algorithm $\mathcal{A}_1$ from Fact 4 is not applicable in this regime. Among the other two algorithms, $\mathcal{A}_3$ always has the best error (because the error of $\mathcal{A}_2$ is in the order of $\eta\sqrt{d}\sqrt{\|\beta\|^2+\sigma^2} \ge \eta\sqrt{d}\|\beta\| \ge 7\|\beta\|$). This explains the last column in Table 2.

Regime $0.49/d \le \eta < 7/\sqrt{d}$ Again, algorithm $\mathcal{A}_1$ is not applicable in this regime. Between $\mathcal{A}_2$ and $\mathcal{A}_3$, we note that whenever $\|\beta\| \ge \eta\sqrt{d}\sigma$, the error of $\mathcal{A}_2$ is $O(\eta\sqrt{d}\sqrt{\|\beta\|^2+\sigma^2}) = O(\eta\sqrt{d}\|\beta\| + \eta\sqrt{d}\sigma) = O(\|\beta\|)$, i.e., $\mathcal{A}_2$ has the best error (up to constant factors). Moreover, when $\|\beta\| < \sigma$, the term that dominates in the error of $\mathcal{A}_2$ is $\eta\sqrt{d}\sigma$ and when $\|\beta\| \ge \sigma$ the dominating $\mathcal{A}_2$ error term is $\eta\sqrt{d}\|\beta\|$. This explains the regimes shown in Table 3.

Regime $0 \le \eta < 0.49/d$ In this regime, all three algorithms are applicable. For the same reason as before, the error of $\mathcal{A}_3$ is better than that of $\mathcal{A}_2$ when $\|\beta\| < \eta\sqrt{d}\sigma$ and $\mathcal{A}_2$ is better than $\mathcal{A}_3$ otherwise. It also outperforms $\mathcal{A}_1$, whose error is on the order of $\eta d\sigma$. In the sub-regime

$\eta\sqrt{d}\sigma \le \|\beta\| < \sqrt{d}\sigma$, the error of $\mathcal{A}_2$ is at most $O(\eta\sqrt{d}\sigma + \eta\sqrt{d}\|\beta\|)$, with both terms being smaller than $O(\eta d\sigma)$. Thus, $\mathcal{A}_2$ achieves the best error in this sub-regime. In the final sub-regime $\|\beta\| \ge \sqrt{d}\sigma$, the error of $\mathcal{A}_1$ is of the order $\eta d\sigma$, which is smaller than that of $\mathcal{A}_3$ (which is of the order $\|\beta\| \ge \sqrt{d}\sigma \ge \sigma \ge \eta d\sigma$). Comparing with $\mathcal{A}_2$, the dominating term in the error of $\mathcal{A}_2$ is the $\eta\sqrt{d}\|\beta\|$ term which is larger than $\eta d\sigma$ whenever $\|\beta\| > \sqrt{d}\sigma$. Thus $\mathcal{A}_1$ has the best error (up to a constant factor) in the sub-regime $\|\beta\| > \sqrt{d}\sigma$. This explains Table 4.

A.3. Unified Algorithm with Error Guarantees as in Tables 2 to 4

We further claim that there exists a single algorithm that can automatically adapt to the best choice among $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ from Fact 4, and achieve the error upper bounds in Tables 2 to 4, without requiring a priori knowledge of the regime in which $\|\beta\|$ lies. The algorithm only needs to know η (which could also be avoided by applying standard techniques like Lepski's method, but is beyond the scope of this work).

Theorem 14 *There is a polynomial time algorithm $\mathcal{A}$ that takes as input a parameter $\eta \in (0,1)$ and $n = O(\mathrm{poly}(d/\eta))$ samples from the d-dimensional linear regression model (Definition 1) after η-fraction of coordinatewise corruptions according to Definition 3, and returns an estimate $\widehat{\beta}$ of β that satisfies the error upper bounds from Table 1 (equivalently Tables 2 to 4) with high constant probability.*

Proof The idea for the meta-algorithm is to first use the knowledge of η to restrict our attention to the relevant table among Tables 2 to 4 and then estimate the error bound of the three algorithms $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ (up to absolute constant factors) in order to return the output of the algorithm corresponding to the smallest error bound. We show how this can be done for the regime $0 < \eta \ll 1/d$; the other regimes can be handled by the same arguments.

Suppose that $\eta \ll 1/d$ and let $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ be the outputs of $\mathcal{A}_1, \mathcal{A}_2$ and $\mathcal{A}_3$ respectively. We first show how to estimate (within a multiplicative absolute constant) the upper bound on the error of $\mathcal{A}_3$ from Table 4 (i.e., the quantity $\eta d\sigma$). Since η and d are known, it suffices to estimate σ. Note that the expectation of the squared labels y^2 under the clean distribution of Definition 1 is $\mathbb{E}[y^2] = \|\beta\|^2 + \sigma^2$. The procedure to estimate σ within a constant factor is then the following: We compute the residuals $y' = y - \widehat{\beta}_1^\top X$ for all samples in the dataset. This transformation makes the clean samples as if they came from the model of Definition 1 with true regressor $\beta - \widehat{\beta}_1$ (instead of β that we started with). The expectation of the new squared labels is now $\mathbb{E}[(y')^2] = \|\beta - \widehat{\beta}_1\|^2 + \sigma^2 = \sigma^2(1 + O(\eta d)) = O(\sigma^2)$. We can use an outlier robust one-dimensional mean estimator to find a u such that $|u - \sigma^2| \le 0.01\sigma^2$. This can be done using Fact 15 below (where $\|(y')^2\|_{\psi_1}$ in our case is at most $O(\sigma^2)$):

Fact 15 (Univariate Trimmed Mean, see, e.g., Diakonikolas and Kane (2023)) *Let ε_0 be a sufficiently small absolute constant. There is an algorithm (trimmed mean) that, for every $\varepsilon \in (0, \varepsilon_0)$ and a univariate distribution D that has ψ_1-norm at most σ^2 (cf. Definition 16), given a set of $n \gg \log(1/\delta)/(\varepsilon^2 \log^2(1/\varepsilon))$ samples from D with corruption at rate ε, outputs a $\widehat{\mu}$ such that $|\widehat{\mu} - \mathbb{E}_{X\sim D}[X]| \lesssim \sigma^2 \varepsilon \log(1/\varepsilon)$ with probability at least $1 - \delta$.*

Definition 16 (Sub-exponential Random Variables) *We call $\|Y\|_{\psi_1} := \sup_{p \ge 1} p^{-1} \mathbb{E}[|Y|^p]$ the sub-exponential norm of the random variable Y.*

By using the above, so far we have shown how to obtain an estimate e_1 such that $|e_1 - \eta d\sigma| \leq 0.1\eta d\sigma$ (i.e., e_1 is a multiplicative approximation to the error bound of $\mathcal{A}_1$ listed in Table 4). By using Fact 15 again we can similarly obtain an estimate for the error bound of $\mathcal{A}_2$ from the table, i.e., e_2 that satisfies $|e_2 - \eta\sqrt{d}\sqrt{\|\beta\|^2 + \sigma^2}| \leq 0.1\eta\sqrt{d}\sqrt{\|\beta\|^2 + \sigma^2}$. Let C, C', C'' be sufficiently large absolute constants with $C \ll C' \ll C''$. The meta-algorithm that chooses which of the outputs $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ to return is the following. If $Ce_1 < e_2$ the meta-algorithm returns $\widehat{\beta}_1$, otherwise it does the following: if $\|\widehat{\beta}_2\| > C'e_2$ the meta-algorithm returns $\widehat{\beta}_2$, otherwise it returns $\widehat{\beta}_3$. The correctness follows easily: In the regime $\|\beta\| \gg \sqrt{d}\sigma$ where $\mathcal{A}_1$ has the best error, the first check yields $e_2 = \Omega(\eta\sqrt{d}\sqrt{\sigma^2 + \|\beta\|^2}) = \Omega(\eta\sqrt{d}\|\beta\|) \gg e_1$ thus the meta-algorithm will indeed return $\widehat{\beta}_1$. If it is the case that that we are in the regime where $\mathcal{A}_3$ has the best error, i.e., $\|\beta\| < Ce_2$ then we can easily see that the meta-algorithm will return $\widehat{\beta}_3$. This is because the first check will yield $e_2 \leq Ce_1$ (since $e_2 = O(\eta\sqrt{d}\sqrt{\sigma^2 + \|\beta\|^2}) = O(\eta\sqrt{d}\sigma) = O(\eta d\sigma) \leq Ce_1$) and the second check will yield $\|\widehat{\beta}_2\| \leq \|\beta\| + \|\beta - \widehat{\beta}_2\| \leq Ce_2 + O(\eta\sqrt{d}\sqrt{\|\beta\|^2 + \sigma^2}) \leq C'e_2$. If we are in the regime where $\mathcal{A}_2$ has the best error, i.e., $\|\beta\| > C''e_2$ the first check will yield $e_2 \leq Ce_1$ (since $e_2 = O(\eta\sqrt{d}\sqrt{\sigma^2 + \|\beta\|^2}) = O(\eta d\sigma) \leq Ce_1$) and the second check will yield $\|\widehat{\beta}_2\| \geq \|\beta\| - \|\beta - \widehat{\beta}_2\| \geq C''e_2 - O(e_2) > C'e_2$ thus the meta-algorithm will indeed yield $\widehat{\beta}_2$. Note that we left out the regime $C'e_2 \leq \|\beta\| \leq C''e_2$. In that case, it does not matter which of $\widehat{\beta}_2, \widehat{\beta}_3$ we output because their errors are within a constant factor from each other.

■

A.4. Other Implications of Our Results

In this brief subsection, we discuss some immediate implications of our results.

Implications on Models with a Restricted Set of Corrupted Coordinates In this paper, we primarily discuss the case in which every coordinate is subject to an η fraction of corruptions. However, another reasonable model would be to consider the case where only a subset $\mathcal{S} \subset \{1, 2, \ldots, d\}$ of $|\mathcal{S}| = r \leq d$ coordinates is subject to adversarial erasures as in Definition 2 or more general corruptions as in Definition 3. It is straightforward to verify that all our results generalize to this model, with estimation error bounds now depending on r in place of d and on $\|\beta_{\mathcal{S}}\|_2$ in place of $\|\beta\|_2$, where $\beta_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \beta_i e_i$ and e_i denotes the i^{th} standard basis vector.

Implications for Generalized Linear Models While our primary focus in this work is on lower and upper bounds for the attainable estimation error $\|\beta - \widehat{\beta}\|_2$, some of our lower bounds have broader implications on the attainable squared error of generalized linear models (GLMs)—where the "ground truth" labels follow $f(\beta^\top X)$ for some possibly more general function f called the activation, instead of $\beta^\top X$ corresponding to the case of linear regression (where $f(t) = t$ is the identity function). In this case, the focus is usually on bounding the mean squared error for a predictor $\widehat{\beta}$, defined by $\mathcal{L}(\widehat{\beta}) = \mathbb{E}_{(X,y)\sim\mathcal{D}}[(y - f(\widehat{\beta}^\top X))^2]$ for (X, y) jointly distributed according to a distribution $\mathcal{D}$. It is immediate that in the case of linear regression model from Definition 1, we have

$$\begin{aligned}\mathcal{L}(\widehat{\beta}) &= \sigma^2 + \mathop{\mathbb{E}}_{X\sim\mathcal{N}}[(\beta^\top X - \widehat{\beta}^\top X)^2] \\ &= \sigma^2 + \mathop{\mathbb{E}}_{X\sim\mathcal{N}}[(\beta - \widehat{\beta})^\top X X^\top (\beta - \widehat{\beta})]\end{aligned}$$

$$= \sigma^2 + \|\beta - \widehat{\beta}\|_2^2.$$

Thus, our lower bounds on $\|\beta - \widehat{\beta}\|_2$ directly imply lower bounds on the mean squared error for linear regression, using the above derivation.

