Robust Statistical Estimators with Bounded Empirical Sensitivity

Valentio Iverson
Massachusetts Institute of Technology
Cambridge, MA, USA
viverson@mit.edu

Gautam Kamath
New York University
New York, NY, USA
g@csail.mit.edu

Argyris Mouzakis
The University of Texas at Austin
Austin, TX, USA
argymouz@cs.utexas.edu

Adam Smith
Boston University
Boston, MA, USA
ads22@bu.edu

Abstract

We introduce a new measure of robustness for statistical estimators, which we call *empirical sensitivity*. An estimator $\hat{\theta}$ has bounded empirical sensitivity if, with high probability over a dataset $X = (X_1, \ldots, X_n) \sim \mathcal{D}^{\otimes n}$, for any dataset Y obtained by modifying at most ηn points in X, we have that $\hat{\theta}(Y)$ is close to $\hat{\theta}(X)$.

We study bounds on this quantity for the prototypical problem of Gaussian mean estimation. We prove new lower bounds, showing that for any estimator $\hat{\mu}$ which achieves an optimal ℓ_2-error bound of $O(\sqrt{d/n})$, the empirical sensitivity is at least $\Omega(\eta + \sqrt{\eta d/n})$. The two terms arise due to obstructions on the mean and variance (via an Efron-Stein argument) of such an estimator. We show that this bound is tight up to logarithmic factors, by employing recent results for robust empirical mean estimation.

1 Introduction

Robustness is a classic topic of study within Statistics [Huber and Ronchetti, 2009]. It tries to understand and bound how much an estimator can fluctuate when data diverges from our assumptions. Such divergences could arise for innocuous reasons such as model misspecification, or more malicious reasons like an adversary manipulating or poisoning the dataset.

A popular style of robustness ensures that an estimator maintains accuracy after an adversary modifies a portion of the input dataset. In more detail, suppose we have an estimator $\hat{\theta}$ for a parameter θ^* of some family of distributions $\mathcal{D}_\theta$. Let $X = (X_1, \ldots, X_n) \sim \mathcal{D}_{\theta^*}$, and let Y be any dataset obtained by modifying ηn points in X. Robustness of the estimator $\hat{\theta}$ is then measured in terms of $\|\hat{\theta}(Y) - \theta^*\|$, the distance between the estimator on the contaminated dataset and the true parameter of interest.

While algorithms satisfying this definition are quite powerful, the notion does not capture some natural properties we might hope for an estimator to possess. We illustrate by considering the estimator $\hat{\mu}$ to be the median, and the dataset X to be sampled from a univariate Gaussian distribution $\mathcal{N}(\mu, 1)^{\otimes n}$. For an uncontaminated dataset X, the median is well known to achieve the near-optimal statistical rate $|\hat{\mu}(X) - \mu| = O_{\mathbb{P}}(1/\sqrt{n})$. The bounds below for contaminated datasets Y hold with high probability over X, simultaneously over all allowed Y (for η bounded away from $1/2$). Furthermore, the median is the canonical example of a robust statistic. For any dataset Y, formed by changing ηn entries of X, we have that $|\hat{\mu}(Y) - \mu| \leq O(\eta + 1/\sqrt{n})$—a mild (but

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

inherent) degradation of the rate from the uncontaminated case. However, we consider how much the estimator can shift due to contamination. For simplicity, focus on the case when $\eta = 1/n$, i.e., a single point is contaminated. The aforementioned robustness guarantee allows us only to say that $|\hat{\mu}(Y) - \hat{\mu}(X)| \le O(1/n + 1/\sqrt{n})$. This is a rather large radius for only a single contamination, with the dominant term being the parametric rate of $1/\sqrt{n}$. In fact, as we show in Section B, the median enjoys a quadratically stronger guarantee on this quantity: $|\hat{\mu}(Y) - \hat{\mu}(X)| \le O(1/n)$. To summarize, the conventional measures of robustness may not give insight into how *sensitive* the estimator may be.

Motivated by this deficiency, we introduce a new measure of robustness for statistical estimators, which we call the *empirical sensitivity*. Informally, we would like for an estimator $\hat{\theta}$ to simultaneously have the following two properties.

- Accuracy: Given a dataset $X \sim \mathcal{D}_{\theta^*}$, we have that $\|\hat{\theta}(X) - \theta^*\|$ is small.
- Bounded empirical sensitivity: For $X \sim \mathcal{D}_\theta$, and for any dataset Y obtained by modifying ηn entries of X, we have that $\|\hat{\theta}(Y) - \hat{\theta}(X)\|$ is small—either in expectation over X or with high probability.

Similar to traditional robustness, we focus on the case where the dataset X is drawn stochastically, but Y is an arbitrary or worst-case contamination of X. As usual, obtainable guarantees will depend on properties of the class of distributions $\mathcal{D}_\theta$. On the other hand, this can be seen as a *stronger* type of robustness: by triangle inequality, an accurate estimator with bounded empirical sensitivity will also enjoy a bound on $\|\hat{\theta}(Y) - \theta^*\|$, the traditional object of study in robust estimation.

The robustness of the estimated value to corruption has been considered before as a *worst-case* constraint on a function, notably in the design of differentially private algorithms. To our knowledge, ours is the first work that studies the interaction of empirical sensitivity with statistical efficiency, and also the first to explicitly study the average-case notion. See Related Work (Section 1.2) for more detail.

1.1 Results and Techniques

We initiate study of empirical sensitivity focusing on one of the most fundamental statistical tasks: *Gaussian mean estimation*. Our main result is the following lower bound on the empirical sensitivity of optimal estimators.

Theorem 1.1. *Consider any estimator $\hat{\mu}$ such that*

$$\sup_{\mu \in \mathbb{R}^d} \mathbb{E}_{X \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}} \|\hat{\mu}(X) - \mu\|_2^2 \lesssim \frac{d}{n}. \tag{1}$$

For every $\eta \in (0, 1)$,[1] *we have that*

$$\sup_{\mu \in \mathbb{R}^d} \left(\mathbb{E}_{X \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}} \left[\sup_{Y \colon d_H(X,Y) \le \lfloor \eta n \rfloor} \|\hat{\mu}(Y) - \hat{\mu}(X)\|_2^2 \right] \right)^{1/2} \ge \Omega\left(\eta + \sqrt{\frac{\eta d}{n}} \right).$$

This lower bound is tight up to logarithmic factors. Very recent work by Chen, Ding, Majid, and McKelvie [Chen et al., 2026] studies robust *empirical* mean estimation. An easy argument shows that their estimator satisfies near-optimal accuracy and also average-case empirical sensitivity (Theorem 3.2) essentially matching Theorem 1.1 above.

We recall that, for robust Gaussian mean estimation, the minimax expected ℓ_2-error rate under η-contamination is $\tilde{\Theta}\left(\eta + \sqrt{\frac{d}{n}}\right)$. By triangle inequality, this implies that one may have hoped for an empirical stability bound as small as $\widetilde{O}(\eta)$.[2] Instead, the larger bound of $\tilde{\Theta}\left(\eta + \sqrt{\frac{\eta d}{n}}\right)$ arises.

[1] Throughout, it is implicit that $\eta \ge 1/n$, i.e., there is at least one contaminated point.

[2] Note that triangle inequality only implies a lower bound of $\Omega(\eta)$ in the regime where $\eta = \Omega\left(\sqrt{d/n}\right)$. Indeed, proving the lower bound of η in all regimes appears to require significant technical work, and is quite different from how this term arises for conventional robust statistics, see Section 3 for more details.

Technically, the two terms in the lower bound are proven by showing obstructions on the mean and variance of the estimator.

The mean obstruction. We first reduce a d-dimensional uniformly accurate estimator to a scalar estimator along a direction and an orthogonal subspace. Then we prove scalar sensitivity lower bounds in two overlapping regimes. In the low corruption regime $\eta \lesssim \frac{1}{\sqrt{n}}$, we compare a valid local corruption, which shifts k random samples, to a statistically indistinguishable global dataset corruption that the estimator is forced to track due to its accuracy guarantees. For $\eta \gtrsim \frac{\log n}{n}$, a coupling argument compares two datasets from two nearby Gaussian means whose total variation distance is of order η, so that the coupled samples differ in at most $\lfloor \eta n \rfloor$ coordinates with high probability. The two regimes overlap for large n, covering the full range of values for η.[3]

The variance obstruction. The $\sqrt{\eta d/n}$ lower bound is driven by clean-sample variability. Indeed, we show that any uniformly MSE-accurate estimator must have output variance $\Omega(d/n)$ at some parameter. Using this, a block-resampling argument and vector Efron-Stein inequality convert this clean output variance into sensitivity to replacing an η fraction of the samples.

1.2 Related Work

Robust estimation is a vibrant area of study, particularly recent work with a focus on minimax statistical rates and computational efficiency in multivariate settings, see, e.g., [Diakonikolas et al., 2016, Lai et al., 2016] and [Diakonikolas and Kane, 2022] for a textbook treatment. As already described, empirical sensitivity is stronger than what is implied by traditional robustness guarantees.

One recent work of Chen, Ding, Majid, and McKelvie [Chen et al., 2026] studies differentially private Bayesian estimation, which they reduce to robust estimation of an empirical mean. While not originally conceived as such, their robust algorithm provides an estimator with bounded empirical sensitivity for a particular statistic that satisfies the desired accuracy guarantee (i.e., the empirical mean).

Stability has long been an object of study in statistics and machine learning, see, e.g., [Kearns and Ron, 1997, Bousquet and Elisseeff, 2000]. This generally focuses on understanding how the error of an algorithm changes if one removes a single point. Our focus instead is on the effect of modifying a *collection* of points (and also on the entire output, not just the error). The recent, contemporaneous work of Chakraborty, Luo, and Barber [Chakraborty et al., 2026] explores minimax risk when an estimator must have low *global sensitivity*—that is, *worst-case* empirical robustness with respect to a single corruption.

They suggest study of a distributional variant, where the dataset is sampled from a distribution belonging to some family, as a direction for future work.

Empirical stability also relates to concepts considered in the literature on differential privacy (DP) [Dwork et al., 2006]. Many DP mechanisms operate by adding noise calibrated to the worst-case sensitivity (called the *global sensitivity*) of a function. (This differs from our notion in that it is for worst-case neighboring datasets and considers only a single corruption.) The literature also considers the *local sensitivity* [Nissim et al., 2007]: the sensitivity of the function at a particular dataset. In particular, our definition can be viewed as an average-case bound on *local sensitivity with step size* s (Definition 4.4 of the full version of Nissim et al. [2007]) for $s = \eta n$.

A similar notion is the *inverse sensitivity* at a data set X, employed in the inverse sensitivity mechanism [Mir et al., 2011, Johnson and Shmatikov, 2013, Asi and Duchi, 2020a,b]. This measures how many corruptions to X would be needed to change the value of an estimator from $f(X)$ to a given target value z. Empirical sensitivity provides an average-case bound on this quantity. That connection was used implicitly in some works relating robustness to privacy, such as Hopkins et al. [2023], Asi et al. [2023].

Several other "local" variants of sensitivity have been considered, for example *down sensitivity* [Chen and Zhou, 2013, Raskhodnikova and Smith, 2015] (related to *resilience* in robust statistics Steinhardt et al. [2018]), which only considers neighboring datasets where points have been *removed*. We believe the full extent of connections between empirical sensitivity and DP are yet to be understood.

[3] In Section 4, we also show this lower bound for the conceptually simpler case of Bernoulli parameter estimation.

Finally, the recent work of Trillos, Jaffe, and Sen [Trillos et al., 2025] introduces a new notion of sensitivity, which measures change in a statistic when infinitesimal Gaussian noise is added to each datapoint. Similar to our notion, it is an average-case (as opposed to worst-case) quantity, but the different contamination models appear to capture different phenomena. In particular, their notion may connect more closely with the literature on measurement error and local differential privacy Warner [1965], Evfimievski et al. [2003], Kasiviswanathan et al. [2011].