Concerning GLMs in broader generality, let us consider the realizable case, where all labeled examples (X, y) satisfy $y = f(\beta^\top X)$ for some fixed vector β and activation function f. Notice first that the lower bounds in Theorem 5(b) and Theorem 5(c) apply even when $\sigma = 0$.[5] Our lower bounds for linear regression are proved by constructing a coupling between the distribution of X conditioned on the value of $y = t$ under two distinct hypotheses corresponding to sufficiently different prediction/weight vectors β and β'. Since for $\sigma = 0$ in the case of linear regression this is equivalent to conditioning on the value of $\beta^\top X = (\beta')^\top X = t$, the same lower bounds (implying impossibility of distinguishing between β and β') generalize to the case where $y = f(t)$. This is because considering more general functions f that are not identity (as in the case of linear regression) can only further obscure information if f is not invertible (as is the case, for example, for ReLU activations where $f(t) = \max\{0, t\}$).

The above discussion implies that the same lower bounds as in Theorem 5(b) and Theorem 5(c) apply for the value of $\|\widehat{\beta} - \beta\|_2$, where β is the "ground truth" predictor and $\widehat{\beta}$ the predictor that can be constructed by an algorithm. A consequence is that the mean squared error can be bounded below as a function of $\|\widehat{\beta} - \beta\|_2^2$ for a broad class of GLMs. In particular, for the class of (a, b)-unbounded activations—namely, monotonically non-decreasing functions f that are b-Lipschitz, satisfy $f(0) = 0$, and are such that $f'(t) \geq a > 0$ for $t > 0$—known results (cf. Li et al. (2024); Wang et al. (2023)) imply that for X distributed according to the standard normal distribution, we have $\mathcal{L}(\widehat{\beta}) = \Omega(\mathrm{poly}(b/a))\|\widehat{\beta} - \beta\|_2^2$. In particular, since ReLU activation is (a, b)-unbounded with $a = b = 1$, we have that for the ReLU activation, $\mathcal{L}(\widehat{\beta}) = C\|\beta - \widehat{\beta}\|_2^2$ for a universal constant $C > 0$. As a consequence, the mean squared error is of the order $\Omega(\|\beta\|_2^2)$ for $\eta \in [7/\sqrt{d}, 1]$ and of the order $\Omega(c^2\eta^2 d\|\beta\|_2^2)$ for $\eta \in [(2 + c)/d, 7/\sqrt{d}]$. In other words, unless both η and $\|\beta\|_2$ are small, no meaningful prediction—as measured by the mean squared error—is possible.

These observations reinforce the argument that linear regression is the most basic model to study when considering limitations imposed by adversarial coordinate-wise corruptions studied in our work.

A.5. Related Work

There is a vast literature on learning with missing data in various forms. To the best of our knowledge, this work is the first to address the fully adversarial missingness model of Definition 2 in the context of high-dimensional regression. The most relevant prior work is Liu et al. (2021), which considers fully adversarial missingness, but for the simpler task of mean estimation. Hu and Reingold (2021) also considers mean estimation and uses a milder setting for the missingness model. We provide a discussion of these works below, followed by a broader (though not exhaustive) review of related literature, with an emphasis on settings most closely aligned with ours.

Comparison with Liu et al. (2021) This work studies the optimal error of Gaussian mean estimation in the same coordinate-wise adversary of Definitions 2 and 3 (among others). Concretely, it considers a slightly more general problem than the vanilla Gaussian mean estimation where data

5. While constructions of other lower bounds (from Theorem 5(a),(d)) may also have implications on the mean squared error of GLMs for $\sigma \neq 0$, here we focus on Theorem 5(b),(c) as the implications are more direct.

has some additional low-dimensional structure: each sample x_i follows $\mathcal{N}(\mu, \Sigma)$ but can be written as $x_i = Az_i$, where $A \in \mathbb{R}^{d\times r}$ and $z \in \mathbb{R}^r$ is some lower dimensional Gaussian. The goal is to estimate μ. Their lower bounds on the estimation error are based on the approach of constructing couplings with small coordinate-wise disagreements (as in the present work). In terms of algorithms, the approach of Liu et al. (2021) for the missing data of Definition 2 is a two-step algorithm that first estimates the missing values and then runs existing estimators on the imputed dataset. They show that the first step becomes NP-hard when data is replaced (Definition 3) instead of missing. However, they are able to get some preliminary results by a randomized algorithm for special cases (like when A is known). In the most general case for the mean estimation problem of Liu et al. (2021), the answer to the question of whether missing data are easier to handle than replaced data remains unclear. In the case of the linear regression problem considered in this paper, we find that the answer is negative. On the technical side, our work builds on the coupling technique of Liu et al. (2021), but significantly extends it to handle new challenges in the linear regression setting. The main difficulty is that the estimation error behaves differently across many distinct regimes, each requiring a separate analysis. As discussed in Section 1.2, a direct application of the Liu et al. (2021) coupling fails when the additive noise σ is small, leaving several regimes unaddressed. We resolve this by coupling only the first $d-1$ coordinates and treating the last one as additive noise. This regime also requires a redesigned hypothesis testing setup. The original test inspired from Liu et al. (2021) that compares $\beta = (1, \ldots, 1)$ and $\beta' = (-1, \ldots, -1)$, is tailored to show an $\Omega(\|\beta\|)$ lower bound, i.e., that no non-trivial estimation is possible. In our case, however, some non-trivial estimation is possible, so we design a finer test: we split the coordinates into two halves with differing values in β and β' and apply separate couplings to each half. Finally, one particularly challenging regime remains, which we address by further splitting both the coordinates and the labels into two parts, and using a more delicate coupling between the parts of covariates and labels.

Comparison with Hu and Reingold (2021) This work also studies mean estimation under a missing data model. Their setting involves a combination of two types of erasures: up to an η-fraction of coordinates may be missing, and up to an ε-fraction of entire samples may be deleted. The second type of deletion (entire samples) can be fully adversarial, while the first type (coordinate-wise) is less adversarial than in our model, because in Hu and Reingold (2021) the adversary must commit to the missingness pattern before seeing the data. The algorithms proposed in Hu and Reingold (2021) are based on imputing missing values, and thus do not naturally extend to the contamination setting, where adversaries may replace values rather than erase them.

Literature on Missing Data The study of missing data has a long history, motivated by concerns like those discussed in Section 1. In a seminal work, Rubin (1976) distinguishes three types of missingness: when missingness is independent of the data, it is called missing completely at random (MCAR); when it depends only on the observed data, it is missing at random (MAR); otherwise, it is missing not at random (MNAR), where missingness depends on both observed and unobserved values. See also Tsiatis (2006); Tang et al. (2003); Little and Rubin (2019); Mohan and Pearl (2021) for a more in-depth review.

Much of the literature on regression with missing data has focused on parameter estimation. Early works include Little (1992, 1993), as well as Rosenbaum and Tsybakov (2010), which proposes a sparse variant of the Lasso estimator. More recent approaches revisit the problem from the perspectives of imputation and collaborative learning (Aladin et al., 2020; Cheng et al., 2023).

Other recent efforts have shifted focus to the label prediction problem, which is a fundamentally different goal, as the test set is also expected to contain missing entries. Notably, the Bayes optimal predictor in this setting decomposes into a sum over predictors for each missingness pattern, and the number of such patterns can be exponential. This pattern-specific structure has been studied in Morvan et al. (2020a,b); Ayme et al. (2022), which characterize minimax-optimal rates under MAR and MCAR (and in Sell et al. (2024) for classification). A body of work also focuses on the idea of using a two-step procedure: first impute the missing entries, then apply algorithms designed for complete data. This method has been shown to be Bayes optimal in various settings (Josse et al., 2024; Bertsimas et al., 2024; Morvan et al., 2021).

In light of the above, our work fits into the MNAR setting, but to the best of our knowledge, is not directly related to any of the aforementioned works. Unlike the prediction-focused literature (which is more challenging and often incurs exponential complexity) we study parameter estimation. Moreover, prior work on parameter estimation for linear regression either assumes MAR (Loh and Wainwright, 2012; Cheng et al., 2023) or considers much more benign forms of MNAR; for instance, Aladin et al. (2020) assumes that the missingness pattern is drawn i.i.d. from an arbitrary distribution.

Truncated Statistics Although technically this literature falls into the Missing Not At Random category from the previous paragraph, it is sufficiently well developed to have its own place in the literature. Truncated statistics refer to the situation where samples falling outside of a fixed set are censored. Parameter estimation under this setting traces back to the early work of Galton, Pearson and Lee (Galton, 1897; Pearson, 1902; Pearson and Lee, 1908). Despite early interest in the problem, computationally efficient algorithms for multivariate Gaussians were developed relatively recently, with some notable works being Daskalakis et al. (2018); Kontonis et al. (2019) for mean estimation and Daskalakis et al. (2019, 2020, 2021) for linear regression. While the previous works operate under a setting where each sample is either entirely censored or entirely visible, Bhattacharyya et al. (2025) extends the setting to partially truncated data, which each coordinate individually visible or censored (for the problem of Gaussian mean estimation). That said, the setting considered in Bhattacharyya et al. (2025) is still milder than ours. It considers either (i) a setting where each coordinate is erased based on membership in a fixed, predetermined set (to which the algorithm has oracle access), or (ii) erasures based on a simple projection rule. Under this model, consistent estimation is possible, as shown in Bhattacharyya et al. (2025), whereas in our fully adversarial erasure setting, it is provably not.

Robust Statistics Robust Statistics was developed in the 1960s (Tukey, 1960; Huber, 1992) to handle scenarios where a small fraction of the data points are arbitrarily corrupted. As in the setting of this work, such corruptions do not allow for consistent estimation. Early work focused on characterizing the information-theoretic error for univariate Gaussian mean estimation. Since then, a large body of work developed estimators for various tasks (Huber and Ronchetti, 2009). However, these estimators were computationally inefficient for high-dimensional tasks. The field saw a resurgence with the work of Lai et al. (2016); Diakonikolas et al. (2019a) that tackled high-dimensional robust mean estimation in polynomial time. In the last decade, a plethora of robust estimators have been introduced for different tasks, including linear regression (Klivans et al., 2018; Diakonikolas et al., 2019b; Pensia et al., 2024; Cherapanamjeri et al., 2020) considered in this work. For a comprehensive treatment, see the book Diakonikolas and Kane (2023). The key difference is that the de facto contamination models in robust statistics treat each sample as either fully clean or fully corrupted, making non-trivial estimation impossible as the corruption rate approaches $1/2$.

This fails to capture cases where each coordinate is only mildly corrupted, i.e., few samples are corrupted per coordinate, even if most samples are affected overall. Such settings may still allow for meaningful estimation, but algorithms from the robust statistics literature are inapplicable.

A.6. Open Problems

While our work addresses the estimation error for the standard model of linear regression with coordinate-wise corruptions by providing matching upper and lower bounds for essentially all possible parameter regimes, there are several avenues for future research that merit further investigation.

First, it would be interesting to consider $X \sim \mathcal{N}(0, \Sigma)$ in the linear regression problem, with $\Sigma \neq I$ and unknown to the algorithm in the linear regression model. Does the conclusion that dealing with missing data is no easier than handling adversarially replaced data still hold in this setting?

Second, although the algorithms are efficient in terms of runtime (i.e., polynomial time) and sample complexity (using $\widetilde{O}(d/\eta^2)$ samples), the optimal sample complexity for the problem remains unclear. In the clean setting (where all samples follow the distribution of Definition 1), the optimal sample complexity is d/u^2, where u is the target estimation error. In some regimes of Tables 2 to 4, the estimation error scales with $\|\beta\|$, which can be very large. This implies that (at least in the noiseless setting) fewer samples are needed to achieve that level of error.

Third, as discussed in Appendix A.4, some of our lower bounds directly imply lower bounds for the mean squared error of a broad class of generalized linear models that includes ReLU as a special case, even with perfect, noise-free labels (i.e., in the realizable case). It is however unclear if those lower bounds are tight or if they can be strengthened further, since non-invertible activations like ReLU can further obscure information available to the algorithm (because the label is zero whenever the argument $\beta^\top X$ is negative). Relatedly, it would be interesting to develop polynomial-time algorithms that can match the information-theoretic lower bounds (or prove no such algorithms exist, even when restricted to classes such as Statistical Query algorithms).

Finally, our results demonstrate that adversarial coordinate-wise deletions make both estimation and prediction challenging, even in the most basic model of Gaussian linear regression and with infinite samples. On the other hand, a fully random model of deletions (where deletions are independent of the observed data) allows for the diminishing estimation error as the sample size is increased. These two extremes beg the question of what lies in between, when the deletions are neither fully adversarial nor fully random.

Appendix B. Preliminaries

We provide the full version of the preliminaries here.

B.1. Notation

Basic Probability Notation We write $X \sim D$ to denote a random variable X that is distributed according to the distribution D. We use $P_X(x)$ to denote the pdf of X. For multiple variables, we use $(X, Y) \sim D$ to denote that X and Y are jointly distributed according to D, and write $P_{X,Y}(x, y)$ for the pdf of that joint distribution. If $(X, Y) \sim D$ we will also use the notation $X|Y = y$ to denote the random variable distributed according to the *conditional* distribution of X given the occurrence of the value y for Y, i.e., the distribution whose pdf is the conditional density function $P_{X|Y}(x|y) = P_{X,Y}(x, y)/P_Y(y)$, and we denote by X the random variable distributed according to the *marginal*

distribution of D, i.e., the distribution with pdf $P_X(x) = \int P_{X,Y}(x,y)\mathrm{d}y$. We use $\mathbb{E}_{X\sim D}[X]$ for the expectation of X. If $X \sim P_X$ and $Y \sim P_Y$ are continuous random variables over some domain $\mathcal{X}$, with pdfs $P_X(x)$ and $P_Y(y)$, we denote by $D_{\mathrm{TV}}(X,Y) = \frac{1}{2}\int_{x\in\mathcal{X}} |P_X(x) - P_Y(x)|\mathrm{d}x$ the total-variation distance between the distributions P_X and P_Y (by slightly abusing notation, $D_{\mathrm{TV}}(X,Y)$ may some times be also denoted as $D_{\mathrm{TV}}(P_X, P_Y)$). We denote the Kullback–Leibler (KL) divergence between distributions $X \sim P_X$ and $Y \sim P_Y$ by $D_{\mathrm{KL}}(X \,\|\, Y) = \int_{x\in\mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{P_Y(x)}\right)\mathrm{d}x$, assuming P_X is absolutely continuous with respect to P_Y.

Basic Notation We use $\mathbb{Z}_+$ for the set of positive integers. We denote $[n] = \{1,\ldots,n\}$. For a vector x we denote by $\|x\|$ its Euclidean norm. Let I_d denote the $d \times d$ identity matrix (omitting the subscript when it is clear from the context). We use $\mathbb{1}_d$ for the all-ones vector in $\mathbb{R}^d$. We use $\top$ for the transpose of matrices and vectors. We use $|A|$ to denote the determinant of matrix A. We use $a \lesssim b$ to denote that there exists an absolute universal constant $C > 0$ (independent of the variables or parameters on which a and b depend) such that $a \le Cb$. In our notation $a = O(b)$ has the same meaning as $a \lesssim b$ (similarly for $\Omega(\cdot)$ notation) We use $\widetilde{O}$ and $\widetilde{\Omega}$ to hide polylogarithmic factors.

Couplings If P and Q are probability distributions over $\mathcal{X}$, then a coupling Π of P and Q is any distribution over $\mathcal{X}\times\mathcal{X}$ such that the marginals of Π coincide with P and Q. We write $(X,Y) \sim \Pi$ to denote random variables that are distributed according to the coupling; X is marginally distributed according to P and Y according to Q.

B.2. Useful Probability and Linear Algebraic Facts

We start with some facts about Gaussian distributions, namely the inner product of a Gaussian and a fixed vector is another Gaussian, and the distribution conditioned on one of the variables is also Gaussian.

Fact 17 (see, e.g., Petersen and Pedersen (2008)) *If $X \sim \mathcal{N}(\mu,\Sigma)$ is a multivariate Gaussian vector in $\mathbb{R}^d$, and $u \in \mathbb{R}^d$ is another fixed (deterministic) vector, then $u^\top X \sim \mathcal{N}(u^\top\mu, u^\top\Sigma u)$.*

Fact 6 (see, e.g., Section 8.1.3. in Petersen and Pedersen (2008)) *If $X \sim \mathcal{N}(\mu, I)$ is a Gaussian vector in $\mathbb{R}^d$, $\xi \sim \mathcal{N}(0,\sigma^2)$ is a univariate Gaussian, and $u \in \mathbb{R}^d$ is a fixed vector, then the distribution of X conditioned on $u^\top X + \xi = r$ is the Gaussian $\mathcal{N}\left(ru/(\|u\|^2+\sigma^2),\ I - uu^\top/(\|u\|^2+\sigma^2)\right)$.*

The following fact provides a closed-form formula for the KL-divergence between two multivariate Gaussians. The formula can be derived by direct computation and using properties from Section 8.2 of Petersen and Pedersen (2008).

Fact 18 (KL-divergence between multivariate Gaussians) *The KL-divergence between $\mathcal{N}(\mu_1,\Sigma_1)$ and $\mathcal{N}(\mu_2,\Sigma_2)$ is*

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mu_1,\Sigma_1)\,\|\,\mathcal{N}(\mu_2,\Sigma_2)\right) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1)\right).$$

We will bound total variation distances using the previous fact combined with Pinsker's inequality, stated below.

Fact 19 (Pinsker's Inequality (see, e.g., Tsybakov (2008))) *Let P and Q be two probability distributions over the same measurable space. Then,*

$$D_{\mathrm{TV}}(P,Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \,\|\, Q)}.$$

Combining the above two facts we obtain the following bound on the TV distance between univariate Gaussians.

Fact 8 (Total variation between univariate Gaussians (see, e.g., Petersen and Pedersen (2008)) *If $D_1 = \mathcal{N}(\mu_1, \sigma^2)$ and $D_2 = \mathcal{N}(\mu_1, \sigma^2)$ then $D_{\mathrm{TV}}(D_1, D_2) \leq (1/\sqrt{2})|\mu_1 - \mu_2|/\sigma$.*

The following fact states that there exists a coupling between two distributions such that the probability of disagreement is at most their total variation distance.

Fact 7 (Maximal coupling (see, e.g., Roch (2024))) *Let P and Q be distributions. There exists a coupling Π between P and Q such that $\mathbb{P}_{(X,Y)\sim\Pi}[X \neq Y] = D_{\mathrm{TV}}(P,Q)$. Moreover for distributions on P, Q on $\mathbb{R}$ where P is a shifted version of Q, if μ_P, μ_Q denote their expectations, Π additionally satisfies $\mathbb{E}_{(X,Y)\sim\Pi}[|X - Y|] = |\mu_P - \mu_Q|$.*

We also require the multiplicative version of the Chernoff-Hoeffding bound for binary random variables:

Fact 20 (Chernoff-Hoeffding Bound (see, e.g., Dubhashi and Panconesi (2009))) *Let $g_1, \ldots, g_n$ be random variables in $\{0,1\}$ such that $\mathbb{E}[g_i] = p$ for all $i \in [n]$. Then, for all $\varepsilon \in (0,1)$ the following holds:*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} g_i > p(1+\varepsilon)\right] \leq \exp\left(-\frac{\varepsilon^2}{3} pn\right).$$

Finally, the following linear algebraic fact provides a useful formula for inverting matrices of the form identity minus a rank-one matrix.

Fact 21 (Sherman-Morrison formula (see, e.g., Fact 3.21.3 in Bernstein (2018))) *The matrix $\Sigma = I + uv^\top$ is invertible if and only if $1 + v^\top u \neq 0$. In this case, $\Sigma^{-1} = I - \frac{uv^\top}{1+v^\top u}$*

Appendix C. Omitted Details from Section 2

C.1. Couplings with Small Coordinate-Wise Disagreements

We restate and prove the following statements.

Lemma 9 *Let D, D' be distributions over labeled examples (X, y) with $X \in \mathbb{R}^d$, $y \in \mathbb{R}$. Assume that for any two permutations $\pi_1, \pi_2 \in S_{d/2}$, the distribution of $(X, y) \sim D$ is the same as that of $(\pi_1(X_1), \pi_2(X_2), y)$, where X_1 and X_2 denote the first and last $d/2$ coordinates of X, respectively. Assume that the same property also holds for D'. Consider the hypothesis testing problem where the null hypothesis is that data are drawn from D under the corruption model of Definition 2, and*

the alternative hypothesis is that data are drawn from D'. If there exists a coupling Π between D, D' such that for some $c \in (0,1)$: $\mathbb{P}_{((X,y),(X',y'))\sim\Pi}[y \neq y'] \leq \frac{\eta}{2}(1-c)$ and

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] \leq \tfrac{\eta}{2} d(1-c), \tag{1}$$

then no test can distinguish D from D' with probability greater than $\frac{1}{2} + \frac{(d+1)}{2} e^{-\Omega(c^2\eta n)}$.

Proof Given a coupling that satisfies Equation (1), we describe below a procedure that generates samples for each hypothesis and uses a simple adversary to edit the samples so that the resulting data set is (with high probability) the same regardless of the hypothesis in effect. The procedure consists of simply drawing paired samples from the coupling and the adversary erases all coordinates where the samples differ.

1. Let Π denote a coupling that satisfies Equation (1).
2. Initialize empty sets $S \leftarrow \emptyset$, $S' \leftarrow \emptyset$.
3. Initialize corruption budgets $r_1 \leftarrow \eta\, n, \ldots, r_{d+1} \leftarrow \eta\, n$ for each of the d coordinates as well as the labels.
4. For $i = 1, 2, \ldots, n$ do:
   (a) Draw $((X, y), (X', y')) \sim \Pi$.
   (b) For every $j = 1, 2, \ldots, d+1$
      i. If $j \leq d$ and $X_j \neq X_j'$ and $r_j > 0$:
         A. $X_j \leftarrow \perp$, and $X_j' \leftarrow \perp$.
         B. Update $r_j \leftarrow r_j - 1$.
      ii. If $j = d$ and $y \neq y'$ and $r_j > 0$:
         A. $y \leftarrow \perp$, and $y' \leftarrow \perp$.
         B. Update $r_j \leftarrow r_j - 1$.
      iii. $S \leftarrow S \cup \{(X, y)\}$
      iv. $S' \leftarrow S' \cup \{(X', y')\}$

Each time the adversary above deletes the j-th coordinate (lines 4(b)iA, 4(b)iiA) it reduces the budget r_j by 1. If the budget reaches zero, the adversary can no longer keep deleting that coordinate. Importantly, if $\mathcal{E}$ denotes the event that none of the r_j's for $j = 1, 2 \ldots, d+1$ reach zero, the dataset S at the end is the same regardless of the hypothesis that is under effect. This means that under that event, no algorithm can distinguish between the two hypotheses. It remains to show that the probability of the event $\mathcal{E}$ is at least $1 - (d+1)e^{-\Omega(c^2 n\eta)}$. For simplicity let us first focus on the corruptions of covariates only (and will discuss label corruptions at the end). That is, let $\mathcal{E}_1$ denote the event that none of the r_j's for $j = 1, 2 \ldots, d$ reach zero, and we will show that this happens with probability at least $1 - de^{-\Omega(c^2 n\eta)}$. A similar argument will work for the labels thus we only focus on showing that the dataset restricted to the covariates becomes indistinguishable.