2 Problem Setup and Definitions

General Notation. We use the notation $A \lesssim B$ to mean that $A \leq CB$ for a universal constant $C > 0$, and $A \gtrsim B$ analogously. We write $A \asymp B$ if both $A \lesssim B$ and $B \lesssim A$ hold. We also note that $\widetilde{O}(\cdot)$ suppresses polylogarithmic factors in the relevant problem parameters. Throughout the whole of this work, we assume that $d, n \geq 1$ are fixed, while all norms are the standard Euclidean norms. We use $O(\cdot)$ and $\Omega(\cdot)$ for upper and lower bounds up to constant factors, respectively, and $\Theta(\cdot)$ for matching upper and lower bounds. These agree with $\lesssim$, $\gtrsim$, and $\asymp$, respectively. Their $\widetilde{O}$, $\widetilde{\Omega}$, and $\widetilde{\Theta}$ variants suppress polylogarithmic factors. For random variables Z_n, the notation $Z_n = O_{\mathbb{P}}(a_n)$ means that Z_n/a_n is bounded in probability. Constants with a subscript, as in $\Omega_{C_{\mathsf{MSE}}}$, may depend on the indicated parameter. For two datasets $x, x' \in (\mathbb{R}^d)^n$, let $d_H(x, x') \coloneqq \#\{i \in [n] \colon x_i \neq x'_i\}$ denote the Hamming distance at the observation level. Thus, $d_H(x, x') \leq k$ means that x' can be obtained from x by replacing at most k observations by arbitrary vectors in $\mathbb{R}^d$. Given a sample space $\mathcal{X}$, we denote the set of all distributions over $\mathcal{X}$ by $\Delta(\mathcal{X})$. For any distribution $\mathcal{D} \in \Delta(\mathcal{X})$, $\mathcal{D}^{\otimes n}$ denotes the product distribution over $\mathcal{X}^n$, where each marginal is $\mathcal{D}$. Given a dataset $X = (X_1, \ldots, X_n)$ drawn i.i.d. from $\mathcal{D}$, we write $X \sim \mathcal{D}^{\otimes n}$. Whenever using the symbols of probability and expectation, we use a subscript to denote what the randomness is over in cases where it might not be clear from the context, e.g., $\mathbb{E}_X[\cdot]$.

Let P, Q be probability measures on a common measurable space. If $Q \ll P$, we define the χ^2-divergence by $\chi^2(Q\|P) := \mathbb{E}_P\left[\left(\frac{dQ}{dP} - 1\right)^2\right]$. The total variation distance is defined by $\mathrm{TV}(P, Q) := \sup_A |P(A) - Q(A)|$.

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a measurable scalar estimator. For each of the following, $\mu' \in \mathbb{R}$ will be fixed and we assume the data is drawn as $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu', 1)^{\otimes n}$. We use the following notation throughout: $a(\mu') := \mathbb{E}_X[g(X)]$ is the mean of the estimator; and $b(\mu') := a(\mu') - \mu'$ is the bias of the estimator.

For $\eta \in (0, 1)$, we consider the η-corruption model where the adversary is allowed to contaminate any of $k \coloneqq \lfloor \eta n \rfloor$ points from the given n sample points.

Definition 2.1 (Uniform MSE Accuracy)**.** Let $\mathcal{D}$ be a distribution class parameterized by $\theta \in \Theta$, and let $D_\theta \in \mathcal{D}$. We assume that f is an estimator that takes n samples drawn i.i.d. from D_θ and outputs an estimate of θ. We say that f is $(C_{\mathsf{MSE}}, \gamma)$*-uniformly MSE-accurate* if

$$\sup_{\theta \in \Theta} \mathbb{E}_{X \sim D_\theta^{\otimes n}}\left[\|f(X) - \theta\|^2\right] \leq C_{\mathsf{MSE}}\gamma. \tag{2}$$

In the above definition, the error rate includes two parameters. These are C_{MSE} and γ. To give a bit more context about this, C_{MSE} should be thought of as an absolute constant, whereas γ should be treated as a component of the rate that may depend on various aspects of the class of distributions under consideration (e.g., the dimension of the data), as well as the number of samples used by our estimator.

Definition 2.2 (Pointwise empirical sensitivity Nissim et al. [2007])**.** For $\eta \in (0, 1)$, we define the η-corruption pointwise empirical sensitivity of an estimator f at dataset X as:

$$S_\eta^f(X) \coloneqq \sup_{X' \colon d_H(X, X') \leq \lfloor \eta n \rfloor} \|f(X) - f(X')\|.$$

The pointwise definition of empirical sensitivity was considered previously, under the name *local sensitivity at step size* $\lfloor \eta n \rfloor$.

Definition 2.3 (Distributional empirical sensitivity)**.** Let $\mathcal{D}$ be a family of distributions parameterized by $\theta \in \Theta$. For $D_\theta \in \mathcal{D}, q \in \{1, 2\}$, and $\eta \in (0, 1)$, we define the η-corruption L^q-distributional empirical sensitivity of an estimator f at distribution D_θ as

$$\mathsf{ES}_{\eta,q}(f; \theta) \coloneqq \left(\mathbb{E}_{X \sim D_\theta^{\otimes n}} \left[(S_\eta^f(X))^q \right] \right)^{1/q}.$$

Thus $\mathsf{ES}_{\eta,1}$ is expected pointwise sensitivity and $\mathsf{ES}_{\eta,2}$ is root-mean-square pointwise sensitivity.

2.1 Adversary Models

The robustness literature frequently studies the relative power of different adversaries [Blanc et al., 2022, Blanc and Valiant, 2025, Blanc et al., 2026, Ben-David et al., 2023, Lechner et al., 2025]. We briefly discuss various adversary models, listed here in increasing amount of strength.

Definition 2.4. Let a dataset $X = (X_1, \ldots, X_n) \sim \mathcal{D}^{\otimes n}$, where $\mathcal{D}$ is a probability distribution. Fix a corruption fraction $\eta \in (0, 1)$ and define $k \coloneqq \lfloor \eta n \rfloor$ to be the number of contaminated points.[4]

1. *Resampling adversary*. A subset $I \subseteq [n]$ of size k is chosen uniformly at random, and the samples $\{X_i : i \in I\}$ are replaced by fresh independent samples from the same clean distribution as X. This is the weakest adversary: both the removed points and the inserted points are stochastic and drawn from the clean distribution.

2. *Stochastic adversary*. A subset $I \subseteq [n]$ of size k is chosen uniformly at random, and the samples $\{X_i : i \in I\}$ are replaced by independent samples from a fixed replacement distribution $\mathcal{Q}$ chosen before seeing X. The replacement distribution may differ from the clean distribution, but it is not chosen adaptively after observing the realized dataset.

3. *Adaptive adversary*. After observing the clean dataset X, the adversary may choose both the corrupted indices and the replacement values arbitrarily, subject only to the constraint $d_H(X, Y) \leq k$. This is the adversary encoded by $S_\eta^f(X)$.

Our discussion thus far has focused on the adaptive adversary. This is reflected in our main result. The result focuses on the class of Gaussian distributions with identity covariance $\{\mathcal{N}(\mu, \mathbb{I}_d)\}_{\mu \in \mathbb{R}^d}$, and shows a lower bound on

$$\sup_{\mu \in \mathbb{R}^d} \mathsf{ES}_{\eta,2}(f; \mu) = \sup_{\mu \in \mathbb{R}^d} \left(\mathbb{E}_{X \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}} \left[(S_\eta^f(X))^2 \right] \right)^{1/2}.$$

The two terms in the lower bound of Theorem 1.1 come from different adversaries. The variance-obstruction term of $\Omega(\sqrt{\eta d/n})$ arises even for the weakest adversary, the resampling adversary. On the other hand, our proof for the mean-obstruction term of $\Omega(\eta)$ requires a stronger adversary.

As we show in Section C, more empirically stable estimators exist under the (weakest) resampling adversary. In this case, the empirical mean achieves an empirical stability of $O(\sqrt{\eta d/n})$, avoiding the $\Omega(\eta)$ term that is necessary under the adaptive adversary.

As the empirical sensitivity is a quantity measured with respect to a realized dataset, all our adversaries are defined as modifying a dataset, rather than a distribution (as is sometimes done in the conventional robustness literature). Our adaptive adversary is the same as what is usually called the *strong contamination* model in the robustness literature [Diakonikolas and Kane, 2022]. The stochastic adversary can be seen as the moral equivalent of Huber contamination [Huber, 1964]. The resampling adversary has no equivalent, as, by construction, the underlying distribution is uncontaminated.

3 Main Results

In this section, we give the bulk of our results. We will be working with the class of Gaussian distributions with unknown mean μ and known covariance matrix Σ which, without loss of generality,

[4] One could consider broader classes of adversaries. For example, an adversary who adaptively chooses k points adversarially, and then resamples them obliviously. Or an adversary who obliviously chooses k points and then replaces them arbitrarily. We omit discussion of such adversaries to streamline the presentation.

we assume to be equal to the identity matrix $\mathbb{I}$. We will consider mean estimators with bounded empirical sensitivity for this class, and give both upper and lower bounds.

The following theorem is the main result of this paper: a tight lower bound of the empirical sensitivity of accurate estimators, complemented by a recent upper bound result in Chen et al. [2026]:

Theorem 3.1. *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and $\left(C_{\mathsf{MSE}}, \frac{d}{n}\right)$-uniformly MSE-accurate for the class of Gaussians $\{\mathcal{N}(\mu, \mathbb{I})\}_{\mu\in\mathbb{R}^d}$. Then for every $\eta \in (0,1)$ and all sufficiently large n depending only on C_{MSE}, we have*

$$\sup_{\mu\in\mathbb{R}^d} \mathsf{ES}_{\eta,2}(f;\mu) \geq \Omega_{C_{\mathsf{MSE}}}\left(\eta + \sqrt{\frac{\eta d}{n}}\right).$$

The lower bound comprises two independent components: a mean obstruction, which gives the η term, and a variance obstruction, which gives the $\sqrt{\eta d/n}$ term. Each part captures the estimator's vulnerability to a different adversarial mechanism. The mean obstruction reflects the estimator's sensitivity to small shifts in the underlying distribution's mean, whereas the variance obstruction isolates the estimator's inherent variability when resampled from the same clean distribution.

3.1 Upper bound from robust empirical-mean recovery

Our lower bound is complemented by a high-probability upper bound, which is a sensitivity consequence of an estimator that remains close to the clean empirical mean simultaneously over all allowed corruptions. The following is a consequence of a result of Chen, Ding, Majid, and McKelvie [Chen et al., 2026].

Theorem 3.2. *Fix $\eta \in (0, 1/3)$ and $\beta \in (0, 1/2)$, there exists a measurable estimator $f_{\eta,\beta} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ such that, for every $\mu \in \mathbb{R}^d$, with probability at least $1-\beta$ over $X \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}$, both of the following hold:*

$$\|f_{\eta,\beta}(X) - \mu\|_2^2 \leq O\left(\frac{d + \log(1/\beta)}{n}\right) \quad \textit{and} \quad S_\eta^{f_{\eta,\beta}}(X) \leq \widetilde{O}\left(\eta + \sqrt{\eta \cdot \frac{d + \log(1/\beta)}{n}}\right).$$

Proof sketch. The primitive from Chen et al. [2026] gives a high-probability event on which, simultaneously for every η-corruption X' of the clean dataset X, every feasible cleaned candidate has empirical mean close to the empirical mean $\overline{X}$. We modify the estimator so that, on clean input, it chooses the feasible candidate with minimum Hamming distance to the input. Therefore, one can show that on the same good event, $f_{\eta,\beta}(X) = \overline{X}$. The complete proof can be found in Section E. ■

Remark 3.3 (Efficient versus inefficient upper bounds). While the estimator of Theorem 3.2 should be viewed as the statistically sharp upper bound needed to match our lower bound up to logarithmic factors, it is not computationally efficient. The same work also gives an efficient estimator with a worse empirical-mean recovery rate. Applying the same triangle-inequality argument as above yields an efficient estimator for our empirical-sensitivity problem with sensitivity of order $\widetilde{O}\left(\eta + \sqrt{\eta \cdot \sqrt{\frac{d}{n}}}\right)$ with high probability, together with clean-data MSE accuracy.

Their efficient estimator is complemented by a computational lower bound against restricted class of algorithms. Informally, they show that one can not do better than this weaker $O\left(\sqrt{\eta\sqrt{\frac{d}{n}}}\right)$ rate with any algorithm implemented by low-degree polynomials. We note that this does *not* imply a computational lower bound for our setting, as their setting is specific to the empirical mean, whereas the problem we are considering allows any estimator which is MSE accurate.

The computational question is not restricted to estimators that recover the empirical mean. For trimmed means, Huber-type estimators, and filtering-based robust estimators [Lai et al., 2016, Diakonikolas et al., 2016], classical contamination-risk guarantees do not by themselves establish optimal empirical sensitivity. We do not rule out other MSE-optimal estimators that attain the optimal sensitivity efficiently; in particular, the low-degree lower bound for empirical-mean recovery does not establish such a lower bound for all accurate estimators.

It remains open whether efficient, uniformly accurate estimators can achieve the statistically optimal empirical-sensitivity scale $\eta + \sqrt{\frac{\eta d}{n}}$, or whether the worse efficient rate $\eta + \sqrt{\eta\sqrt{\frac{d}{n}}}$ is inherent for computationally-efficient estimators under our empirical sensitivity condition.

3.2 Mean obstruction

We establish the mean obstruction by separating the analysis into two distinct regimes that collectively cover the entire spectrum of corruption levels. In the low-corruption regime, where $\eta = O(1/\sqrt{n})$, our argument relies on bounding the χ^2-divergence between two distinct adversaries: a local shift of exactly k random samples by a constant δ (which represents a valid adversarial η-contamination), and a global shift of all samples by $\eta\delta$ (which represents a genuine change in the underlying parameter). Because the χ^2-divergence between these two distributions scales as $O(k^2\delta^4/n)$, choosing a sufficiently small δ renders the adversarial contamination statistically indistinguishable from a true parameter shift. Since the MSE-accuracy guarantee forces the estimator to track the global shift, this indistinguishability ensures the estimator is similarly displaced by the local corruption, yielding the $\Omega(\eta)$ lower bound.