Let $g_{i,j} \in \{0,1\}$ be 1 if and only if lines 4(b)iA, 4(b)iiA caused a deletion of the j-th coordinate of the i-th sample during the process described above. By assumption we have that for every $i \in [n]$,

$$\mathbb{E}\left[\sum_{j=1}^{d} g_{i,j}\right] \leq \frac{\eta}{2} d(1-c).$$

By the assumption in the lemma statement, we can assume that the first half of the coordinates undergo a random permutation and that the second half of the coordinates also undergo another random permutation. If we denote by $g'_{i,j} \in \{0,1\}$ the random variable that is 1 if and only if the i-th sample has its j-th coordinate corrupted after the aforementioned two random permutations, then we have the following for every $i \in [n]$ and $j \in [d/2]$ (i.e., we are only analyzing the first half of coordinates for now as the analysis is the same for the second half):

$$\begin{aligned}\mathbb{E}[g'_{i,j}] = \mathbb{P}[g'_{i,j} = 1] &= \sum_{k=0}^{d} \mathbb{P}\left[g'_{i,j} = 1 \middle| \sum_{j=1}^{d} g_{i,j} = k\right] \mathbb{P}\left[\sum_{j=1}^{d} g_{i,j} = k\right] \\ &\leq \sum_{k=0}^{d} \frac{k}{d/2} \mathbb{P}\left[\sum_{j=1}^{d} g_{i,j} = k\right] = \frac{2}{d}\mathbb{E}\left[\sum_{j=1}^{d} g_{i,j}\right] \leq \eta(1-c).\end{aligned}$$

where the second line used that in the worst case where the k disagreements all happen within the first half of the coordinates, after randomly permuting these coordinates, the probability of our fixed coordinate j to experience a disagreement is $\frac{k}{d/2}$. This means that fixing a coordinate $j \in [d/2]$, the number of corruptions in that coordinate across all n samples, $\sum_{i=1}^{n} g'_{i,j}$, is a sum of independent binary variables with expectation $\eta(1-c)$ each. By Chernoff-Hoeffding bounds (Fact 20), we obtain

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} g'_{i,j} > \eta\right] \leq e^{-(1-c)\eta n\left(\frac{1}{1-c}-1\right)^2/3} = e^{-\Omega(c^2 n\eta)}.$$

The same analysis applies for coordinates j in the second half of the coordinates. By a union bound, the probability that there exists a coordinate $j \in [d]$ with $\frac{1}{n}\sum_{i=1}^{n} g'_{i,j} > \eta$ is at most $de^{-\Omega(c^2 n\eta)}$. This means that the event $\mathcal{E}_1$ defined earlier has probability at least $1 - de^{-\Omega(c^2 n\eta)}$. Finally, we can define a similar event $\mathcal{E}_2$ for the label corruptions, i.e., $\mathcal{E}_2$ being the even that r_{d+1} does not reach zero. With another application of the Chernoff bound we can also conclude that $\mathcal{E}'$ happens with probability at least $1 - e^{-\Omega(c^2 n\eta)}$. Combining with a union bound, the probability of both $\mathcal{E}$ and $\mathcal{E}_2$ happening is at most $1 - (d+1)e^{-\Omega(c^2 n\eta)}$. This means that the distributions of the sets S, S' output by the pseudocode have total variation distance at most $1 - (d+1)e^{-\Omega(c^2 n\eta)}$. Consequently, by Le Cam's inequality (LeCam, 1973), any test that distinguishes between the two distributions has probability of failure at least $\frac{1}{2}(1 - (d+1)e^{-\Omega(c^2 n\eta)})$. ■

Lemma 10 (Hybrid argument for constructing couplings) *Let $D = \mathcal{N}((\mu_1, \ldots, \mu_d), \Sigma)$ and $D' = \mathcal{N}((\mu'_1, \ldots, \mu'_d), \Sigma)$. There exists a coupling Π between D and D' such that*

$$\mathop{\mathbb{E}}_{(X,X')\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X'_i)\right] = \sum_{i=0}^{d-1} D_{\mathrm{TV}}\left(Q_i, Q_{i+1}\right),$$

where $Q_i = \mathcal{N}((\mu'_1\mu'_2, \ldots, \mu'_i, \mu_{i+1}, \ldots, \mu_d), \Sigma)$ is the Gaussian whose mean has the same values as μ' in the first i coordinates and the same as μ elsewhere ($Q_0 = D$ and $Q_d = D'$).

Before presenting the proof, we note that although the statement is written in terms of Gaussian distributions (which is the setting we will apply it to later), the argument applies to any distributions D and D', where D' is a shifted version of D (in the sense that a random variable from D' can be written as a random variable from D plus a deterministic vector).

Proof

This lemma essentially follows from the observation that two Gaussians with the same covariance and means differing in only one coordinate can be coupled so that the disagreement occurs only on that coordinate (and with probability at most equal to their total variation distance) while all other coordinates always agree. This is formalized in the claim below:

Claim 11 *Consider the two d-dimensional Gaussians $Q = \mathcal{N}((\mu_1, \mu_2, \ldots, \mu_d), \Sigma)$ and $Q' = \mathcal{N}((\mu_1, \ldots, \mu_{i-1}, \mu'_i, \mu_{i+1}, \ldots, \mu_d), \Sigma)$. There is a coupling Π between the distributions Q and Q' such that $\mathbb{P}_{(X,X')\sim\Pi}[X_i \neq X'_i] = D_{\mathrm{TV}}(Q, Q')$ and $\mathbb{P}_{(X,X')\sim\Pi}[X_j \neq X'_j] = 0$ for all $j \neq i$.*

Proof (Proof of Claim 11) Without loss of generality we use $i = 1$ in this proof. Denote by $(X_1, \ldots, X_d)$ a random vector distributed as $Q = \mathcal{N}((\mu_1, \mu_2, \ldots, \mu_d), \Sigma)$ and by $(X'_1, \ldots, X'_d)$ a random vector distributed as $Q' = \mathcal{N}((\mu'_1, \mu_2 \ldots, \mu_d), \Sigma)$. Also, for any $Z \in \mathbb{R}^{d-1}$ denote by P_Z the distribution of X_1 conditioned on $(X_2, \ldots, X_d) = Z$ and by P'_Z the distribution of X'_1 conditioned on $(X'_2, \ldots, X'_d) = Z$. The coupling Π that satisfies the guarantee in the claim statement is the distribution between the pair of vectors $(\widetilde{X}_1, \ldots, \widetilde{X}_d), (\widetilde{Y}_1, \ldots, \widetilde{Y}_d)$ created as follows:

1. Draw $Z \in \mathbb{R}^{d-1}$, according to the marginal distribution of Q in the coordinates $2, 3, \ldots, d$. (Note that this marginal is the same under Q and Q').
2. Set $(\widetilde{X}_2, \ldots, \widetilde{X}_d) = Z$ and $(\widetilde{Y}_2, \ldots, \widetilde{Y}_d) = Z$.
3. Draw $\widetilde{X}_1, \widetilde{Y}_1$ from the maximal coupling Π^*_Z (given in Fact 7) between the distributions P_Z and P'_Z.

By construction, the marginal of $(\widetilde{X}_1, \ldots, \widetilde{X}_d)$ is Q and that of $(\widetilde{Y}_1, \ldots, \widetilde{Y}_d)$ is Q' thus Π is a valid coupling. We also trivially have that $\mathbb{P}_{(X,X')\sim\Pi}[\widetilde{X}_j \neq \widetilde{Y}_j] = 0$ for all $j \neq 1$. For the first coordinate, we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{(\widetilde{X},\widetilde{Y})\sim\Pi}[\widetilde{X}_1 \neq \widetilde{Y}_1] &= \mathop{\mathbb{E}}_{Z}\left[\mathop{\mathbb{P}}_{(\widetilde{X}_1,\widetilde{Y}_1)\sim\Pi^*_Z}[X_1 \neq \widetilde{Y}_1 \mid Z]\right] \\
&= \mathop{\mathbb{E}}_{Z}\left[D_{\mathrm{TV}}\left(P_Z, P'_Z\right)\right] && \text{(using Fact 7)} \\
&= D_{\mathrm{TV}}\left(Q, Q'\right). && \text{(law of total expectation)}
\end{aligned}$$

■

We now show how Lemma 10 follows given Claim 11. We will show a procedure to generate random variables $X^{(i)} \in \mathbb{R}^d$ for $i = 0, \ldots, d$ and the coupling Π that realizes Lemma 10 will be the joint distribution of $X^{(0)}$ and $X^{(d)}$. The generating procedure is the following:

1. Let Π_1 be the coupling that Claim 11 gives for the Gaussians $Q_0 = \mathcal{N}((\mu_1, \mu_2, \ldots, \mu_d), \Sigma)$ and $Q_1 = \mathcal{N}((\mu_1', \mu_2, \ldots, \mu_d), \Sigma)$. Draw $X^{(0)}$ and $X^{(1)}$ from Π_1.

2. Let Π_2 be the coupling that Claim 11 gives for the Gaussians $Q_1 = \mathcal{N}((\mu_1', \mu_2, \mu_3, \ldots, \mu_d), \Sigma)$ and $Q_2 = \mathcal{N}((\mu_1', \mu_2', \mu_3, \ldots, \mu_d), \Sigma)$. Draw $X^{(2)}$ from Π_2 conditioned on the value of $X^{(1)}$ from the previous step.

3. In general, in the i-th step, let Π_i be the coupling that Claim 11 gives for the Gaussians $Q_{i-1} = \mathcal{N}((\mu_1', \ldots, \mu_{i-1}', \mu_i, \ldots, \mu_d), \Sigma)$ and $Q_i = \mathcal{N}((\mu_1', \ldots, \mu_{i-1}', \mu_i', \mu_{i+1}, \ldots, \mu_d), \Sigma)$. Draw $X^{(i)}$ from Π_i conditioned on the value of $X^{(i-1)}$ from the previous step.

The expected number of coordinates that disagree is bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{d} \mathbb{1}(X_i^{(0)} \neq X_i^{(d)})\right] &= \sum_{i=1}^{d} \mathop{\mathbb{E}}_{X^{(i-1)}}\left[\mathop{\mathbb{E}}_{X^{(i)}}[\mathbb{1}(X_i^{(i)} \neq X_i^{(i-1)}) \,|\, X^{(i-1)}]\right] \\
&= \sum_{i=1}^{d} \mathop{\mathbb{E}}_{X^{(i-1)}, X^{(i)}}[\mathbb{1}(X_i^{(i)} \neq X_i^{(i-1)})] \\
&= \sum_{i=1}^{d} D_{\mathrm{TV}}(Q_{i-1}, Q_i),
\end{aligned}$$

where the first line follows from the fact that a disagreement between $X^{(0)}$ and $X^{(d)}$ in the i-th coordinate can occur only during the i-th step of the generation process (by the design of the couplings in Claim 11). The transition from the second to the third line uses the law of total probability to rearrange the expectations; and the final line follows from the guarantee provided by the couplings in Claim 11.

■

C.2. Proof of Theorem 5(a)

We restate and prove the following lower bound.

Theorem 22 (Lower Bound for regime $\|\beta\| \leq \sigma$) *Let c be a sufficiently small positive absolute constant. For any $d \in \mathbb{Z}_+$, $\sigma, \eta, b \in \mathbb{R}_+$ with $\eta \in [0, 1]$, $\sigma > 0$ and $b \leq \sigma$ the following statement holds. For every algorithm $\mathcal{A}$ that takes as input η, σ, b as well as n labeled examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ with $x^{(i)}$ and $y^{(i)} \in \mathbb{R}$ and outputs a vector $\widehat{\beta} \in \mathbb{R}^d$, there exists a $\beta \in \mathbb{R}^d$ with $\|\beta\| = b$ such that running $\mathcal{A}$ on input η, σ, b and n labeled examples from the model of Definition 1 with regressor β, standard deviation σ for the additive noise, and η-fraction of missing data per coordinate according to the contamination model of Definition 2, the output $\widehat{\beta}$ satisfies*

$$\left\|\widehat{\beta} - \beta\right\| \geq c \min(\|\beta\|, \eta\sqrt{d}\sigma). \tag{7}$$

with probability at least $\frac{1}{2}(1 - (d+1)e^{-\Omega(\eta n)})$.

Proof We first focus on proving the result for $\|\beta\| \le \eta\sqrt{d}\sigma$, in which case $\min(\|\beta\|, \eta\sqrt{d}\sigma) = \|\beta\|$. We describe at the end how to extend it for $\|\beta\| > \eta\sqrt{d}\sigma$.

For any $b \in [0, \eta\sqrt{d}\sigma]$ we define the hypothesis testing problem of distinguishing between the regression vectors $\beta^{(0)}, \beta^{(1)}$ defined below from n samples from the models of Definitions 1 and 2:

1. (Null Hypothesis) $\beta^{(0)} = (b/\sqrt{d}, \dots, b/\sqrt{d})$.
2. (Alternative Hypothesis) $\beta^{(1)} = -\beta^{(0)}$.

Note that $\|\beta^{(0)}\| = \|\beta^{(1)}\| = b$ and $\|\beta^{(0)} - \beta^{(1)}\| = \sqrt{2}b$. By the standard reduction from estimation to hypothesis testing, showing that no algorithm can solve the above testing problem will imply that no estimator has Euclidean error smaller than $b/\sqrt{2}$. By Fact 6, for each of the two hypotheses $i \in \{0, 1\}$ above, the conditional distribution $X|(y = t)$ of the covariates given that the label is t is a Gaussian $\mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$ where

$$\mu^{(i)} := \frac{t}{\sigma^2 + b^2}\beta^{(i)}, \quad \text{and} \quad \Sigma^{(1)} = \Sigma^{(2)} = \Sigma := I - \frac{\beta^{(1)}(\beta^{(1)})^\top}{\sigma^2 + b^2}. \tag{8}$$

In order to prove that the testing problem is not solvable, it suffices to find a coupling between $\mathcal{N}(\mu^{(0)}, \Sigma), \mathcal{N}(\mu^{(1)}, \Sigma)$ such that for each $i \in [d]$, only up to η-fraction of samples have their i-th coordinate corrupted. By Lemma 10, for each $i \in [d]$ we need to upper bound the TV-distance between the two Gaussians $\mathcal{N}(m^{(i)}, \Sigma), \mathcal{N}(m^{(i+1)}, \Sigma)$, where

$$m^{(i)} = (\mu_1^{(1)}, \dots, \mu_{i-1}^{(1)}, \mu_i^{(0)}, \mu_{i+1}^{(0)}, \dots, \mu_d^{(0)}), \tag{9}$$

$$m^{(i+1)} = (\mu_1^{(1)}, \dots, \mu_{i-1}^{(1)}, \mu_i^{(1)}, \mu_{i+1}^{(0)}, \dots, \mu_d^{(0)}). \tag{10}$$

We will do this by bounding the KL-divergence and relating it to the total variation distance via Pinsker's inequality (Fact 19). By Fact 18, we have

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathcal{N}(m^{(i)}, \Sigma) \,\|\, \mathcal{N}(m^{(i+1)}, \Sigma)) && (11)\\
&\le \frac{1}{2}(m^{(i+1)} - m^{(i)})^\top \Sigma^{-1}(m^{(i+1)} - m^{(i)}) && (12)\\
&\le \frac{1}{2}(m^{(i+1)} - m^{(i)})^\top \left(I + \frac{\beta^{(1)}\beta^{(1)\top}}{\sigma^2}\right)(m^{(i+1)} - m^{(i)}) && \text{(using Fact 21)}\\
&\le \frac{1}{2}\|m^{(i+1)} - m^{(i)}\|^2 + \frac{1}{2}\frac{((m^{(i+1)} - m^{(i)})^\top \beta^{(1)})^2}{\sigma^2} && (13)\\
&\le \frac{2t^2}{d}\frac{b^2}{(\sigma^2 + b^2)^2} + \frac{1}{2\sigma^2}\left(\frac{2t}{\sqrt{d}}\frac{b}{\sigma^2 + b^2}\frac{b}{\sqrt{d}}\right)^2 && \text{(see explanation below)}\\
&= \frac{2t^2}{d}\frac{b^2}{(\sigma^2 + b^2)^2} + \frac{1}{\sigma^2}\frac{2t^2}{d}\frac{b^2}{(\sigma^2 + b^2)^2}\frac{b^2}{d} && (14)\\
&= \frac{2t^2}{d}\frac{b^2}{(\sigma^2 + b^2)^2}\left(1 + \frac{b^2}{d\sigma^2}\right) && (15)\\
&\le \frac{2t^2}{d}\frac{b^2}{(\sigma^2 + b^2)^2}(1 + \eta^2) && \text{(using assumption } b \le \eta\sqrt{d}\sigma)\\
&\le \frac{4t^2}{d}\frac{b^2}{(\sigma^2 + b^2)^2}, && \text{(using } \eta \le 1)
\end{aligned}$$

where the fourth inequality above follows from equations (8) to (10) as follows: the vector $m^{(i+1)}-m^{(i)}$ has zero in all coordinates except the i-th, where it equals $\frac{t}{\sigma^2+b^2}\cdot\frac{b}{\sqrt{d}}$.

Combining the above with Pinsker's inequality (Fact 19), we obtain

$$D_{\mathrm{TV}}(\mathcal{N}(m^{(i)},\Sigma),\mathcal{N}(m^{(i+1)},\Sigma)) \le \frac{\sqrt{2}|t|}{\sqrt{d}}\frac{b}{\sigma^2+b^2}. \tag{16}$$

By Lemma 10, there exists a coupling Π between $\mathcal{N}(m^{(0)},\Sigma),\mathcal{N}(m^{(d)},\Sigma)$ such that

$$\mathop{\mathbb{E}}_{(X,X')\sim\Pi}\left[\sum_{i=1}^{d}\mathbb{1}(X_i\neq X'_i)\right] \le \frac{\sqrt{2}\sqrt{d}|t|b}{\sigma^2+b^2}. \tag{17}$$

As explained earlier, the Gaussians involved in the above coupling are the conditional distributions of the covariates in our linear regression model (Definition 1), given that the label y equals t. We can easily extend this coupling to a new coupling Π', which couples the entire labeled examples from our linear regression model under the two hypotheses, as follows (recall that the distribution of labels under both hypotheses is $\mathcal{N}(0,\sigma^2+\|\beta^{(i)}\|^2)=\mathcal{N}(0,\sigma^2+b^2)$:

1. Draw $t\sim\mathcal{N}(0,\sigma^2+b^2)$ and set $y=y'=t$.
2. Draw (X,X') from the coupling Π that satisfies Equation (17).
3. Return $(X,y),(X',y)$.

By construction, (X,y) is marginally distributed according to the linear regression model with regressor $\beta^{(0)}$ and (X',y') is marginally distributed according to the linear regression model with regressor $\beta^{(1)}$, thus Π' is a valid coupling. Moreover, the expected number of disagreements is

$$\begin{aligned}
\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi'}\left[\sum_{i=1}^{d}\mathbb{1}(X_i\neq X'_i)\right] &\le \frac{\sqrt{2}\sqrt{d}\,b\,\mathbb{E}_{t\sim\mathcal{N}(\sigma^2+b^2)}[|t|]}{\sigma^2+b^2} \le \frac{\sqrt{2}\sqrt{d}\,b}{\sqrt{\sigma^2+b^2}} \\
&\qquad\qquad \text{(using Equation (17))} \\
&\le \frac{\sqrt{2}\sqrt{d}\,b}{\sigma} \qquad\qquad \text{(using } b\ge 0\text{)} \\
&= \frac{\sqrt{2}\sqrt{d}\eta\sigma\sqrt{d}}{\sigma} \qquad\qquad \text{(using } b\le\sqrt{d}\eta\sigma\text{)} \\
&\le \sqrt{2}\eta d \le \sqrt{2}\eta d.
\end{aligned}$$

The labels always agree in this coupling. The constant $\sqrt{2}$ in front of η above is not particularly important; we could obtain the same bound with a smaller constant, such as $1/3$, in front of d by simply redefining η as $\eta/(2\sqrt{2})$ at the beginning of the proof. This would allow us to apply Lemma 9 to conclude that no algorithm can solve the hypothesis testing problem (except with probability $\frac{1}{2}(1-(d+1)e^{-\Omega(\eta n)})$), thus completing the proof of Theorem 22 for the regime $\|\beta\|\le\eta\sqrt{d}\sigma$.

For the remaining regime $\eta\sqrt{d}\sigma\le\|\beta\|$, consider instead the hypothesis testing problem over $\mathbb{R}^d$: distinguish $\beta=(r,b/\sqrt{d},\dots,b/\sqrt{d})$ from $\beta'=(r,-b/\sqrt{d},\dots,-b/\sqrt{d})$, where $b\le\eta\sqrt{d}\sigma$ is as before and r is a tunable parameter allowing $\|\beta\|$ to get any desired value larger than $\eta\sqrt{d}\sigma$. The labels can be written as $y=rX_1+y_0$ and $y'=rX'_1+y'_0$, where $y_0=(b/\sqrt{d},\dots,b/\sqrt{d})^\top X_{2:d}+\xi$

and $y'_0 = (-b/\sqrt{d}, \ldots, -b/\sqrt{d})^\top X'_{2:d} + \xi'$, with $X_{2:d}$ denoting the last d coordinates of X. As shown earlier, there exists a coupling between $(X_{2:d}, y_0)$ and $(X'_{2:d}, y'_0)$ with expected disagreements at most $\eta d/2$ per sample. This extends trivially to a coupling between (X, y) and (X', y') (by adding shared Gaussian noise $\mathcal{N}(0, r^2)$ to the labels), preserving the disagreement bound. Applying Lemma 9 as before completes the proof. ■

Appendix D. Omitted Details from Section 4

In this section we restate and prove the lower bounds corresponding to parts (b),(c) and (d) of Theorem 5. Since these bounds depend on Lemma 13, we restate that lemma below for convenience:

Lemma 13 (Improved core coupling) *For any $d \in \mathbb{Z}_+$, $t, t' \in \mathbb{R}$, the following hold. If D denotes the distribution of $(X_1, \ldots, X_d) \sim \mathcal{N}(0, I_d)$ conditioned on $\sum_{i=1}^d X_i = t$ and D' the distribution of $(X'_1, \ldots, X'_d) \sim \mathcal{N}(0, I_d)$ conditioned on $\sum_{i=1}^d X'_i = t'$, then there exists a coupling Π between D, D' such that $\mathbb{E}_{(X,X'))\sim\Pi}\left[\sum_{i=1}^d \mathbb{1}\left(X_i \neq X'_i\right)\right] \leq 1 + |t - t'|$.*

We start with Theorem 5(b)

Theorem 23 (Lower bound for regime $\eta \geq 7/\sqrt{d}$) *The following holds for every $d \in \mathbb{Z}_+$ and $\sigma, \eta, b \in \mathbb{R}_+$ with $7/\sqrt{d} \leq \eta \leq 1$. For every algorithm $\mathcal{A}$ that takes as input η, σ, b as well as n labeled examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ with $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$ and outputs a vector $\widehat{\beta} \in \mathbb{R}^d$, there exists a $\beta \in \mathbb{R}^d$ with $\|\beta\| = b$ such that running $\mathcal{A}$ on input η, σ, b and n labeled examples from the model of Definition 1 with regressor β, standard deviation σ for the additive noise, and η-fraction of missing data per coordinate according to the contamination model of Definition 2, the output $\widehat{\beta}$ satisfies*

$$\left\|\widehat{\beta} - \beta\right\| \geq \frac{1}{\sqrt{2}}\|\beta\|, \tag{18}$$

with probability at least $\frac{1}{2}(1 - (d+1)e^{-\Omega(\eta n)})$.

Proof Let $\beta = s(1, 1 \ldots, 1)$ be the all-ones vector and $\beta' = -\beta$ the vector with -1 in every coordinate where s is a tunable parameter so that $\|\beta\|$ can have any desired value (we want to prove that the lower bound holds for any $\|\beta\|$). We consider the following hypothesis testing problem:

- (Null Hypothesis) The regression vector is β.
- (Alternative Hypothesis) The regression vector is β'.

It suffices to show that no algorithm distinguishes between the two hypotheses with probability better than $\frac{1}{2}(1 + (d+1)e^{-\Omega(\eta n)})$. Since $\|\beta - \beta'\| = \sqrt{2d} \geq \sqrt{2}\max(\|\beta\|, \|\beta'\|)$, by the standard reduction between estimation and hypothesis testing, this would imply that no algorithm can estimate β with error smaller than $\|\beta\|/\sqrt{2}$.