More precisely, after averaging over the prior, the integrated risk and χ^2 comparison give an error of order $O(\eta\delta^2)$ in tracking the global shift of size $\eta\delta$. Choosing $\delta > 0$ as a sufficiently small constant leaves a displacement of order η.

For the high-corruption regime, where $\eta = \Omega(\log n/n)$, we shift to a coordinate-wise maximal coupling argument. We couple two Gaussian distributions whose means are separated by exactly η. By showing that samples drawn from these coupled distributions differ in at most k coordinates with high probability in this regime, we show that an η-bounded adversary can seamlessly bridge the two datasets. Since the estimator's accuracy forces its mean response to be displaced by $\Omega(\eta)$ between these two parameter values, the empirical sensitivity is similarly bounded below. Crucially, these two regimes are not isolated; they overlap for all sufficiently large n, ensuring that our combined lower bound fully covers the entire spectrum of corruption allowances without any gaps.

To prove both high-dimensional theorems, we reduce them to the single-dimensional case using a projection and conditioning lemma. Informally, any uniformly MSE-accurate estimator $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ can be projected along a carefully chosen direction to yield a one-dimensional estimator $g\colon \mathbb{R}^n \to \mathbb{R}$ with scalar parameter μ' that inherits the $O(1/n)$ Bayes risk, allowing us to lower-bound the high-dimensional sensitivity $\mathsf{ES}_{\eta,2}(f;\mu)$ via the scalar empirical sensitivity $S^g_\eta(X)$.

Theorem 3.4 (Mean obstruction: low-corruption regime)**.** *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and $\big(C_{\mathsf{MSE}}, \frac{d}{n}\big)$-uniformly MSE-accurate for the class of Gaussians $\{\mathcal{N}(\mu, \mathbb{I})\}_{\mu\in\mathbb{R}^d}$. There exist constants $\kappa = \kappa(C_{\mathsf{MSE}}) > 0$ and $C_3 = C_3(C_{\mathsf{MSE}}) > 0$ such that, for every $\eta \in (0,1)$ such that $1 \le \lfloor \eta n \rfloor \le \kappa\sqrt{n}$ and for all sufficiently large n, we have*

$$\sup_{\mu\in\mathbb{R}^d} \mathsf{ES}_{\eta,2}(f;\mu) \ge C_3\eta.$$

Proof sketch. We establish the low-corruption lower bound through an indistinguishability argument between two adversaries. As previously stated, we may assume that $\mu' \in [0,1]$ and thus it suffices to analyze a clipped scalar estimator $h : \mathbb{R}^n \to [0,1]$.

1. First, because the estimator inherits a low $O(1/n)$ integrated risk under a smooth prior, its expected output is forced to accurately track the true parameter. Specifically, if the entire data distribution is subjected to a global mean shift of $\eta\delta$, the estimator's mean output must naturally shift by approximately $\eta\delta$ to maintain its low average risk. We note that this global shift modifies *all* n points, meaning that it is *not* a valid adversarial η-contamination.
2. Next, we consider a local shift: an adversary selects a uniformly random subset of exactly $k := \lfloor \eta n \rfloor$ coordinates and shifts them by δ.[5] Because this modifies at most k samples, it *is* a valid η-corruption. In the low-corruption regime $k = O(\sqrt{n})$, we bound the χ^2-divergence

[5] In Section 2.1 we defined the stochastic adversary as replacing part of the dataset with points drawn independently from another distribution. In the present case, the points are not resampled from scratch, but the δ-shift results in a Gaussian distribution with different mean. While technically not exactly the same, this is arguably morally equivalent to a stochastic adversary.

between the global shift and the local shift by $O(k^2\delta^4/n)$. By choosing the constant δ to be sufficiently small, this divergence becomes of lower order than the estimator's precision, making the valid local corruption statistically indistinguishable from the global shift.

3. Since the estimator is forced to track the global shift to preserve its accuracy, and it cannot distinguish the valid local corruption from the global shift, the adversary's valid η-corruption forces the estimator to have a large displacement of $\Omega(\eta\delta) = \Omega(\eta)$. Taking the supremum over the prior establishes the $\Omega(\eta)$ empirical sensitivity lower bound.

The complete proof can be found in Section F. ■

Theorem 3.5 (Mean obstruction: high-corruption regime)**.** *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and $\big(C_{\mathsf{MSE}}, \frac{d}{n}\big)$-uniformly MSE-accurate for the class of Gaussians $\{\mathcal{N}(\mu, \mathbb{I})\}_{\mu\in\mathbb{R}^d}$. There exist constants $C_{\mathsf{high}} = C_{\mathsf{high}}(C_{\mathsf{MSE}}) > 0$ and $C_4 = C_4(C_{\mathsf{MSE}}) > 0$ such that, for all sufficiently large n and every $\eta \in (0, \frac{1}{10})$[6] such that $\lfloor \eta n \rfloor \geq C_{\mathsf{high}} \log n$, we have*

$$\sup_{\mu\in\mathbb{R}^d} \mathsf{ES}_{\eta,2}(f;\mu) \geq C_4\eta.$$

Proof sketch. We establish the high-corruption lower bound through a coupling argument. As previously stated, we may assume that $\mu' \in [0,1]$ and thus it suffices to analyze a clipped scalar estimator $h\colon \mathbb{R}^n \to [0,1]$.

1. We exploit the estimator's required accuracy. Because h maintains $O(1/n)$ integrated risk over $[0,1]$, its expected output must track the true parameter at the boundaries of the interval. We show that evaluating the estimator across the interval forces its mean response to increase by $\Omega(\eta)$ on average when the underlying distribution is shifted by η.
2. We construct our adversary using coordinate-wise maximal coupling as in Theorem A.4. For any parameter μ', we couple samples from $\mathcal{N}(\mu', 1)^{\otimes n}$ and $\mathcal{N}(\mu' + \eta, 1)^{\otimes n}$. The expected number of different coordinates will be governed by the TV distance between their Gaussian marginals, which is strictly less than η. Therefore, for $k = \Omega(\log n)$, a Chernoff bound (Theorem A.5) can guarantee that the samples differ by at most k points with high probability. Averaging the displacement across the parameter interval establishes the $\Omega(\eta)$ empirical sensitivity lower bound.

The complete proof can be found in Section G. ■

3.3 Variance obstruction

The variance obstruction is rooted in a fundamental statistical trade-off: any estimator achieving uniform MSE accuracy must exhibit an output variance of order $\Omega(d/n)$ at some parameter point. An adversary can thus exploit this inherent variability, leveraging the estimator's own structural fluctuations to induce empirical sensitivity. This is made precise by the following theorem.

Theorem 3.6 (Variance obstruction)**.** *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and $\big(C_{\mathsf{MSE}}, \frac{d}{n}\big)$-uniformly MSE-accurate for the class of Gaussians $\{\mathcal{N}(\mu, \mathbb{I})\}_{\mu\in\mathbb{R}^d}$. There is a universal constant $C_5 > 0$ such that for every $\eta \in (0,1)$, we have*

$$\sup_{\mu\in\mathbb{R}^d} \mathsf{ES}_{\eta,2}(f;\mu) \geq C_5\sqrt{\frac{\eta d}{n}}.$$

Proof sketch. The proof of the variance obstruction proceeds in three stages: establishing a scalar variance bound, lifting it to high dimensions, and converting variance into empirical sensitivity.

1. First, we show that any one-dimensional uniformly MSE-accurate estimator must have output variance $\Omega(1/n)$ at some parameter. Assume that the one-dimensional estimator is g with parameter μ'. Let $a(\mu') \coloneqq \mathbb{E}_X[g(X)]$. Uniform MSE accuracy thus implies $|a(\mu') - \mu'| \lesssim n^{-1/2}$. We consider an interval of large length which we assume the parameter μ' lies in, and we partition this interval such that each part is roughly of size

[6] The $\frac{1}{10}$ is an artifact of the technical aspects of the proof, and could potentially be boosted to a larger constant.

$n^{-1/2}$, i.e., $\mu'_j = jn^{-1/2}$. Then $a(\mu')$ must move by a constant fraction of the interval length across the grid. For adjacent grid points, the two distributions $\mathcal{N}(\mu'_j, 1)^{\otimes n}$ and $\mathcal{N}(\mu'_{j+1}, 1)^{\otimes n}$ have bounded χ^2-divergence. One can then use Hammersley-Chapman-Robbins inequality (Theorem A.2) to get

$$\mathrm{Var}_{\mu'_j}(g) \gtrsim (a(\mu'_{j+1}) - a(\mu'_j))^2.$$

Summing over j and applying Cauchy-Schwarz forces the average scalar variance over the grid to be $\Omega(1/n)$.

2. Second, we lift this scalar variance lower bound to d dimensions by applying this one-dimensional finite-difference argument along each coordinate direction, with all other mean coordinates fixed on a common finite grid. Doing this gives us some parameter μ^* such that $\mathrm{Var}_{\mu^*}(f(X)) \gtrsim \frac{d}{n}$.
3. Finally, we convert this clean-sample variance bound into empirical sensitivity by block resampling. Work on parameter μ^*. Partition the sample into blocks of size $k := \lfloor \eta n \rfloor$. Let $X^{(i)}$ be obtained from X by replacing the ith block by an independent fresh block from the same distribution $\mathcal{N}(\mu^*, \mathbb{I}_d)$. Each $X^{(i)}$ differs from X in at most k samples, and so pointwise

$$S^f_\eta(X) \geq \|f(X^{(i)}) - f(X)\|.$$

The vector Efron–Stein inequality (Theorem A.1) then gives $\mathsf{ES}_{\eta,2}(f; \mu^*) \gtrsim \sqrt{\frac{\eta d}{n}}$, as desired.

The complete proof can be found in Section H. ■

4 Related Bernoulli Results

The same empirical-sensitivity question can be asked for Bernoulli mean estimation. We record a simple one-dimensional analogue of the mean obstruction.

Indeed, for $x \in \{0,1\}^n$, we let $|x| := \sum_{i=1}^n x_i$ denote its Hamming weight. For a scalar estimator $f\colon \{0,1\}^n \to \mathbb{R}$, define

$$S^f_\eta(x) := \sup_{y \in \{0,1\}^n\colon\, d_H(x,y) \leq \lfloor \eta n \rfloor} |f(y) - f(x)|.$$

Theorem 4.1. *Let $f\colon \{0,1\}^n \to \mathbb{R}$ be measurable and suppose that*

$$\sup_{p \in [0,1]} \mathbb{E}_{X \sim \mathsf{Bern}(p)^{\otimes n}} |f(X) - p| \leq \frac{C_{\mathsf{Bern}}}{\sqrt{n}}.$$

Then for every $\eta \in (0,1)$ such that $k = \lfloor \eta n \rfloor \geq 1$, and for all sufficiently large n depending on C_{Bern}, we have

$$\sup_{p \in [0,1]} \mathbb{E}_{X \sim \mathsf{Bern}(p)^{\otimes n}} [S^f_\eta(X)] \geq C_6 \eta.$$

Proof sketch. We will show the empirical sensitivity lower bound through a layer-telescoping style argument on the Boolean hypercube. As shown previously, clipping preserves accuracy and only decreases sensitivity, so it suffices to analyze a bounded scalar estimator $f\colon \{0,1\}^n \to [0,1]$.

1. We exploit the estimator's required accuracy. By decomposing the uniform prior on the Bernoulli parameter into discrete Hamming-weight layers, we can evaluate the estimator's expected output conditionally per layer. Because f maintains $O(1/\sqrt{n})$ expected risk, binomial anti-concentration bounds guarantee that its expected output at widely separated boundary layers must closely track the true proportion of ones. This forces an $\Omega(\eta)$ gap in the estimator's mean between these boundaries.
2. We then distribute this global gap to local neighborhoods using a transport coupling. We couple elements between layers separated by a distance $\ell = \Theta(\eta n)$ by uniformly flipping exactly ℓ zero coordinates to ones, ensuring the coupled instances are strictly within the allowed Hamming radius $\lfloor \eta n \rfloor$. This then shows that the required boundary gap must be bridged by significant intermediate jumps. Averaging this forced variation across the discrete layers establishes the strict $\Omega(\eta)$ lower bound on expected empirical sensitivity.

The complete proof can be found in Section I. ■

The Bernoulli case also has an immediate matching upper bound: the empirical-mean estimator $f_{\mathrm{mean}}(x) = \frac{1}{n}\sum_i x_i$ satisfies $S_\eta^{f_{\mathrm{mean}}}(x) \le \lfloor \eta n \rfloor / n \le \eta$. Thus, this simpler case isolates the mean-obstruction phenomenon before the Gaussian analysis.