First, it is easy to see that the coupling construction from Lemma 13 allows us to couple the distributions of labeled examples for the two hypotheses while ensuring a small number of coordinate-wise disagreements. This is shown in the lemma below:

Lemma 24 *Let $\sigma \geq 0$ and a scaling parameter $s \geq 0$. Define $\beta = s(1, 1, \ldots, 1)$ as a scaled version of the all-ones vector in $\mathbb{R}^d$ and $\beta' = -\beta$. If D denotes the distribution of a labeled example (X, y) drawn from the linear regression model of Definition 1 with regressor β and standard deviation of additive noise σ, and D' denotes the distribution of a labeled example according to a linear regression model with regressor β' and standard deviation of additive noise σ, then there exists a coupling Π between D and D' such that $\mathbb{P}_{((X,y),(X',y'))\sim\Pi}[y = y'] = 1$ and*

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] \leq 3\sqrt{d}. \tag{19}$$

Proof (Proof of Lemma 24) It suffices to prove the lemma for $s = 1$ and $\sigma = 0$. This is because if $((X, y), (X', y'))$ is distributed according to a coupling with the desired properties ($\mathbb{P}[y = y'] = 1$ and $\mathbb{E}\left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] \leq 3\sqrt{d}$) for the case of $s = 1$ and $\sigma = 0$, then we can let $\xi \sim \mathcal{N}(0, \sigma^2)$ and the pair $((sX, sy + \xi), (sX', sy' + \xi))$ will be the final coupling that corresponds to the case with positive σ and $s \neq 1$ and continues to satisfy $\mathbb{P}[y = y'] = 1$ and $\mathbb{E}\left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] \leq 3\sqrt{d}$. The coupling for the case $s = 1, \sigma = 0$ is the following:

- Draw $z \sim \mathcal{N}(0, d)$ and set $y = y' = z$.
- Draw $((X_1, \ldots, X_d), (X_1', \ldots, X_d'))$ from the coupling of Lemma 13 applied with $t = z$ and $t' = -z$.

It is easy to verify that the labeled example $(X_1, \ldots, X_d, y)$ is marginally distributed according to the linear model that uses $\beta = (1, \ldots, 1)$ as the regressor, and the example $(X_1', \ldots, X_d', y)$ is distributed according to the linear model using $\beta' = (-1, \ldots, -1)$. Moreover, by Lemma 13, the expected number of disagreements is

$$\mathop{\mathbb{E}}_{(X_1,\ldots,X_d),(X_1',\ldots,X_d')}\left[\mathbb{1}(X_i \neq X_i')\right] \leq 1 + \mathop{\mathbb{E}}_{t,t'}[|t - t'|] = 1 + \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,d)}[|2z|] \leq 3\sqrt{d}.$$

The proof of Lemma 24 is completed. ■

When $\eta \geq 7/\sqrt{d}$ the right hand side of Equation (19) is at most $\eta d 3/7$. Thus, by Lemma 9 , no algorithm can solve the hypothesis testing defined in the beginning with probability higher than $\frac{1}{2}(1 + (d + 1)e^{-\Omega(\eta n)})$.

■

We now move to Theorem 5(c) which is restated and proved below.

Theorem 25 (Lower bound for regime $\frac{2+c}{d} \leq \eta \leq \frac{7}{\sqrt{d}}$) *There exists a sufficiently large absolute constant C such that the following holds for every $d \in \mathbb{Z}_+$, every $c \in (0, 1)$, every $\sigma \geq 0$, every $\eta \in [\frac{2+c}{d}, \frac{7}{\sqrt{d}}]$, and every $b \in \mathbb{R}_+$. For every algorithm $\mathcal{A}$ that takes as input η, σ, b as well as n labeled examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ with $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$ and outputs a vector $\widehat{\beta} \in \mathbb{R}^d$, there exists a $\beta \in \mathbb{R}^d$ with $\|\beta\| = b$ such that running $\mathcal{A}$ on input η, σ, b and n labeled examples from the model of Definition 1 with regressor β, standard deviation σ for the additive noise, and η-fraction*

of missing data per coordinate according to the contamination model of Definition 2, the output $\widehat{\beta}$ *satisfies*

$$\left\|\widehat{\beta} - \beta\right\| \geq \eta \frac{c\sqrt{d}}{C} \|\beta\|, \tag{20}$$

with probability at least $\frac{1}{2}(1 - (d+1)e^{-\Omega(c^2\eta n)})$.

Proof

Let $s \geq 0$ be a scaling factor to control the norm of the regressor. We will use a parameter $\varepsilon \in (0,1)$ that will be specified later. Fix $\beta = s(\varepsilon, \ldots, \varepsilon, 1, \ldots, 1)$ and $\beta' = s(-\varepsilon, \ldots, -\varepsilon, 1, \ldots, 1)$ as two vectors in $\mathbb{R}^d$, where the first $d/2$ coordinates are εs for β and $-\varepsilon s$ for β', while the last $d/2$ coordinates are all equal to s. By the estimation to hypothesis testing reduction discussed in Section 2, it suffices to show that no algorithm that uses n η-corrupted samples can solve the following hypothesis testing problem:

- (Null Hypothesis) The regression vector is β.
- (Alternative Hypothesis) The regression vector is β'.

Note that we can use s to make the norm of $\|\beta\|$ and $\|\beta'\|$ have any desired value b (as in the statement of Theorem 25). Thus we focus on showing hardness of the testing problem in what follows.

Using the coupling construction from Section 3, we can construct a coupling between the linear regression distributions corresponding to the two hypotheses such that the number of disagreements per coordinate is upper bounded as stated in Lemma 26 below.

Lemma 26 *Let* d *be a power of* 2, $\varepsilon \in [0,1]$, $\sigma \geq 0$, *and* $s \geq 0$. *Let* $\beta = s(\varepsilon, \ldots, \varepsilon, 1, \ldots, 1)$ *and* $\beta' = s(-\varepsilon, \ldots, -\varepsilon, 1, \ldots, 1)$ *be the two vectors in* $\mathbb{R}^d$, *where the first* $d/2$ *coordinates are* $s\,\varepsilon$ *for* β *and* $-s\,\varepsilon$ *for* β', *while the last* $d/2$ *coordinates are all equal to* s. *If* D *denotes the distribution of a labeled example drawn from the linear model of Definition 1 with regressor* β *and standard deviation of additive noise* σ, *and* D' *denotes the distribution of a labeled example according to a linear model with regressor* β' *and standard deviation of additive noise* σ, *then there exists a coupling* Π *between* D *and* D' *such that* $\mathbb{P}_{((X,y),(X',y'))\sim\Pi}[y = y'] = 1$ *and*

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi} \left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X_i'\right)\right] = 2 + O\left(\varepsilon\sqrt{d}\right). \tag{21}$$

The proof of Lemma 26 is somewhat tedious, so we defer it until after we show how the theorem follows from Lemma 26. Denote by C the absolute constant hidden inside the big-O notation in Equation (21), and denote by c' an additional parameter that is less than 0.25. We will re-parameterize things as follows: $\eta = \frac{2}{(1-c')d} + \frac{C\varepsilon}{(1-c')\sqrt{d}}$, or equivalently $\varepsilon := (\eta - \frac{2}{(1-c')d})\frac{\sqrt{d}}{C}$. This is so that the right-hand side of Equation (21) becomes equal to $\eta d(1-c')$ and we can use Lemma 9 to conclude that no algorithm can solve the hypothesis testing problem with probability better than $\frac{1}{2}(1 + (d+1)e^{-\Omega(c'\eta n)})$. By the estimation to hypothesis testing reduction, this means that every

estimation algorithm has error at least $\Omega(\|\beta - \beta'\|)$. For the precise form of this error we have the following lower bounds:

$$\begin{aligned}
\|\beta - \beta'\| &= s\,\varepsilon\sqrt{d/2} && \text{(by definition of } \beta, \beta') \\
&= \frac{\varepsilon}{\sqrt{1+\varepsilon^2}} \max(\|\beta\|, \|\beta'\|) && (\|\beta\| = \|\beta'\| = s\sqrt{d(1+\varepsilon^2)/2}) \\
&\gtrsim \varepsilon \max(\|\beta\|, \|\beta'\|) && (\varepsilon := (\eta - \tfrac{2}{(1-c')d})\tfrac{\sqrt{d}}{C} \le 1 \text{ since } \eta \le 7/\sqrt{d}) \\
&\gtrsim \left(\eta - \frac{2}{(1-c')d}\right)\sqrt{d} \max\left(\|\beta\|, \|\beta'\|\right) && (\text{using } \varepsilon := (\eta - \tfrac{2}{(1-c')d})\tfrac{\sqrt{d}}{C}) \\
&\gtrsim c'\eta\sqrt{d} \max\left(\|\beta\|, \|\beta'\|\right), && (\text{this step holds if } \eta \ge \tfrac{2+10c'}{d}.)
\end{aligned}$$

where the last step uses that $\eta \ge \frac{2+10c'}{d}$ implies $\eta \ge \frac{2+10c'}{d} \ge \frac{2}{(1-c')^2 d}$ and means that the terms in the parentheses from the previous step can be lower bounded as $\eta - \frac{2}{(1-c')d} \ge c'\eta$. The fact that $\eta \ge \frac{2+10c'}{d}$ is due to the assumption in the theorem statement that $\eta \ge (2+c)/d$ (and the fact that c' is a parameter that we can choose to be $c' = c/10$).

We now prove Lemma 26 that was used earlier.

Proof (Proof of Lemma 26) We first note that it suffices to prove the claim for $s = 1$ and $\sigma = 0$. To see that, let us momentarily use the notation $D_{s,\sigma}, D'_{s,\sigma}$ for the two distributions of the statement explicitly indicating the values of s and σ. Suppose that there exists the desired coupling Π between $D_{1,0}$ and $D'_{1,0}$ (i.e., for $s = 1$ and $\sigma = 0$). If $((X, y), (X', y')) \sim \Pi$ then $((sX, sy), (sX', sy'))$ is a coupling with the desired properties for $D_{s,0}, D'_{s,0}$, i.e., $\mathbb{P}[sy = sy'] = 1$ and $\mathbb{E}[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X'_i)] = 2 + O(\varepsilon\sqrt{d})$. If we further consider $\xi \sim \mathcal{N}(0, \sigma^2)$, then the pair $((sX, sy + \xi), (sX', sy' + \xi))$ is the desired coupling between $D_{s,\sigma}, D'_{s,\sigma}$.

For the remainder of the proof, we thus use $s = 1$ and $\sigma = 0$. We now define how the coupling that generates the pair $((X, y), (X', y'))$. While the definition may initially appear complicated, we will argue below that the resulting distribution is indeed a valid coupling and ultimately show that it satisfies $y = y'$ as well as Equation (21).

1. We first sample $z \sim \mathcal{N}(0, (1+\varepsilon^2)d/2)$ and set $y = y' = z$.
2. We then sample t and t' from a coupling such that the marginals are $t \sim \mathcal{N}(\varepsilon z/(1+\varepsilon^2), d/(2(1+\varepsilon^2))$ and $t' \sim \mathcal{N}(-\varepsilon z/(1+\varepsilon^2), d/(2(1+\varepsilon^2))$ and it also holds $t - t' = 2\varepsilon z/(1+\varepsilon^2)$ always (i.e., with probability 1).
3. We sample $(X_1, \ldots, X_{d/2}), (X'_1, \ldots, X'_{d/2})$ according to the coupling from Lemma 13, i.e., the distribution of $(Z_1, \ldots, Z_{d/2}) \sim \mathcal{N}(0, I)$ conditioned on $\sum_{i\in[d/2]} Z_i = t$ and the distribution of $(Z'_1, \ldots, Z'_{d/2}) \sim \mathcal{N}(0, I)$ conditioned on $\sum_{i\in[d/2]} Z'_i = t'$.
4. We sample $(X_{d/2}, \ldots, X_d), (X'_{d/2}, \ldots, X'_d)$ from the coupling of Lemma 13 but now using the conditioning $\sum_{i=d/2+1}^{d} Z_i = z - \varepsilon t$ and $\sum_{i=d/2+1}^{d} Z'_i = z + \varepsilon t'$ respectively.

We start by verifying that this is indeed a valid coupling, i.e., that $(X_1, \ldots, X_d, y)$ follows the linear model with regressor β and $(X'_1, \ldots, X'_d, y')$ the linear model with regressor β'. We will perform the check for the first part (i.e., check that $(X_1, \ldots, X_d, y)$ indeed follows the linear model with regressor β). Checking the other part can be done with the same argument and replacing ε

with $-\varepsilon$. For notational purposes only, we let another example $(\widetilde{X}_1, \dots \widetilde{X}_d, \widetilde{y})$ be a random labeled example distributed according to the distribution D from the lemma statement.

First, the marginal distribution of $\widetilde{y}$ is $\widetilde{y} = \beta^\top \widetilde{X} \sim \mathcal{N}(0, \|\beta\|^2) = \mathcal{N}(0, (1+\varepsilon^2)d/2)$. This explains the first step (Item 1 in our coupling procedure). To explain the rest of the steps, we want to argue that $(X_1, \dots, X_d)|(y = z)$ in our procedure is distributed in the same way as $(\widetilde{X}_1, \dots, \widetilde{X}_d)|(\widetilde{y} = z)$ (i.e, in the same way as under the distribution D from the lemma statement). The distribution under D can be viewed as follows: First we sample a value t from the distribution of $\sum_{i\in[d/2]} \widetilde{X}_i$ conditioned on $\widetilde{y} = z$ (let us call this distribution D_1 for later referencing), then we sample the coordinates $(\widetilde{X}_1, \dots, \widetilde{X}_{d/2})$ from the distribution of a standard Gaussian vector, conditioned on $\sum_{i\in[d/2]} \widetilde{X}_i = t$ and finally we sample the rest of the coordinates $(\widetilde{X}_{d/2}, \dots, \widetilde{X}_d)$ from the distribution of a standard Gaussian vector, conditioned on $\sum_{i=d/2+1}^{d} \widetilde{X}_i = z - \varepsilon t$ (so that $\widetilde{y} = \beta^\top \widetilde{X} = \sum_{i=1}^{d/2} \varepsilon \widetilde{X}_i + \sum_{i=d/2+1}^{d} \widetilde{X}_i = \varepsilon t + z - \varepsilon t = z$). It can be checked that the distribution D_1 mentioned earlier is $\mathcal{N}(\varepsilon z/(1+\varepsilon^2), d/(2(1+\varepsilon^2))$. This can be seen as follows: First, by Fact 6 applied with $u = \beta$ and $\sigma = 0$, the conditional distribution of the entire vector, given $y = z$ is

$$\widetilde{X}_1, \dots, \widetilde{X}_{d/2}, \dots, \widetilde{X}_d | (\widetilde{y} = z) \sim \mathcal{N}\left(z\frac{\beta}{\|\beta\|^2}, I_d - \frac{\beta\beta^\top}{\|\beta\|^2} \right).$$

This means that

$$\widetilde{X}_1, \dots, \widetilde{X}_{d/2} | \widetilde{y} = z \sim \mathcal{N}\left(z\frac{\beta_{1:d/2}}{\|\beta\|^2}, I_{d/2} - \frac{\beta_{1:d/2}\beta_{1:d/2}^\top}{\|\beta\|^2} \right),$$

where the notation $\beta_{1:d/2}$ denotes the vector formed by taking the first $d/2$ coordinates of β and $I_{d/2}$ is the $(d/2) \times (d/2)$ identity matrix. Note that $\beta_{1:d/2}$ is the vector in $\mathbb{R}^{d/2}$ with value ε in every coordinate. Then, by Fact 17 applied with u being the all-ones vector, we have

$$\sum_{i=1}^{d/2} \widetilde{X}_i \big| (\widetilde{y} = z) \sim \mathcal{N}\left(\frac{\varepsilon z}{(1+\varepsilon^2)}, \frac{d}{2}\frac{1}{(1+\varepsilon^2)} \right).$$

This completes the proof that the procedure generating $\widetilde{X}_1, \dots, \widetilde{X}_d, y$ that is described in the bullets at the start of this proof matches the distribution of a labeled example under the linear model with regressor β. The check for the marginal of $X'_1, \dots, X'_d$ is similar and we skip it. So far we have thus shown that the generating process from the bullets defines a valid coupling Π between D and D'.

Finally, we analyze the expected number of disagreements to show Equation (21). Because of our two applications of Lemma 13, this expected number of disagreements is

$$\begin{aligned} \mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi} \left[\sum_{i=1}^{d} \mathbb{1}\left(X_i \neq X'_i\right) \right] &= \mathop{\mathbb{E}}_{t,t'} \left[2 + O\left(|t - t'| + |z - \varepsilon t - (z + \varepsilon t')| \right) \right] \\ &= 2 + O\left(\mathop{\mathbb{E}}_{t,t'} \left[|t - t'| \right] \right) + \varepsilon\, O\left(\mathop{\mathbb{E}}_{t,t'} \left[|t + t'| \right] \right). \end{aligned} \tag{22}$$

the second term above is $O(\varepsilon\, \mathbb{E}[|z|]/(1+\varepsilon^2))$ because we have defined the coupling between t and t' to always satisfy $t - t' = 2\varepsilon z/(1+\varepsilon^2)$. The term can be further bounded by $O(\varepsilon\sqrt{d})$ using that $z \sim \mathcal{N}(0, (1+\varepsilon^2)d/2)$ and $\varepsilon \leq 1$. Regarding the last term in Equation (22), we have

$$\varepsilon \mathop{\mathbb{E}}_{t,t'} \left[|t + t'| \right] \lesssim \varepsilon \mathop{\mathbb{E}}_{t,t'} \left[|t| \right] + \varepsilon \mathop{\mathbb{E}}_{t,t'} \left[|t'| \right] \lesssim \varepsilon\sqrt{d},$$

where we used the triangle inequality and the fact that the variance of t and t' is at most $d/2$. This concludes the proof of Equation (21) and Lemma 26. ■

Since the proof of Lemma 26 is complete, this also completes the proof of Theorem 25. ■

We restate and prove Theorem 5(d) below.

Theorem 27 (Lower bound for regime $\eta \leq 1/d$) *There exists a sufficiently large absolute constant C such that the following holds for every $d \in \mathbb{Z}_+$, $\sigma \geq 0$, $\eta \in [0, 1/d]$, and $b \geq 0$. For every algorithm $\mathcal{A}$ that takes as input η, σ, b as well as n labeled examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ with $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$ and outputs a vector $\widehat{\beta} \in \mathbb{R}^d$, there exists a $\beta \in \mathbb{R}^d$ with $\|\beta\| = b$ such that running $\mathcal{A}$ on η, σ, b and n labeled examples from the model of Definition 1 with regressor β, standard deviation σ for the additive noise, and η-fraction of missing data per coordinate according to the contamination model of Definition 2, the output $\widehat{\beta}$ satisfies*

$$\left\|\widehat{\beta} - \beta\right\| \geq \frac{1}{C} \min\left(\eta d\sigma,\, \eta\sqrt{d}\|\beta\|\right), \tag{23}$$

with probability at least $\frac{1}{2}(1 - (d+1)e^{-\Omega(\eta n)})$.

Proof We let $B \in \mathbb{R}_+, E \in (0,1)$ be parameters where E will be specified later on as a function of B and η, d, σ, and B is a parameter that is chosen to ensure that the norm of β and β' below is equal to b (recall that as stated in Theorem 27, we want our lower bound to hold for any value of $\|\beta\|$). We will consider the following hypothesis testing:

- (Null Hypothesis) The regressor is $\beta := \left(\frac{B}{\sqrt{d/2}}\mathbb{1}_{d/2}, \frac{E}{\sqrt{d/2}}\mathbb{1}_{d/2}\right)$.
- (Alternative Hypothesis) The regressor is $\beta' := \left(\frac{B}{\sqrt{d/2}}\mathbb{1}_{d/2}, -\frac{E}{\sqrt{d/2}}\mathbb{1}_{d/2}\right)$.

To clarify the notation, $\mathbb{1}_{d/2}$ is the vector in $\mathbb{R}^{d/2}$ having the value 1 in all its coordinates, thus β above is the vector having $B/\sqrt{d/2}$ in the first $d/2$ coordinates and $E/\sqrt{d/2}$ in the second $d/2$ coordinates and β' is defined similarly, but with $-E$ instead of E.

By the standard reduction from estimation to hypothesis testing (described at the beginning of Section 2), showing that the above hypothesis testing problem is unsolvable implies an estimation error lower bound of $\Omega(\|\beta - \beta'\|) = \Omega(E)$.

In order to prove hardness of the hypothesis testing problem we prove the following lemma:

Lemma 28 *Let $B \in \mathbb{R}_+$ and $E \in (0,1)$ with $E \lesssim B$. Let $d \in \mathbb{Z}_+$ $\sigma \geq 0$ $\beta := \left(\frac{B}{\sqrt{d/2}}\mathbb{1}_{d/2}, \frac{E}{\sqrt{d/2}}\mathbb{1}_{d/2}\right)$ and $\beta' := \left(\frac{B}{\sqrt{d/2}}\mathbb{1}_{d/2}, -\frac{E}{\sqrt{d/2}}\mathbb{1}_{d/2}\right)$. If D denotes the distribution of a labeled example (X, y) drawn from the linear regression model of Definition 1 with regressor β and standard deviation of additive noise σ, and D' denotes the distribution of a labeled example according to the linear regression model with regressor β' and standard deviation of additive noise σ, then there exists a coupling Π between D and D' such that*

$$\mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi}\left[\sum_{i=1}^{d} \mathbb{1}(X_i \neq X_i')\right] = O\left(\frac{E}{\sigma} + \frac{\sqrt{d}E}{B}\right), \tag{24}$$

$$\textit{and} \qquad \mathop{\mathbb{E}}_{((X,y),(X',y'))\sim\Pi} \left[\mathbb{1}(y \neq y')\right] = 0. \tag{25}$$

We will prove the lemma at the end. In order to use the conclusion of Lemma 28 in combination with Lemma 9 we need the right hand sides in Equations (24) and (25) to be at most $\eta d/2$. For this it suffices to assume that $CE/\sigma < \eta d/2$ and $C\sqrt{d}E/B < \eta d/2$, where C is the hidden constant in the big-O notation. Equivalently, we need to assume $E < \frac{1}{2C}\min(\eta d\sigma, \eta\sqrt{d}B)$. Also note that our assumption $\eta \leq 1/d$ from the lemma statement also ensures that $E/\sigma < 1$, which is required by Lemma 28. Using the value $E := \frac{1}{2C}\min(\eta d\sigma, \eta\sqrt{d}B)$ we can thus apply Lemma 9 and obtain that the hypothesis testing defined in the beginning is not solvable with probability better than $\frac{1}{2}(1 + (d+1)e^{-\Omega(-\eta n)})$. By the estimation to hypothesis testing reduction, the error lower bound against any estimator is at least a constant multiple of

$$\begin{aligned}
\|\beta - \beta'\| &\gtrsim E \\
&\gtrsim \min(\eta d\sigma, \eta\sqrt{d}B) \\
&= \min\left(\eta d\sigma, \eta\sqrt{d}(B+E)\frac{B}{B+E}\right) \\
&\gtrsim \min\left(\eta d\sigma, \eta\sqrt{d}(B+E)\right) && \text{(see explanation below)} \\
&\gtrsim \min(\eta d\sigma, \eta\sqrt{d}\|\beta\|) && \text{(by construction } \|\beta\| = \Theta(B+E)\text{)}
\end{aligned}$$

where the fourth line used the following: first, since $\eta \lesssim 1/\sqrt{d}$ we have $E \lesssim \eta\sqrt{d}B \lesssim B$, and then this further implies $B/(B+E) \gtrsim B$. We now move to showing Lemma 28.

Proof (Proof of Lemma 28)

The coupling we claim satisfies the guarantee in the lemma statement is the procedure that generates (X, y) and (X', y'), which we will describe shortly in pseudocode form. Before presenting the pseudocode of that procedure, we introduce some notation: let X_1 denote the first $d/2$ coordinates of X, and X_2 the second half; we use similar notation, X'_1 and X'_2, for the corresponding coordinates of X'. Additionally, for reference purposes only, we will use another pair of labeled examples, $(\widetilde{X}, \widetilde{y})$ and $(\widetilde{X}', \widetilde{y}')$, drawn from the distributions D and D' defined in the lemma statement (the goal in our generating procedure will be for the output example (X, y) to match the distribution of $(\widetilde{X}, \widetilde{y})$ and for the second output example (X', y') to match the distribution of $(\widetilde{X}', \widetilde{y}')$). We will also let the notation $\widetilde{X}_1, \widetilde{X}_2$ for denoting the first half and second half of coordinates of $\widetilde{X}$, and similar notation for $\widetilde{X}'$. Finally we will denote $\widetilde{S}_2 := \mathbb{1}_{d/2}^\top \widetilde{X}_2/\sqrt{d/2}$ and $\widetilde{S}'_2 := \mathbb{1}_{d/2}^\top \widetilde{X}'_2/\sqrt{d/2}$ (where $\mathbb{1}_{d/2}$ is the all-ones vector of length $d/2$). We will use similar notation $\widetilde{S}_1, \widetilde{S}'_1$ for the scaled sum of the first half of the coordinates. The procedure defining our coupling is as follows.