5 Conclusion

We introduced the notion of *empirical sensitivity* as a data-dependent measure of robustness for statistical estimators. Unlike traditional contamination robustness, which asks whether an estimator remains close to the underlying parameter after corruptions, empirical sensitivity asks whether the estimator remains close to its own clean-data output under local modifications of the realized sample.

For Gaussian mean estimation, we characterized the optimal empirical-sensitivity scale, up to logarithmic factors, among estimators that retain the optimal clean-data MSE rate. Our lower bound shows that every such estimator must incur empirical sensitivity at least $\Omega\left(\eta + \sqrt{\eta d/n}\right)$. The two terms arise from different mechanisms: a mean obstruction, showing that accurate estimators must respond to small distributional shifts, and a variance obstruction, showing that clean-sample variability can be converted into sensitivity through resampling. We also showed that this rate is essentially achievable using a recent robust empirical-mean recovery result. Finally, we also complemented the theory for Gaussian with a lower bound for the analogous questions for Bernoulli.

Several questions remain open. First, while the statistically optimal upper bound follows from a computationally inefficient empirical-mean recovery primitive, it is unclear whether efficient estimators can achieve the same empirical-sensitivity rate. Second, our results focus on Gaussian and Bernoulli mean estimation; extending the theory to broader statistical models, such as covariance estimation and other parameter estimation problems. Other open questions include the sensitivity–accuracy tradeoff when the clean-data MSE requirement is relaxed, whether the upper-bound logarithmic factors are necessary, and what happens under adversaries intermediate between resampling and adaptive contamination.

More broadly, empirical sensitivity suggests a refined perspective on robustness: beyond asking whether an estimator remains accurate after contamination, one can ask how stable its output is around typical clean datasets.

Acknowledgments

The authors are listed in alphabetical order. Valentio Iverson was supported by the University of Waterloo through an MURA and a URF project. Gautam Kamath was supported by a Canada CIFAR AI Chair, an NSERC Discovery Grant, and an Ontario Early Researcher Award. Argyris Mouzakis was supported by an Ontario Early Researcher Award and a David R. Cheriton Graduate Scholarship. Adam Smith was supported in part by US NSF award 2232694 and a gift from Apple Research.

References

Hilal Asi and John C Duchi. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020a.

Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 14106–14117. Curran Associates, Inc., 2020b.

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 1121–1146. JMLR, Inc., 2023.

Shai Ben-David, Alex Bie, Gautam Kamath, and Tosca Lechner. Distribution learnability and robustness. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23, pages 52732–52758. Curran Associates, Inc., 2023.

Guy Blanc and Gregory Valiant. Adaptive and oblivious statistical adversaries are equivalent. In *Proceedings of the 57th Annual ACM Symposium on the Theory of Computing*, STOC '25, pages 2031–2042. ACM, 2025.

Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. On the power of adaptivity in statistical adversaries. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 5030–5061, 2022.

Guy Blanc, Yizhi Huang, Tal Malkin, and Rocco A Servedio. Is nasty noise actually harder than malicious noise? In *Proceedings of the 37th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '26, pages 6767–6787. SIAM, 2026.

Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In *Advances in Neural Information Processing Systems 13*, NIPS '00, pages 196–202. Curran Associates, Inc., 2000.

Abhinav Chakraborty, Yuetian Luo, and Rina Foygel Barber. Stability and accuracy trade-offs in statistical estimation. *arXiv preprint arXiv:2601.11701*, 2026.

Shixi Chen and Shuigeng Zhou. Recursive mechanism: Towards node differential privacy and unrestricted joins. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 653–664. ACM, 2013.

Sitan Chen, Jingqiu Ding, Mahbod Majid, and Walter McKelvie. Computation-utility-privacy tradeoffs in Bayesian estimation. In *Proceedings of the 58th Annual ACM Symposium on the Theory of Computing*, STOC '26. ACM, 2026.

Ilias Diakonikolas and Daniel Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2022.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 655–664. IEEE Computer Society, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 211–222. ACM, 2003.

Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23, pages 497–506. ACM, 2023.

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

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2009.

Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1079–1087. ACM, 2013.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the 10th Annual Conference on Learning Theory*, COLT '97, pages 152–162, 1997.

Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 665–674. IEEE Computer Society, 2016.

Tosca Lechner, Alex Bie, and Gautam Kamath. On the learnability of distribution classes with adaptive adversaries. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML '25. JMLR, Inc., 2025.

Darakhshan Mir, Shan Muthukrishnan, Aleksandar Nikolov, and Rebecca N Wright. Pan-private algorithms via statistics on sketches. In *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '11, pages 37–48. ACM, 2011.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, STOC '07, pages 75–84. ACM, 2007.

Ronald Pyke. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(3): 395–449, 1965.

Sofya Raskhodnikova and Adam Smith. Private analysis of graph data. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*, pages 1–6. Springer, Berlin, Heidelberg, 2015.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 45:1–45:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Nicolás García Trillos, Adam Quinn Jaffe, and Bodhisattva Sen. Wasserstein-Cramr-Rao theory of unbiased estimation. *arXiv preprint arXiv:2511.07414*, 2025.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

A Standard Facts

We collect the standard facts used throughout the paper.

Theorem A.1 (Efron–Stein inequality)**.** *Suppose that $X_1, \dots, X_n, X_1', \dots, X_n'$ are independent, with X_i' distributed identically to X_i for every i. Let $X = (X_1, \dots, X_n)$ and*

$$X^{(i)} = (X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n), \quad i = 1, \dots, n.$$

Then, for every scalar function f with finite variance,

$$\mathrm{Var}(f(X)) \le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[(f(X) - f(X^{(i)}))^2 \right].$$

More generally, if f is vector-valued and $\mathbb{E}\|f(X)\|_2^2 < \infty$, then

$$\mathbb{E}\left[\|f(X) - \mathbb{E}f(X)\|_2^2 \right] \le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\|f(X) - f(X^{(i)})\|_2^2 \right].$$

Theorem A.2 (Hammersley–Chapman–Robbins inequality)**.** *Let P, Q be probability measures with $Q \ll P$ and $\chi^2(Q\|P) < \infty$. Let T be a statistic such that $T \in L^2(P)$ and $T \in L^1(Q)$. Then*

$$\mathrm{Var}_P(T) \ge \frac{(\mathbb{E}_Q T - \mathbb{E}_P T)^2}{\chi^2(Q\|P)}.$$

Theorem A.3 (Cramér–Rao inequality)**.** *Let $\{P_{\mu'}\}_{\mu' \in \Theta}$ be a one-dimensional parametric family with densities $p_{\mu'}$. Assume the family and a given statistic $T(X)$ satisfy the following regularity conditions:*

- *The support $\{x : p_{\mu'}(x) > 0\}$ is independent of the parameter μ'.*
- *The score function $s_{\mu'}(X) \coloneqq \frac{\partial}{\partial \mu'} \log p_{\mu'}(X)$ exists almost everywhere.*
- *The Fisher information $I(\mu') \coloneqq \mathbb{E}_{\mu'}[s_{\mu'}(X)^2]$ satisfies $0 < I(\mu') < \infty$.*
- *The statistic T has finite variance, its mean response $m_T(\mu') \coloneqq \mathbb{E}_{\mu'}[T]$ is differentiable, and it satisfies the score identity $m_T'(\mu') = \mathbb{E}_{\mu'}[T(X) s_{\mu'}(X)]$.*

Then, the variance of the statistic is lower bounded by

$$\mathrm{Var}_{\mu'}(T) \ge \frac{(m_T'(\mu'))^2}{I(\mu')}.$$

In particular, if $X = (X_1, \dots, X_n) \sim \mathcal{N}(\mu', 1)^{\otimes n}$, then $I(\mu') = n$, and hence $\mathrm{Var}_{\mu'}(T) \ge \frac{(m_T'(\mu'))^2}{n}$.

Theorem A.4. *For any two probability measures P, Q on a common measurable space, there exists a coupling (X, Y) with $X \sim P$ and $Y \sim Q$ such that $\mathbb{P}(X \ne Y) = \mathrm{TV}(P, Q)$. Consequently, if $(X_i, Y_i)_{i=1}^n$ are independent maximal couplings of P and Q, then*

$$\#\{i \in [n] : X_i \ne Y_i\} \sim \mathrm{Bin}(n, \mathrm{TV}(P, Q)).$$

Theorem A.5 (Chernoff bounds)**.** *Let $Z \sim \mathrm{Bin}(n, p)$ and write $\lambda = np$. Then, for every $\delta > 0$, $\mathbb{P}(Z \ge (1+\delta)\lambda) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\lambda$. Equivalently, for every $t > \lambda$, $\mathbb{P}(Z \ge t) \le \left(\frac{e\lambda}{t} \right)^t$. In particular, for every fixed $\rho \in (0, 1)$, if $\lambda \le \rho k$, then $\mathbb{P}(Z \ge k) \le \exp(-c_\rho k)$, where $c_\rho > 0$ is a constant depending only on ρ.*

Lemma A.6. *There exists a universal constant $c > 0$ such that for every integer $n \ge 1$ and every $r \in \{0, 1, \dots, n\}$, $\mathbb{P}\left(\mathrm{Bin}\left(n, \frac{r}{n}\right) = r\right) \ge \frac{c}{\sqrt{r+1}}$.*

B Empirical Sensitivity of Median

Throughout this section, we assume for simplicity that n is odd. Write $n = 2m-1$ and for $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, let $x_{(1)} \le \cdots \le x_{(n)}$ denote the order statistics. Therefore, $\mathrm{med}(x) \coloneqq x_{(m)}$ and

$$S_\eta^{\mathrm{med}}(x) \coloneqq \sup_{y \in \mathbb{R}^n : d_H(x,y) \le k} |\mathrm{med}(x) - \mathrm{med}(y)|$$

where $k \coloneqq \lfloor \eta n \rfloor$.

Lemma B.1. *Let $0 \le k \le m-1$. Then, for every $x \in \mathbb{R}^n$,*

$$S_\eta^{\mathrm{med}}(x) \le \max\{x_{(m+k)} - x_{(m)}, x_{(m)} - x_{(m-k)}\}.$$

Proof. Fix $y \in \mathbb{R}^n$ with $d_H(x,y) \le k$. Thus y is obtained from x by changing at most k coordinates. We claim that $\mathrm{med}(y) \in [x_{(m-k)}, x_{(m+k)}]$. Indeed, there are $m+k$ entries of x that are at most $x_{(m+k)}$. Since at most k coordinates are changed, at least m of these entries remain in y. Hence at least m entries of y are at most $x_{(m+k)}$, which implies $\mathrm{med}(y) \le x_{(m+k)}$. Similarly, there are $m+k$ entries of x that are at least $x_{(m-k)}$. After changing at most k coordinates, at least m of these entries remain in y. Hence at least m entries of y are at least $x_{(m-k)}$, which implies $\mathrm{med}(y) \ge x_{(m-k)}$. Therefore $\mathrm{med}(y) \in [x_{(m-k)}, x_{(m+k)}]$. Since $\mathrm{med}(x) = x_{(m)}$, we obtain $|\mathrm{med}(y) - \mathrm{med}(x)| \le \max\{x_{(m+k)} - x_{(m)}, x_{(m)} - x_{(m-k)}\}$. Taking the supremum over all y with $d_H(x,y) \le k$ proves the claim. □

We will also need some results from uniform spacings, see Pyke [1965].

Theorem B.2. *Let $U_1, \dots, U_n \overset{iid}{\sim} \mathsf{Unif}[0,1]$, and let $0 = U_{(0)} \le U_{(1)} \le \cdots \le U_{(n)} \le U_{(n+1)} = 1$ denote the order statistics with endpoints added. Define the spacings $D_i \coloneqq U_{(i)} - U_{(i-1)}$ for $i = 1, \dots, n+1$. Then $(D_1, \dots, D_{n+1}) \sim \mathrm{Dirichlet}(1, \dots, 1)$. Equivalently, if $E_1, \dots, E_{n+1} \overset{iid}{\sim} \mathsf{Exp}(1)$ and $T \coloneqq \sum_{i=1}^{n+1} E_i$, then $(D_1, \dots, D_{n+1}) \overset{d}{=} \left(\frac{E_1}{T}, \dots, \frac{E_{n+1}}{T}\right)$.*

We can now state the Gaussian empirical sensitivity of the median.