1. Draw t from the distribution of $\widetilde{y}$ and set $y = t$ and $y' = t$.

2. Draw s_2, s'_2 from a coupling that couples the distribution of $\widetilde{S}_2$ conditioned on $\widetilde{y} = t$ and the distribution of $\widetilde{S}'_2$ conditioned on $\widetilde{y}' = t$, and also has the additional property that $\mathbb{P}[s_2 \neq s'_2] = D_{\mathrm{TV}}(\widetilde{S}_2, \widetilde{S}'_2)$ and $\mathbb{E}[|s_2 - s'_2|] = |\mathbb{E}[s_2] - \mathbb{E}[s'_2]|$. Note that such a coupling exists by Fact 7.

3. If $s_2 = s'_2$:

(a) Draw s_1, s_1' using a coupling from Fact 7 that couples the distribution of $\widetilde{S}_1$, conditioned on $\widetilde{S}_2 = s_2$ and $\widetilde{y} = t$ and the distribution of $\widetilde{S}_1'$ conditioned on $\widetilde{S}_2' = s_2$ and $\widetilde{y}' = t$.

(b) If $s_1 = s_1'$: Draw X_1 from the distribution of $\widetilde{X}_1$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_1/\sqrt{d/2} = s_1$. Set $X_1' = X_1$. Draw X_2 from the distribution of $\widetilde{X}_2$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_2/\sqrt{d/2} = s_2$. Set $X_2' = X_2$.

(c) If $s_1 \neq s_1'$: Use Lemma 13 to couple the distribution of $\widetilde{X}_1$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_1/\sqrt{d/2} = s_1$ and the distribution of $\widetilde{X}_1'$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_1'/\sqrt{d/2} = s_1'$ and draw the pair (X_1, X_1') from that coupling. Draw X_2 from the distribution of $\widetilde{X}_2$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_2/\sqrt{d/2} = s_2$. Set $X_2' = X_2$.

4. If $s_2 \neq s_2'$:

(a) Draw s_1, s_1' using a coupling that couples the distribution of $\widetilde{S}_1$, conditioned on $\widetilde{S}_2 = s_2$ and $\widetilde{y} = t$ and the distribution of $\widetilde{S}_1'$ conditioned on $\widetilde{S}_2' = s_2'$ and $\widetilde{y}' = t$, and satisfies $\mathbb{E}[|s_1 - s_1'|] \leq |\mathbb{E}[s_1] - \mathbb{E}[s_1']|$ (by Fact 7).

(b) Use Lemma 13 to couple the distribution of $\widetilde{X}_1$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_1/\sqrt{d/2} = s_1$ and the distribution of $\widetilde{X}_1'$ conditioned on $\mathbb{1}_{d/2}^\top \widetilde{X}_1'/\sqrt{d/2} = s_1'$ and draw the pair (X_1, X_1') from that coupling.

5. Return $(X_1, X_2, y), (X_1', X_2', y')$.

By construction (X, y) follows the same distribution as $(\widetilde{X}, \widetilde{y})$ (i.e., D from the lemma statement) and (X', y') follows the same distribution as $(\widetilde{X}', \widetilde{y}')$ (i.e., D') thus it is a valid coupling between D and D'. This can be seen by tracking how (X, y) was created (and identical argument will apply to (X', y')). The construction samples (X, y) by following the chain rule decomposition of the joint distribution under D: we first sample y, then sample S_2 and X_2 conditioned on y, and finally sample S_1 then X_1 conditioned on all previously sampled quantities. Since each step uses the correct conditional distribution under the (null) model, the resulting joint distribution matches that of $(\widetilde{X}, \widetilde{y})$.

We now bound the coordinate-wise disagreements. Towards that end, we will need to derive the forms of all the conditional distributions mentioned in the pseudocode. This can be done using the following fact about Gaussians.

Fact 29 *If* $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$, *then* $y_1|y_2 \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, *with* $\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$ *and* $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Using this fact, the conditional distributions of $\widetilde{S}_2$ and $\widetilde{S}_2'$ conditioned on the label values are

$$\widetilde{S}_2|\widetilde{y} = t \sim \mathcal{N}\left(\frac{Et}{B^2 + E^2 + \sigma^2}, 1 - \frac{E^2}{B^2 + E^2 + \sigma^2}\right), \tag{26}$$

$$\widetilde{S}_2'|\widetilde{y}' = t \sim \mathcal{N}\left(\frac{-Et}{B^2 + E^2 + \sigma^2}, 1 - \frac{E^2}{B^2 + E^2 + \sigma^2}\right). \tag{27}$$

Again, using the fact, the conditional distributions of the sum of the first half of coordinates $\widetilde{S}_1$ and $\widetilde{S}'_1$ given the sums of the second half of coordinates and the label values are:

$$\widetilde{S}_1|\widetilde{S}_2 = s_2, \widetilde{y} = t \sim \mathcal{N}\left(\frac{B(t - Es_2)}{B^2 + \sigma^2}, 1 - \frac{B^2}{B^2 + \sigma^2}\right), \tag{28}$$

$$\widetilde{S}'_1|\widetilde{S}'_2 = s'_2, \widetilde{y}' = t \sim \mathcal{N}\left(\frac{B(t + Es'_2)}{B^2 + \sigma^2}, 1 - \frac{B^2}{B^2 + \sigma^2}\right). \tag{29}$$

We are now ready to analyze the expected number of disagreeing coordinates. We will start with the analysis of the disagreeing coordinates, conditioned on the value of the label being t (and will take expectation over t at the end). We also consider the first half of coordinates and second half of coordinates separately. We will also denote by $\mathcal{E}_{\text{bad}}$ the event that $s_1 \neq s'_1$ or $s_2 \neq s'_2$.

For the second half we have the following:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{d/2} \mathbb{1}(X_2(i) \neq X_2(i)') \,\middle|\, y = t\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d/2} \mathbb{1}(X_2(i) \neq X_2(i)') \,\middle|\, \mathcal{E}_{\text{bad}}, y = t\right] \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] \\
&\qquad\qquad \text{(since disagreements occur only if } \mathcal{E}_{\text{bad}} \text{ happens)} \\
&\leq \left(1 + \mathbb{E}\left[|s_2 - s'_2|\sqrt{\frac{d}{2}} \,\middle|\, \mathcal{E}_{\text{bad}}, y = t\right]\right) \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] \qquad \text{(using Lemma 13)} \\
&= \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] + \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] \mathbb{E}\left[|s_2 - s'_2|\sqrt{\frac{d}{2}} \,\middle|\, \mathcal{E}_{\text{bad}}, y = t\right] \\
&= \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] + \mathbb{E}\left[|s_2 - s'_2|\sqrt{\frac{d}{2}} \,\middle|\, y = t\right].
\end{aligned} \tag{30}$$

Regarding the two terms above, first, we have that by Fact 7, (in particular the last part of the conclusion of that fact)

$$\mathbb{E}\left[|s_2 - s'_2| \mid y = t\right] = |\mathbb{E}[s_2] - \mathbb{E}[s_2]| = \frac{2E|t|}{B^2 + E^2 + \sigma^2}. \tag{31}$$

For the first term in Equation (30) we will use a union bound to relate it to the probabilities of $s_2 \neq s'_2$ and $s_1 \neq s'_1$. Then we note that the former is equal to the total variation distance between the distributions of s_2 and s'_2 (conditioned on $y = t$), which are the same distributions as the Gaussians in Equations (26) and (27). Similarly the probability of $s_1 \neq s'_1$ is the total variation distance between the Gaussians in Equations (28) and (29). Thus,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] &\leq \mathbb{P}\left[s_2 \neq s'_2 \mid y = t\right] + \mathbb{P}\left[s_1 \neq s'_1 \mid y = t\right] \\
&= \frac{2E|t|}{B^2 + E^2 + \sigma^2} + \frac{E\ B\ \mathbb{E}[|s_2 - s'_2| \mid y = t]}{B^2 + E^2} \\
&= \frac{2E|t|}{B^2 + E^2 + \sigma^2} + \frac{E\ B}{B^2 + E^2} \frac{2E|t|}{B^2 + E^2 + \sigma^2}.
\end{aligned}$$

Taking expectation over $t \sim \mathcal{N}(0, B^2 + E^2 + \sigma^2)$ (which is the distribution of the label values in our linear model) in the above inequality we obtain:

$$\mathbb{P}\left[\mathcal{E}_{\text{bad}}\right] \leq \frac{2E}{\sqrt{B^2+E^2+\sigma^2}} + \frac{E\ B}{B^2+E^2}\frac{2E}{\sqrt{B^2+E^2+\sigma^2}} \lesssim \frac{E}{\sigma}, \tag{32}$$

where the last step uses that $E \lesssim B, \sigma, E, B \geq 0$.

Combining Equations (30) to (32) and taking expectation over the label value $t \sim \mathcal{N}(0, B^2 + E^2 + \sigma^2)$ for the terms that we have not done it already, we have that the expected number of disagreeing coordinates in the second half of the vector is

$$\mathbb{E}\left[\sum_{i=1}^{d/2} \mathbb{1}(X_2(i) \neq X_2(i)')\right] \lesssim \frac{E}{\sqrt{B^2+E^2+\sigma^2}} + \frac{\sqrt{d}E}{\sqrt{B^2+E^2+\sigma^2}} \leq \frac{E}{\sigma} + \frac{\sqrt{d}E}{B}.$$

We now work similarly in order to bound the expected number of disagreements in the first half of the coordinates. First, conditioned on t we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{d/2} \mathbb{1}(X_1(i) \neq X_1(i)') \,\middle|\, y = t\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d/2} \mathbb{1}(X_1(i) \neq X_1(i)') \,\middle|\, \mathcal{E}_{\text{bad}}, y = t\right] \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] \\
&\qquad\qquad \text{(since disagreements occur only during } \mathcal{E}_{\text{bad}}) \\
&\leq \left(1 + \mathbb{E}\left[|s_1 - s_1'|\sqrt{\frac{d}{2}} \,\middle|\, \mathcal{E}_{\text{bad}}, y = t\right]\right) \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] \qquad \text{(using Lemma 13)} \\
&= \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] + \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] \mathbb{E}\left[|s_1 - s_1'|\sqrt{\frac{d}{2}} \,\middle|\, \mathcal{E}_{\text{bad}}, y = t\right] \\
&= \mathbb{P}\left[\mathcal{E}_{\text{bad}} \mid y = t\right] + \mathbb{E}\left[|s_1 - s_1'|\sqrt{\frac{d}{2}} \,\middle|\, y = t\right].
\end{aligned} \tag{33}$$

The probability term (after taking expectation over t) has been bounded in Equation (32). For the expectation term, we can again take into consideration that s_1, s_1' have been generated from the coupling of Fact 7, and because the means of the distributions are as shown in Equations (28) and (29), we have

$$\mathbb{E}\left[|s_1 - s_1'| \,\middle|\, y = t\right] = \mathbb{E}\left[\frac{B\ E\ |s_2 - s_2'|}{B^2+\sigma^2} \,\middle|\, y = t\right] = \frac{BE}{B^2+\sigma^2}\frac{2E|t|}{B^2+E^2+\sigma^2}. \tag{34}$$

Combining Equations (32) to (34) and taking expectation over $t \sim \mathcal{N}(0, B^2 + E^2 + \sigma^2)$ we have that

$$\mathbb{E}\left[\sum_{i=1}^{d/2} \mathbb{1}(X_1(i) \neq X_1(i)')\right] \lesssim \frac{E}{\sigma} + \frac{\sqrt{d}BE}{B^2+\sigma^2}\frac{E\ \mathbb{E}_{t\sim\mathcal{N}(0,B^2+E^2+\sigma^2)}[|t|]}{B^2+E^2+\sigma^2}$$

$$\lesssim \frac{E}{\sigma} + \frac{\sqrt{d}BE^2}{(B^2+\sigma^2)\sqrt{B^2+E^2+\sigma^2}} \lesssim \frac{E}{\sigma} + \frac{\sqrt{d}E}{B}.$$

This completes the proof of Lemma 28. ■

The proof of Theorem 27 is now complete. ■