Theorem B.3. *Let $n = 2m-1$ be odd, and let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Fix $\eta \in (0, 1/3)$ and set $k = \lfloor \eta n \rfloor$. There exists universal constants $c, C > 0$ such that, for every $\beta \in (0, 1/2)$, with probability at least $1 - \beta - Ce^{-cn}$, we have*

$$S_\eta^{\mathrm{med}}(X) \le C\left(\eta + \frac{\log(1/\beta)}{n}\right).$$

Proof. By translation invariance, it suffices to prove the result for $\mu = 0$. The case $k = 0$ or $\eta = 0$ directly gives us $S_\eta^{\mathrm{med}}(X) = 0$. So we will assume $k \ge 1$. Let Φ denote the standard Gaussian distribution function and define $U_i \coloneqq \Phi(X_i)$. Then $U_1, \dots, U_n \overset{\mathrm{iid}}{\sim} \mathsf{Unif}[0,1]$. Since Φ is strictly increasing, $X_{(j)} = \Phi^{-1}(U_{(j)})$ for every $j \in [n]$. We first localize the relevant order statistics away from the tails. Since $\eta < 1/3$ and $k = \lfloor \eta n \rfloor$, we have $m - k \ge \frac{n}{6}$ and $m + k \le \frac{5n}{6} + 1$.

Let $N_- \coloneqq \sum_{i=1}^n \mathbf{1}\{U_i \le 1/8\}$ and $N_+ \coloneqq \sum_{i=1}^n \mathbf{1}\{U_i \le 7/8\}$. Then $N_- \sim \mathrm{Bin}(n, 1/8)$ and $N_+ \sim \mathrm{Bin}(n, 7/8)$. By the binomial Chernoff bound in Theorem A.5, there exist universal constants $c, C > 0$ such that $\mathbb{P}\left(N_- \ge \frac{n}{6}\right) \le Ce^{-cn}$ and $\mathbb{P}\left(N_+ \le \frac{5n}{6}\right) \le Ce^{-cn}$. Therefore, with probability at least $1 - Ce^{-cn}$, $U_{(m-k)} \ge \frac{1}{8}$ and $U_{(m+k)} \le \frac{7}{8}$. On this event, all the order statistics $U_{(m-k)}, U_{(m)}, U_{(m+k)}$ lie in $[1/8, 7/8]$.

The inverse Gaussian CDF is Lipschitz on $[1/8, 7/8]$. Indeed, $\frac{d}{du}\Phi^{-1}(u) = \frac{1}{\varphi(\Phi^{-1}(u))}$, where φ is the standard Gaussian density, and the right-hand side is bounded on the compact interval $[1/8, 7/8]$. Hence there exists a universal constant $L < \infty$ such that, for all $u, v \in [1/8, 7/8]$, $|\Phi^{-1}(u) - \Phi^{-1}(v)| \le L|u - v|$. Consequently, on the localization event, $X_{(m+k)} - X_{(m)} \le L\big(U_{(m+k)} - U_{(m)}\big)$ and $X_{(m)} - X_{(m-k)} \le L\big(U_{(m)} - U_{(m-k)}\big)$.

It remains to control the two uniform order-statistic spacings. By Theorem B.2, each k-spacing, such as $U_{(m+k)} - U_{(m)}$ and $U_{(m)} - U_{(m-k)}$, has the same marginal distribution as A/T, where

$A \sim \mathrm{Gamma}(k,1)$ is the sum of k iid exponential random variables, and $T \sim \mathrm{Gamma}(n+1,1)$ is the sum of all $n+1$ such variables. We now prove a high-probability bound for A/T. Let $t \geq 0$. By Chernoff's inequality applied to the moment generating function of A, we have

$$\mathbb{P}(A \geq 2(k+t)) \leq e^{-(k+t)}\mathbb{E}e^{A/2} = e^{-(k+t)}2^k = \exp\{-t + k(\log 2 - 1)\} \leq e^{-t}.$$

Also, another Chernoff bound gives $\mathbb{P}(T \leq n/2) \leq e^{-cn}$ for a universal constant $c > 0$. Hence, with probability at least $1 - e^{-t} - e^{-cn}$, $\frac{A}{T} \leq 4\frac{k+t}{n}$. Applying this bound to the two spacings and taking a union bound gives $\mathbb{P}\left(\max\left\{U_{(m+k)} - U_{(m)},\, U_{(m)} - U_{(m-k)}\right\} > 4\frac{k+t}{n}\right) \leq 2e^{-t} + 2e^{-cn}$. Taking $t = \log(4/\beta)$, we obtain $\max\left\{U_{(m+k)} - U_{(m)},\, U_{(m)} - U_{(m-k)}\right\} \leq C\frac{k+\log(1/\beta)}{n}$ with probability at least $1 - \beta/2 - Ce^{-cn}$.

Using union bound, we get that with probability at least $1 - \beta - Ce^{-cn}$,

$$\max\left\{X_{(m+k)} - X_{(m)},\, X_{(m)} - X_{(m-k)}\right\} \leq C\frac{k+\log(1/\beta)}{n} \leq C\left(\eta + \frac{\log(1/\beta)}{n}\right).$$

Finally, by Theorem B.1, $S_\eta^{\mathrm{med}}(X) \leq \max\left\{X_{(m+k)} - X_{(m)},\, X_{(m)} - X_{(m-k)}\right\}$. Therefore, with probability at least $1 - \beta - Ce^{-cn}$, we conclude $S_\eta^{\mathrm{med}}(X) \leq C\left(\eta + \frac{\log(1/\beta)}{n}\right)$, which proves the theorem. □

C Resampling Adversary Upper Bound

In this section, we record a simple observation showing that the linear-in-η term in our corruption lower bound should not be expected under the resampling adversary.

Recall the resampling adversary:

Definition C.1 (Resampling Adversary)**.** Fix $\eta \in (0,1)$ and let $k := \lfloor \eta n \rfloor$. Given $X = (X_1, \ldots, X_n)$ with $X \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}$. Let $I \subset [n]$ be a uniformly random subset of size k, independent of X. Let $(X'_i)_{i\in I}$ be independent fresh samples from $\mathcal{N}(\mu, \mathbb{I}_d)$, also independent of (X, I). The resampled dataset X^{res} is defined by

$$X_i^{\mathrm{res}} := \begin{cases} X'_i, & \text{if } i \in I, \\ X_i, & \text{if } i \notin I. \end{cases}$$

The following proposition quantifies how much the empirical mean shifts under a resampling adversary.

Proposition C.2. *Let* $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ *and* $\overline{X}^{\mathrm{res}} := \frac{1}{n}\sum_{i=1}^n X_i^{\mathrm{res}}$. *Then*

$$\left(\mathbb{E}_{X,I,X'}\|\overline{X}^{\mathrm{res}} - \overline{X}\|_2^2\right)^{1/2} \asymp \sqrt{\frac{\eta d}{n}}.$$

Consequently, under the resampling adversary, the empirical mean has empirical stability of order $\sqrt{\eta d/n}$, and does not exhibit an η term.

Proof. By definition, we have

$$\overline{X}^{\mathrm{res}} - \overline{X} = \frac{1}{n}\sum_{i\in I}(X'_i - X_i).$$

Conditional on I, the random vectors $(X'_i - X_i)_{i\in I}$ are independent, and each has law $\mathcal{N}(0, 2\mathbb{I}_d)$. Therefore,

$$(\overline{X}^{\mathrm{res}} - \overline{X}) \mid I \sim \mathcal{N}\left(0, \frac{2k}{n^2}\mathbb{I}_d\right)$$

Hence,

$$\mathbb{E}_{X,X'|I}\|\overline{X}^{\mathrm{res}} - \overline{X}\|_2^2 = \mathrm{tr}\left(\frac{2k}{n^2}\mathbb{I}_d\right) = \frac{2kd}{n^2}.$$

Since RHS is deterministic in I, averaging over I gives the same value. Therefore, taking square root yields

$$\left(\mathbb{E}_{X,I,X'}\|\overline{X}^{\mathrm{res}} - \overline{X}\|_2^2\right)^{1/2} = \sqrt{\frac{2kd}{n^2}} \asymp \sqrt{\frac{\eta d}{n}}.$$

□

D Lemmas

In this section, we isolate the main lemma that allow us to prove the high-dimensional theorem for the mean obstruction by reducing the problem to a scalar setting. Ultimately, we claim that it suffices to prove the lower bound for a bounded, one-dimensional estimator $h\colon \mathbb{R}^n \to [0,1]$ that inherits $O(1/n)$ integrated risk.

This restriction to compact interval $[0,1]$ is rigorously justified by the following reduction: first, we restrict the parameter space of the projected 1D problem to the interval $\mu' \in [0,1]$. Because the original high-dimensional estimator f is uniformly MSE-accurate over all $\mathbb{R}^d$, its 1D projection must maintain low average risk over this specified bounded sub-interval under a Gaussian prior. Because the true parameter μ is now restricted to $[0,1]$, any output of the estimator outside this interval is strictly suboptimal, and thus we may clip the estimator's output to $[0,1]$ which strictly decreases or maintains both its estimation and its empirical sensitivity. Consequently, we may establish an empirical sensitivity lower bound for the clipped version of the estimator h to guarantee a valid lower bound for the original unrestricted high-dimensional estimator f.

Let $V = \mathsf{span}(u)$. If $\lambda \in u^\perp$ and $Z \sim \mathcal{N}(0, \mathbb{I}_d)$, then $\lambda + (\mathbb{I}_d - uu^\top)Z$ is drawn from a Gaussian distribution on the affine subspace $\lambda + V^\perp$ with mean λ.

Lemma D.1. *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable. Fix $u \in \mathbb{S}^{d-1}$ and $\lambda \in u^\perp$. For $t = (t_1, \dots, t_n) \in \mathbb{R}^n$, define*

$$g_\lambda(t) \coloneqq \mathbb{E}_Z\left[\langle u, f(t_1 u + V_1, \dots, t_n u + V_n)\rangle\right], \qquad V_i \coloneqq \lambda + (\mathbb{I}_d - uu^\top)Z_i,$$

where $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbb{I}_d)$. Then for every $\eta \in (0,1)$, every $q \in \{1,2\}$, and every $\mu' \in \mathbb{R}$,

$$\left(\mathbb{E}_{T \sim \mathcal{N}(\mu',1)^{\otimes n}}\left[(S_\eta^{g_\lambda}(T))^q\right]\right)^{1/q} \le \mathsf{ES}_{\eta,q}(f; \mu' u + \lambda).$$

Moreover,

$$\mathbb{E}_{T \sim \mathcal{N}(\mu',1)^{\otimes n}}[(g_\lambda(T) - \mu')^2] \le \mathbb{E}_{X \sim \mathcal{N}(\mu' u + \lambda, \mathbb{I}_d)^{\otimes n}}\left[\langle u, f(X) - (\mu' u + \lambda)\rangle^2\right].$$

Proof. Set $k \coloneqq \lfloor \eta n \rfloor$. Fix $t, t' \in \mathbb{R}^n$ with $d_H(t,t') \le k$, and use the same orthogonal noise to lift both scalar samples: $x_i = t_i u + V_i$ and $x_i' = t_i' u + V_i$. Then $d_H(x,x') \le k$ and so

$$\begin{aligned} |g_\lambda(t') - g_\lambda(t)| &= |\mathbb{E}_Z \langle u, f(x') - f(x)\rangle| \\ &\le \mathbb{E}_Z \|f(x') - f(x)\| \\ &\le \mathbb{E}_Z S_\eta^f(x). \end{aligned}$$

Taking the supremum over all such t' gives

$$S_\eta^{g_\lambda}(t) \le \mathbb{E}_Z S_\eta^f(t_1 u + V_1, \dots, t_n u + V_n),$$

which implies the sensitivity comparison we want by Jensen's inequality and averaging over $T \sim \mathcal{N}(\mu', 1)^{\otimes n}$.

For the risk comparison, Jensen gives

$$\begin{aligned} \mathbb{E}_T[(g_\lambda(T) - \mu')^2] &= \mathbb{E}_T\left[\left(\mathbb{E}_Z[\langle u, f(T_1 u + V_1, \dots, T_n u + V_n)\rangle - \mu']\right)^2\right] \\ &\le \mathbb{E}_{T,Z}\left[(\langle u, f(T_1 u + V_1, \dots, T_n u + V_n)\rangle - \mu')^2\right]. \end{aligned}$$

For each i,

$$T_i u + V_i = T_i u + \lambda + (\mathbb{I}_d - uu^\top)Z_i \sim \mathcal{N}(\mu' u + \lambda, \mathbb{I}_d).$$

Also $\langle u, \mu' u + \lambda\rangle = \mu'$, since $\lambda \in u^\perp$. Therefore the last term above is just

$$\mathbb{E}_{X \sim \mathcal{N}(\mu' u + \lambda, \mathbb{I}_d)^{\otimes n}}\left[\langle u, f(X) - (\mu' u + \lambda)\rangle^2\right],$$

which proves the claim. □

Lemma D.2. *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and $\big(C_{\mathsf{MSE}}, \frac{d}{n}\big)$-uniformly MSE-accurate. Fix $\rho > 0$. Then there exist $u \in \mathbb{S}^{d-1}$ and $\lambda_* \in u^\perp$ such that the scalar estimator g_{λ_*} from Theorem D.1 satisfies*

$$\mathbb{E}_{\mu' \sim \mathcal{N}(0,\rho^2)} \mathbb{E}_{T \sim \mathcal{N}(\mu',1)^{\otimes n}} [(g_{\lambda_*}(T) - \mu')^2] \le \frac{C_{\mathsf{MSE}}}{n}.$$

Proof. Let $M \sim \mathcal{N}(0, \rho^2 \mathbb{I}_d)$ and let $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, independently of all other randomness. For every fixed $z \in \mathbb{R}^d$,

$$\mathbb{E}_U[\langle U, z\rangle^2] = \frac{\|z\|^2}{d}.$$

Hence, by uniform MSE accuracy,

$$\begin{aligned}&\mathbb{E}_U \mathbb{E}_M \mathbb{E}_{X \sim \mathcal{N}(M, \mathbb{I}_d)^{\otimes n}} [\langle U, f(X) - M\rangle^2] \\ &\quad = \frac{1}{d} \mathbb{E}_M \mathbb{E}_{X \sim \mathcal{N}(M, \mathbb{I}_d)^{\otimes n}} \|f(X) - M\|^2 \le \frac{C_{\mathsf{MSE}}}{n}.\end{aligned}$$

Therefore there exists $u \in \mathbb{S}^{d-1}$ such that

$$\mathbb{E}_M \mathbb{E}_{X \sim \mathcal{N}(M, \mathbb{I}_d)^{\otimes n}} [\langle u, f(X) - M\rangle^2] \le \frac{C_{\mathsf{MSE}}}{n}.$$

Decompose

$$M = \mu' u + \lambda, \qquad \mu' \coloneqq \langle M, u\rangle, \qquad \lambda \coloneqq M - \langle M, u\rangle u.$$

Then $\mu' \sim \mathcal{N}(0, \rho^2)$, $\lambda \in u^\perp$, and μ' is independent of λ. Applying Theorem D.1 and averaging over λ gives

$$\begin{aligned}&\mathbb{E}_\lambda \mathbb{E}_{\mu' \sim \mathcal{N}(0,\rho^2)} \mathbb{E}_{T \sim \mathcal{N}(\mu',1)^{\otimes n}} [(g_\lambda(T) - \mu')^2] \\ &\quad \le \mathbb{E}_M \mathbb{E}_{X \sim \mathcal{N}(M, \mathbb{I}_d)^{\otimes n}} [\langle u, f(X) - M\rangle^2] \le \frac{C_{\mathsf{MSE}}}{n}.\end{aligned}$$

Thus some $\lambda_* \in u^\perp$ satisfies the desired bound. □

Lemma D.3. *Let $K \subseteq \mathbb{R}$ be a closed interval and let $\mathrm{clip}_K(t)$ denote Euclidean projection onto K. For any scalar estimator g, define $h \coloneqq \mathrm{clip}_K \circ g$. Then, for every $x \in \mathbb{R}^n$, we have $S^h_\eta(x) \le S^g_\eta(x)$. Moreover, if $\mu' \in K$, then for every x, $|h(x) - \mu'| \le |g(x) - \mu'|$.*

Proof. The projection map clip_K is 1-Lipschitz and fixes every point of K. Thus, if $\mu \in K$, we have

$$|h(x) - \mu'| = |\mathrm{clip}_K(g(x)) - \mathrm{clip}_K(\mu')| \le |g(x) - \mu'|.$$

Similarly, for any x, x', we obtain $|h(x') - h(x)| \le |g(x') - g(x)|$. Taking the supremum over all x' with $d_H(x, x') \le \lfloor \eta n \rfloor$ proves the sensitivity bound we want. □

Lemma D.4. *Let $f\colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and $\big(C_{\mathsf{MSE}}, \frac{d}{n}\big)$-uniformly MSE-accurate. There exists $u \in \mathbb{S}^{d-1}$, $\lambda_* \in u^\perp$, and a scalar estimator $h\colon \mathbb{R}^n \to [0,1]$ such that, for every $\eta \in (0,1)$, every $q \in \{1,2\}$, and every $\mu' \in [0,1]$,*

$$(\mathbb{E}_{X \sim \mathcal{N}(\mu',1)^{\otimes n}} [S^h_\eta(X)^q])^{1/q} \le \mathsf{ES}_{\eta,q}(f; \mu' u + \lambda_*)$$

and

$$\int_0^1 \mathbb{E}_{X \sim \mathcal{N}(\mu',1)^{\otimes n}} [(h(X) - \mu')^2] \, d\mu' \le \frac{C_{\mathsf{int}}}{n},$$

where $C_{\mathsf{int}} = C_{\mathsf{int}}(C_{\mathsf{MSE}})$.

Proof. Apply Theorem D.2 with $\rho = 1$ to obtain u and λ_*. Let $g \coloneqq g_{\lambda_*}$ be the scalar estimator from Theorem D.1, and define $h \coloneqq \mathrm{clip}_{[0,1]} \circ g$. By Theorem D.3, $S^h_\eta(x) \le S^g_\eta(x)$ for every x, and for every $\mu' \in [0,1]$, $|h(x) - \mu'| \le |g(x) - \mu'|$. The sensitivity comparison thus follow from

Theorem D.1. It remains to prove the risk bound. Indeed, let φ denote the density of $\mathcal{N}(0,1)$ and set $c_0 \coloneqq \inf_{\mu' \in [0,1]} \varphi(\mu') > 0$. Then,

$$\begin{aligned}\int_0^1 \mathbb{E}_{X \sim \mathcal{N}(\mu',1)^{\otimes n}}[(h(X)-\mu')^2]\, d\mu' &\le \int_0^1 \mathbb{E}_{X \sim \mathcal{N}(\mu',1)^{\otimes n}}[(g(X)-\mu')^2]\, d\mu' \\ &\le c_0^{-1}\mathbb{E}_{\mu' \sim \mathcal{N}(0,1)}\mathbb{E}_{X \sim \mathcal{N}(\mu',1)^{\otimes n}}[(g(X)-\mu')^2] \\ &\le \frac{c_0^{-1} C_{\mathsf{MSE}}}{n}.\end{aligned}$$

Therefore, the claim holds with $C_{\mathsf{int}} \coloneqq c_0^{-1} C_{\mathsf{MSE}}$. □

E Proof of Theorem 3.2

We will rely on the following estimator that is recently studied in Chen et al. [2026] in the context of empirical mean recovery.

Theorem E.1 (Theorem 5.5 from Chen et al. [2026])**.** *Suppose $0 \le \eta < \frac{1}{2}$ is bounded away from $\frac{1}{2}$. There exists a (computationally inefficient) η-robust estimator that, for all $\mu \in \mathbb{R}^d$, given n iid samples from $\mathcal{N}(\mu, \mathbb{I}_d)$ with empirical mean $\overline{x}$, outputs $\hat{x}$ such that $\|\hat{x} - \overline{x}\|_2 \le \alpha$, with probability at least $1-\beta$, for*

$$\alpha = O\left(\eta\sqrt{\log(1/\eta)} + \sqrt{\eta \cdot \frac{d + \log(1/\beta)}{n}}\right)$$

We will use a slightly strengthened formulation that follows immediately from the proof of Theorem E.1. This estimator operates by searching over candidate cleaned datasets $Y \in (\mathbb{R}^d)^n$ that differ from the corrupted input $\widetilde{X}$ in at most ηn coordinates that satisfy some constraints. We denote the set of feasible candidates in this search as $\mathcal{F}_{\eta,\beta}(\widetilde{X})$. The proof of Theorem E.1 establishes the following two uniform properties in the setup of Theorem E.1: for every η-corruption $\widetilde{X}$ of X,

1. the clean dataset X is feasible, i.e., $X \in \mathcal{F}_{\eta,\beta}(\widetilde{X})$, and
2. every feasible candidate $Y \in \mathcal{F}_{\eta,\beta}(\widetilde{X})$ satisfies $\|\overline{Y} - \overline{X}\|_2 \le \alpha$, where $\overline{Y} \coloneqq \frac{1}{n}\sum_{i=1}^n Y_i$.

The original estimator in Theorem E.1 may output $\perp$ if the feasibility system is empty. We will slightly modify this as follow to ensure $f_{\eta,\beta}$ is a measurable map. Given an input dataset $Z = (Z_1, \ldots, Z_n)$, if $\mathcal{F}_{\eta,\beta}(Z) = \varnothing$, we set $f_{\eta,\beta}(Z) \coloneqq \overline{Z}$. Otherwise, we select a feasible candidate $Y^*(Z)$ that minimizes the Hamming distance to the input:

$$Y^*(Z) \in \operatorname*{argmin}_{Y \in \mathcal{F}_{\eta,\beta}(Z)} |\{i \in [n] \colon Y_i \ne Z_i\}|,$$

where ties are broken arbitrarily, and we set $f_{\eta,\beta}(Z) \coloneqq \overline{Y^*(Z)}$. We note that this tie-breaking procedure preserves the guarantee of Theorem E.1 because the guarantee there applies to *all* feasible solutions.

With this modification in mind, we see that on the high probability event established in Theorem E.1, if the input is not corrupted, i.e., $Z = X$, then $f_{\eta,\beta}(X) = \overline{X}$. Indeed, to see this, note that $X \in \mathcal{F}_{\eta,\beta}(X)$ and since this candidate has zero disagreements with the input X, it uniquely minimizes the Hamming distance, and thus $Y^*(X) = X$, giving us $f_{\eta,\beta}(X) = \overline{X}$.

Using this, we can now bound the empirical sensitivity and clean MSE bound. Indeed, assume the high probability event from Theorem E.1 holds with the bad event happening with probability at most $\frac{\beta}{2}$. We see that

$$\begin{aligned}S_\eta^{f_{\eta,\beta}}(X) &= \sup_{X' \colon d_H(X,X') \le \lfloor \eta n \rfloor} \|f_{\eta,\beta}(X') - f_{\eta,\beta}(X)\|_2 \\ &= \sup_{X' \colon d_H(X,X') \le \lfloor \eta n \rfloor} \|f_{\eta,\beta}(X') - \overline{X}\|_2 \le \alpha = \widetilde{O}\left(\eta + \sqrt{\eta \cdot \frac{d + \log(1/\beta)}{n}}\right),\end{aligned}$$

which is what we wanted. Similarly, on the same event one can note that

$$\|f_{\eta,\beta}(X) - \mu\|_2^2 = \|\overline{X} - \mu\|_2^2 = O\left(\frac{d + \log(1/\beta)}{n}\right).$$

with probability $1 - \frac{\beta}{2}$. Union bound establish what we want.

F Proof of Theorem 3.4

Throughout this section, for a scalar parameter $\mu' \in \mathbb{R}$, we write $P_{\mu'} \coloneqq \mathcal{N}(\mu', 1)^{\otimes n}$. By the reduction established in Theorem D.4, to prove the high-dimensional theorem, it suffices to prove the following scalar claim: if the induced one-dimensional estimator $h\colon \mathbb{R}^n \to [0,1]$ satisfies the integrated risk bound

$$\int_0^1 \mathbb{E}_{X\sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu' \le \frac{C_{\mathsf{int}}}{n},$$

then, for a corruption budget $1 \le k = \lfloor \eta n \rfloor \le \kappa\sqrt{n}$, its average empirical sensitivity must be large:

$$\sup_{\mu'\in[0,1]} \mathbb{E}_{X\sim P_{\mu'}}[S_\eta^h(X)] \ge c_{\mathrm{low}}\eta.$$

To see exactly why this scalar claim implies the full high-dimensional theorem, recall that Theorem D.4 guarantees the existence of a "good" unit direction $u \in \mathbb{S}^{d-1}$ and an orthogonal shift $\lambda_* \in u^\perp$. When we evaluate the original high-dimensional estimator f along the one-dimensional affine line parameterized by $\mu' u + \lambda_*$, its empirical sensitivity strictly upper-bounds the sensitivity of the induced scalar estimator h. Together with Jensen's inequality, this gives us

$$\mathsf{ES}_{\eta,2}(f; \mu' u + \lambda_*) \ge \left(\mathbb{E}_{X\sim P_{\mu'}}[S_\eta^h(X)^2]\right)^{1/2} \ge \mathbb{E}_{X\sim P_{\mu'}}[S_\eta^h(X)].$$

Taking the supremum over $\mu' \in [0,1]$ on both sides demonstrates that establishing the $c_{\mathrm{low}}\eta$ lower bound for h immediately yields the required lower bound for f, completing the reduction.

F.1 Statistical Indistinguishability of the Local and Global Shift

WLOG we may choose η such that $\eta = \frac{\lfloor \eta n \rfloor}{n}$, i.e., $\eta n \in \{1, \dots, n\}$. For $\delta \ge 0$ and $\mu' \in \mathbb{R}$, we define the local-shift distribution $Q_{\mu',\delta}$ via the following generative process: draw $X \sim P_{\mu'}$, draw a k-subset $I \subset [n]$ uniformly at random (independent of X), and output $X^{(I,\delta)} \coloneqq X + \delta e_I$, i.e., this means that

$$X_i^{(I,\delta)} = \begin{cases} X_i + \delta, & \text{if } i \in I, \\ X_i, & \text{if } i \notin I. \end{cases}$$

The key observation is to show that this local adversarial corruption is statistically indistinguishable from a global mean shift $P_{\mu'+\eta\delta}$.

Lemma F.1. *For every $\mu' \in \mathbb{R}$ and every $\delta \ge 0$,*

$$\chi^2(Q_{\mu',\delta} \| P_{\mu'+\eta\delta}) \le \exp\left(\frac{k^2}{n}(e^{\delta^2} - 1 - \delta^2)\right) - 1.$$

To prove Theorem F.1, we require two elementary auxiliary lemmas.

Lemma F.2 (Gaussian likelihood-ratio identity)**.** *For any $\theta, \mu, \nu \in \mathbb{R}^n$, if $X \sim \mathcal{N}(\theta, \mathbb{I}_n)$, then*

$$\mathbb{E}_{X\sim\mathcal{N}(\theta,\mathbb{I}_n)}\left[\frac{d\mathcal{N}(\mu,\mathbb{I}_n)}{d\mathcal{N}(\theta,\mathbb{I}_n)}(X) \cdot \frac{d\mathcal{N}(\nu,\mathbb{I}_n)}{d\mathcal{N}(\theta,\mathbb{I}_n)}(X)\right] = \exp(\langle \mu - \theta, \nu - \theta\rangle).$$

Proof. Let $a \coloneqq \mu - \theta$ and $b \coloneqq \nu - \theta$. If $X \sim \mathcal{N}(\theta, \mathbb{I}_n)$, then $X = \theta + G$ with $G \sim \mathcal{N}(0, \mathbb{I}_n)$. The likelihood ratio of $\mathcal{N}(\mu, \mathbb{I}_n)$ with respect to $\mathcal{N}(\theta, \mathbb{I}_n)$ is $\frac{d\mathcal{N}(\mu,\mathbb{I}_n)}{d\mathcal{N}(\theta,\mathbb{I}_n)}(X) = \exp\left(\langle a, X - \theta\rangle - \frac{1}{2}\|a\|_2^2\right) =$

$\exp\left(\langle a, G\rangle - \frac{1}{2}\|a\|_2^2\right)$. Similarly, $\frac{d\mathcal{N}(\nu,\mathbb{I}_n)}{d\mathcal{N}(\theta,\mathbb{I}_n)}(X) = \exp\left(\langle b, G\rangle - \frac{1}{2}\|b\|_2^2\right)$. Therefore, the expected product of the likelihood ratios is

$$\mathbb{E}_{X\sim\mathcal{N}(\theta,\mathbb{I}_n)}\left[\frac{d\mathcal{N}(\mu,\mathbb{I}_n)}{d\mathcal{N}(\theta,\mathbb{I}_n)}(X)\frac{d\mathcal{N}(\nu,\mathbb{I}_n)}{d\mathcal{N}(\theta,\mathbb{I}_n)}(X)\right] = \exp\left(-\frac{1}{2}\|a\|_2^2 - \frac{1}{2}\|b\|_2^2\right)\mathbb{E}_G\exp\left(\langle a+b, G\rangle\right).$$

Using the standard Gaussian moment-generating function $\mathbb{E}\exp(\langle t, G\rangle) = \exp(\frac{1}{2}\|t\|_2^2)$ with $t = a + b$, the expectation evaluates to $\exp\left(-\frac{1}{2}\|a\|_2^2 - \frac{1}{2}\|b\|_2^2 + \frac{1}{2}\|a+b\|_2^2\right)$. Expanding the squared norm gives the identity $\frac{1}{2}\|a+b\|_2^2 - \frac{1}{2}\|a\|_2^2 - \frac{1}{2}\|b\|_2^2 = \langle a, b\rangle$. Substituting back $a = \mu - \theta$ and $b = \nu - \theta$ yields the final expected value $\exp(\langle \mu - \theta, \nu - \theta\rangle)$, as claimed. □

Lemma F.3. *Let A and B be independent uniformly random k-subsets of $[n]$, and let $H = |A \cap B|$. For every $\lambda \geq 0$,*

$$\mathbb{E}\exp\left(\lambda\left(H - \frac{k^2}{n}\right)\right) \leq \exp\left(\frac{k^2}{n}(e^\lambda - 1 - \lambda)\right).$$

Proof. Since A and B are independent and exchangeable, we may condition on A throughout the proof. Fix a realization of A with $|A| = k$. For each $i \in A$, define $X_i := \mathbf{1}\{i \in B\}$. Then $H = |A \cap B| = \sum_{i\in A} X_i$. Set $c := e^\lambda - 1$. Since $\lambda \geq 0$, we have $c \geq 0$. Moreover, because each X_i is $\{0,1\}$-valued, $e^{\lambda X_i} = 1 + cX_i$. Therefore $e^{\lambda H} = \prod_{i\in A} e^{\lambda X_i} = \prod_{i\in A}(1 + cX_i)$. Expanding the product gives

$$e^{\lambda H} = \sum_{R\subseteq A} c^{|R|}\prod_{i\in R} X_i = \sum_{R\subseteq A} c^{|R|}\mathbf{1}\{R \subseteq B\}.$$

Taking expectation over the random set B yields $\mathbb{E}\left[e^{\lambda H} \mid A\right] = \sum_{R\subseteq A} c^{|R|}\mathbb{P}(R \subseteq B)$. If $R \subseteq A$ has size s, then $\mathbb{P}(R \subseteq B) = \frac{\binom{n-s}{k-s}}{\binom{n}{k}} = \frac{(k)_s}{(n)_s}$, where $(r)_s := r(r-1)\cdots(r-s+1)$ denotes the falling factorial.

Since there are $\binom{k}{s}$ subsets $R \subseteq A$ of size s, we get $\mathbb{E}\left[e^{\lambda H} \mid A\right] = \sum_{s=0}^k \binom{k}{s} c^s \frac{(k)_s}{(n)_s}$. For each $0 \leq s \leq k$, $\frac{(k)_s}{(n)_s} = \prod_{j=0}^{s-1}\frac{k-j}{n-j} \leq \left(\frac{k}{n}\right)^s$, because $\frac{k-j}{n-j} \leq \frac{k}{n}$ whenever $0 \leq j \leq s-1$ and $k \leq n$. Hence

$$\mathbb{E}\left[e^{\lambda H} \mid A\right] \leq \sum_{s=0}^k \binom{k}{s} c^s \left(\frac{k}{n}\right)^s = \left(1 + \frac{k}{n}c\right)^k = \left(1 + \frac{k}{n}(e^\lambda - 1)\right)^k.$$

The right-hand side no longer depends on A, so the same bound holds unconditionally: $\mathbb{E}e^{\lambda H} \leq \left(1 + \frac{k}{n}(e^\lambda - 1)\right)^k$.

Therefore, $\mathbb{E}\exp\left(\lambda\left(H - \frac{k^2}{n}\right)\right) = \exp\left(-\lambda\frac{k^2}{n}\right)\mathbb{E}e^{\lambda H} \leq \exp\left(-\lambda\frac{k^2}{n}\right)\left(1 + \frac{k}{n}(e^\lambda - 1)\right)^k$. Using $\log(1+x) \leq x$ for $x \geq 0$, we further obtain $\left(1 + \frac{k}{n}(e^\lambda - 1)\right)^k = \exp\left[k\log\left(1 + \frac{k}{n}(e^\lambda - 1)\right)\right] \leq \exp\left[\frac{k^2}{n}(e^\lambda - 1)\right]$. Combining the last two displays gives

$$\mathbb{E}\exp\left(\lambda\left(H - \frac{k^2}{n}\right)\right) \leq \exp\left(-\lambda\frac{k^2}{n} + \frac{k^2}{n}(e^\lambda - 1)\right) = \exp\left(\frac{k^2}{n}(e^\lambda - 1 - \lambda)\right),$$

which proves the claim. □

Proof of Theorem F.1. By translation invariance, it suffices to set $\mu' = 0$. Let $P := P_{\eta\delta} = \mathcal{N}(\eta\delta, \mathbb{I}_n)$. The local-shift distribution $Q_{0,\delta}$ is a mixture over the subset I, so its likelihood ratio with respect to P is

$$\frac{dQ_{0,\delta}}{dP}(X) = \mathbb{E}_I\left[\frac{d\mathcal{N}(\delta e_I, \mathbb{I}_n)}{dP}(X)\right].$$

Using Theorem F.2 to compute the χ^2 divergence, we expand the square of the mixture:

$$\chi^2(Q_{0,\delta}\|P) = \mathbb{E}_{X\sim P}\left[\left(\frac{dQ_{0,\delta}}{dP}(X)\right)^2\right] - 1$$

$$= \mathbb{E}_{I,J}\left[\exp(\langle \delta e_I - \eta\delta, \delta e_J - \eta\delta\rangle)\right] - 1.$$

We evaluate the inner product: $\langle \delta e_I - \eta\delta, \delta e_J - \eta\delta\rangle = \delta^2|I \cap J| - 2\eta\delta^2 k + n\eta^2\delta^2$. Since $\eta = k/n$, this simplifies to $\delta^2(|I \cap J| - k^2/n)$. Letting $H = |I \cap J|$ and applying Theorem F.3 with $\lambda = \delta^2$ yields:

$$\chi^2(Q_{0,\delta}\|P) = \mathbb{E}_{I,J}\left[\exp\left(\delta^2\left(H - \frac{k^2}{n}\right)\right)\right] - 1 \le \exp\left(\frac{k^2}{n}(e^{\delta^2} - 1 - \delta^2)\right) - 1,$$

concluding the proof. □

Corollary F.4. *Let $h : \mathbb{R}^n \to [0,1]$ be any bounded measurable estimator. Then*

$$\left|\mathbb{E}_{Q_{\mu',\delta}}[h] - \mathbb{E}_{P_{\mu'+\eta\delta}}[h]\right| \le \sqrt{\mathbb{E}_{P_{\mu'+\eta\delta}}[(h(X) - (\mu' + \eta\delta))^2] \cdot \chi^2(Q_{\mu',\delta}\|P_{\mu'+\eta\delta})}.$$

Proof. Let $P \coloneqq P_{\mu'+\eta\delta}$ and let $R(X) = h(X) - (\mu' + \eta\delta)$. Applying the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\mathbb{E}_{Q_{\mu',\delta}}[h] - \mathbb{E}_P[h]\right| &= \left|\mathbb{E}_P\left[R(X)\left(\frac{dQ_{\mu',\delta}}{dP}(X) - 1\right)\right]\right| \\ &\le \sqrt{\mathbb{E}_P[R(X)^2] \cdot \mathbb{E}_P\left[\left(\frac{dQ_{\mu',\delta}}{dP}(X) - 1\right)^2\right]},\end{aligned}$$

which perfectly matches the desired statement. □

F.2 Tracking the Global Shift

Having established indistinguishability, we will now show that the estimator must track the global shift to maintain its uniform MSE accuracy assumption. Indeed, let $t \coloneqq \eta\delta$, where $\delta > 0$ is a small constant chosen later. We let $a(\mu')$ and $b(\mu')$ be the expected output and bias of h respectively.

To lower-bound the supremum over μ', we lower-bound the average over a carefully chosen prior. Let $w\colon \mathbb{R} \to \mathbb{R}_+$ be a continuously differentiable probability density function supported strictly on $[\epsilon, 1-\epsilon]$ for some fixed $\epsilon > 0$. We assume $t < \epsilon$ so that the shifted density $w(\cdot - t)$ remains entirely supported within $[0,1]$.

By our risk assumption, the average squared bias over the interval satisfies $\int_0^1 b(\mu')^2\, d\mu' \le \int_0^1 \mathbb{E}_{X\sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu' \le C_{\mathsf{int}}/n$. Because w is smooth and compactly supported, its derivative w' has finite L^2 norm, ensuring $\|w(\cdot - t) - w(\cdot)\|_{L^2} \le t\|w'\|_{L^2}$. Therefore, by change of variables, we have

$$\begin{aligned}\int w(\mu')(a(\mu' + t) - a(\mu'))\, d\mu' &= t + \int b(v)(w(v - t) - w(v))\, dv \\ &\ge t - \|b\|_{L^2}\|w(\cdot - t) - w(\cdot)\|_{L^2} \\ &\ge t - \sqrt{\frac{C_{\mathsf{int}}}{n}} \cdot t\|w'\|_{L^2} \\ &= t - O\left(\frac{t}{\sqrt{n}}\right) \ge t - o(t).\end{aligned}$$

Thus, the accuracy guarantee implies that, on average over the prior w, the estimator's mean output shifts by at least $t - o(t)$, where $t = \eta\delta$.

To finish the proof, notice that because $Q_{\mu',\delta}$ is generated by shifting exactly $k = \lfloor \eta n \rfloor$ coordinates of $X \sim P_{\mu'}$, the resulting dataset $X^{(I,\delta)}$ is a valid η-corruption. Therefore, we have $S_\eta^h(X) \ge |h(X^{(I,\delta)}) - h(X)|$. By Triangle Inequality, we have

$$\begin{aligned}\mathbb{E}_{X\sim P_{\mu'}}[S_\eta^h(X)] &\ge \mathbb{E}_{X\sim P_{\mu',I}}[|h(X^{(I,\delta)}) - h(X)|] \\ &\ge \mathbb{E}_{X\sim P_{\mu',I}}[h(X^{(I,\delta)})] - \mathbb{E}_{X\sim P_{\mu'}}[h(X)]\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbb{E}_{X\sim Q_{\mu',\delta}}[h(X)] - \mathbb{E}_{X\sim P_{\mu'}}[h(X)] \\
&= (\mathbb{E}_{X\sim P_{\mu'+t}}[h(X)] - \mathbb{E}_{X\sim P_{\mu'}}[h(X)]) + (\mathbb{E}_{X\sim Q_{\mu',\delta}}[h(X)] - \mathbb{E}_{X\sim P_{\mu'+t}}[h(X)]) \\
&\geq a(\mu'+t) - a(\mu') - \left|\mathbb{E}_{X\sim Q_{\mu',\delta}}[h(X)] - \mathbb{E}_{X\sim P_{\mu'+t}}[h(X)]\right|
\end{aligned}
$$

We now average this inequality over the prior $w(\mu')$ to establish the final lower bound. The first term satisfies $\int w(\mu')(a(\mu'+t)-a(\mu'))\,d\mu' \geq \eta\delta - o(\eta\delta)$, where the $o(\eta\delta)$ term vanishes as $n\to\infty$ due to the smoothness of w. For the second term, we apply Theorem F.4 and the Cauchy-Schwarz inequality over the prior integral to decouple the estimator's accuracy from the statistical distance between distributions:

$$
\begin{aligned}
&\int w(\mu')\left|\mathbb{E}_{Q_{\mu',\delta}}[h] - \mathbb{E}_{P_{\mu'+t}}[h]\right|\,d\mu' \\
&\qquad\leq \int w(\mu')\sqrt{\mathbb{E}_{P_{\mu'+t}}[(h(X)-(\mu'+t))^2]\cdot\chi^2(Q_{\mu',\delta}\|P_{\mu'+t})}\,d\mu' \\
&\qquad\leq \sqrt{\int w(\mu')\mathbb{E}_{P_{\mu'+t}}[(h(X)-(\mu'+t))^2]\,d\mu'}\cdot\sqrt{\int w(\mu')\chi^2(Q_{\mu',\delta}\|P_{\mu'+t})\,d\mu'}.
\end{aligned}
$$

We evaluate the two components under the square root independently:

1. By the uniform MSE assumption on the scalar estimator h, we have that variance is bounded by MSE, and thus
$$\int w(\mu')\mathbb{E}_{P_{\mu'+t}}[(h(X)-(\mu'+t))^2]\,d\mu' \leq \int w(v-t)\mathbb{E}_{P_v}[(h(X)-v)^2]\,dv \leq \frac{C_{\mathsf{int}}}{n}\|w\|_\infty.$$
2. Applying Lemma F.1 for the regime $k\leq\kappa\sqrt{n}$, the divergence satisfies
$$\chi^2(Q_{\mu',\delta}\|P_{\mu'+t}) \leq \exp\left(\frac{k^2}{n}(e^{\delta^2}-1-\delta^2)\right)-1 \lesssim \frac{k^2\delta^4}{n}.$$

Substituting these bounds into the Cauchy-Schwarz product yields an error term of:

$$\sqrt{\frac{C_{\mathsf{int}}}{n}\|w\|_\infty\cdot\frac{k^2\delta^4}{n}} = \sqrt{C_{\mathsf{int}}\|w\|_\infty}\cdot\frac{k\delta^2}{n} = O(\eta\delta^2).$$

Putting this together, the average empirical sensitivity is bounded by:

$$\int w(\mu')\mathbb{E}_{X\sim P_{\mu'}}[S_\eta^h(X)]\,d\mu' \geq \eta\delta - o(\eta\delta) - O(\eta\delta^2).$$

By choosing the absolute constant $\delta>0$ to be sufficiently small, the $O(\eta\delta^2)$ term is made strictly smaller than $\eta\delta/2$. This guarantees an average expected sensitivity of $\Omega(\eta\delta)=\Omega(\eta)$. Since the average sensitivity is lower-bounded by $c_{\mathrm{low}}\eta$, the supremum over the interval must also satisfy

$$\sup_{\mu'\in[0,1]}\mathbb{E}_{X\sim\mathcal{N}(\mu',1)^{\otimes n}}[S_\eta^h(X)] \geq c_{\mathrm{low}}\eta.$$

This establishes the scalar claim, which transfers to the high-dimensional estimator f via the orthogonal projection, completing the proof of Theorem 3.4.

G Proof of Theorem 3.5

Throughout this section, for a scalar parameter $\mu'\in\mathbb{R}$, we write $P_{\mu'} := \mathcal{N}(\mu',1)^{\otimes n}$. By the reduction established in Theorem D.4, to prove the high-dimensional theorem, it suffices to prove the following scalar claim: if the induced one-dimensional estimator $h\colon\mathbb{R}^n\to[0,1]$ satisfies the integrated risk bound

$$\int_0^1 \mathbb{E}_{X\sim P_{\mu'}}[(h(X)-\mu')^2]\,d\mu' \leq \frac{C_{\mathsf{int}}}{n},$$

then, for a corruption budget $k=\lfloor\eta n\rfloor\geq C_{\mathrm{high}}\log n$, its average empirical sensitivity must be large:

$$\sup_{\mu'\in[0,1]}\mathbb{E}_{X\sim P_{\mu'}}[S_\eta^h(X)] \geq c_{\mathrm{high}}\eta.$$

To see exactly why this scalar claim implies the full high-dimensional theorem, recall that Theorem D.4 guarantees the existence of a "good" unit direction $u \in \mathbb{S}^{d-1}$ and an orthogonal shift $\lambda_* \in u^\perp$. When we evaluate the original high-dimensional estimator f along the one-dimensional affine line parameterized by $\mu' u + \lambda_*$, its empirical sensitivity strictly upper-bounds the sensitivity of the induced scalar estimator h. Together with Jensen's inequality, this gives us

$$\mathsf{ES}_{\eta,2}(f; \mu' u + \lambda_*) \geq \left(\mathbb{E}_{X \sim P_{\mu'}}[S_\eta^h(X)^2]\right)^{1/2} \geq \mathbb{E}_{X \sim P_{\mu'}}[S_\eta^h(X)].$$

Taking the supremum over $\mu' \in [0,1]$ on both sides demonstrates that establishing the $c_{\mathrm{high}}\eta$ lower bound for h immediately yields the required lower bound for f, completing the reduction.

We start with the following lemma about endpoint average bound for scalar estimator bound with integrated MSE bound.

Lemma G.1. *Assume $\eta \in \left(0, \frac{1}{10}\right)$. If the scalar estimator $h\colon \mathbb{R}^n \to [0,1]$ satisfies the integrated MSE bound*

$$\int_0^1 \mathbb{E}_{X \sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu' \leq \frac{C_{\mathsf{int}}}{n}$$

Assume moreover that n is large enough such that $C_{\mathsf{high}} \log n \geq 100 C_{\mathsf{int}}$. Then we have the following endpoint average bounds:

$$\frac{1}{\eta}\int_0^\eta a(\mu')\, d\mu' \leq \frac{1}{3} \text{ and } \frac{1}{\eta}\int_{1-\eta}^1 a(\mu')\, d\mu' \geq \frac{2}{3}.$$

Proof. By the integrated MSE assumption,

$$\frac{1}{\eta}\int_0^\eta \mathbb{E}_{X \sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu' \leq \frac{C_{\mathsf{int}}}{\eta n} \leq \frac{1}{100},$$

by our choice of n large enough so that $\eta n \geq k \geq C_{\mathsf{high}} \log n \geq 100 C_{\mathsf{int}}$. Similarly, we obtain

$$\frac{1}{\eta}\int_{1-\eta}^1 \mathbb{E}_{X \sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu' \leq \frac{1}{100}.$$

Now, by noticing that for $\mu' \in [0, \eta]$, we have $a(\mu') \leq \mu' + \mathbb{E}_{X \sim P_{\mu'}}[|h(X) - \mu'|]$. Therefore,

$$\begin{aligned}
\frac{1}{\eta}\int_0^\eta a(\mu')\, d\mu' &\leq \frac{1}{\eta}\int_0^\eta \mu'\, d\mu' + \frac{1}{\eta}\int_0^\eta \mathbb{E}_{X \sim P_{\mu'}}[|h(X) - \mu'|]\, d\mu' \\
&\leq \frac{\eta}{2} + \left(\frac{1}{\eta}\int_0^\eta \mathbb{E}_{X \sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu'\right)^{1/2} \\
&\leq \frac{\eta}{2} + \frac{1}{10} \leq \frac{1}{3}.
\end{aligned}$$

Similarly, for $\mu' \in [1-\eta, 1]$, we have $a(\mu') \geq \mu' - \mathbb{E}_{X \sim P_{\mu'}}[|h(X) - \mu|]$. Therefore,

$$\begin{aligned}
\frac{1}{\eta}\int_{1-\eta}^1 a(\mu')\, d\mu' &\geq \frac{1}{\eta}\int_{1-\eta}^1 \mu'\, d\mu' - \frac{1}{\eta}\int_{1-\eta}^1 \mathbb{E}_{X \sim P_{\mu'}}[|h(X) - \mu'|]\, d\mu' \\
&\geq 1 - \frac{\eta}{2} - \left(\frac{1}{\eta}\int_{1-\eta}^1 \mathbb{E}_{X \sim P_{\mu'}}[(h(X) - \mu')^2]\, d\mu'\right)^{1/2} \\
&\geq 1 - \frac{\eta}{2} - \frac{1}{10} \geq \frac{2}{3}.
\end{aligned}$$

This proves the lemma. □

G.1 Average Mean Displacement

Take n sufficiently large such that the condition in Theorem G.1 holds. We first quantify the global displacement of the estimator's expected mean under a shift of η. Indeed, we have

$$\int_0^{1-\eta} (a(\mu' + \eta) - a(\mu'))\, d\mu' = \int_\eta^1 a(\mu')\, d\mu' - \int_0^{1-\eta} a(\mu')\, d\mu'$$

$$= \int_{1-\eta}^{1} a(\mu')\, d\mu' - \int_{0}^{\eta} a(\mu')\, d\mu' \geq \frac{\eta}{3}.$$

Let $\mathcal{U}$ denote the uniform probability measure over the interval $[0, 1-\eta]$. We have

$$\mathbb{E}_{\mu' \sim \mathcal{U}}[a(\mu'+\eta) - a(\mu')] = \frac{1}{1-\eta} \int_{0}^{1-\eta} (a(\mu'+\eta) - a(\mu'))\, d\mu' \geq \frac{\eta}{3}.$$

G.2 Coordinate-wise Maximal Coupling

For any fixed $\mu' \in [0, 1-\eta]$, we construct a joint distribution over (X, X') via independent coordinatewise maximal couplings as in Theorem A.4 of the marginals $\mathcal{N}(\mu', 1)$ and $\mathcal{N}(\mu'+\eta, 1)$ respectively. In this case, $X \sim P_{\mu'}$ and $X' \sim P_{\mu'+\eta}$. Let $D \coloneqq \#\{i \in [n] \colon X_i \neq X_i'\}$ be the number of coordinates where the samples differ. By Theorem A.4, $D \sim \mathrm{Bin}(n, p)$ where the success probability p satisfies $p < \eta$. Let $k \coloneqq \lfloor \eta n \rfloor$, we see that there exists an absolute constant $\rho \in (0,1)$ such that $\mathbb{E}[D] \leq \rho k$ for all $\eta \in (0, 1/10]$. Therefore, a standard Chernoff bound argument as in Theorem A.5 gives us a constant $c_{\mathsf{Ch}} > 0$ such that $\mathbb{P}(D > k) \leq \exp(-c_{\mathsf{Ch}} k)$. We may choose n and C_{high} large enough such that in the regime $k \geq C_{\mathsf{high}} \log n$, we are guaranteed that $\mathbb{P}(D > k) \leq \eta/6$ for all n.

On the event $\{D \leq k\}$, the datasets X and X' differ by at most k points. Therefore, X' is a valid η-corruption of X. By definition, $S_\eta^h(X) \geq |h(X') - h(X)| \cdot \mathbf{1}\{D \leq k\}$. Since $h \in [0,1]$, we have $|h(X') - h(X)| \leq 1$. Therefore,

$$\begin{aligned}
\mathbb{E}_{X \sim P_{\mu'}}[S_\eta^h(X)] &\geq \mathbb{E}\left[|h(X') - h(X)| \cdot \mathbf{1}\{D \leq k\}\right] \\
&\geq \mathbb{E}[(h(X') - h(X)) \cdot \mathbf{1}\{D \leq k\}] \\
&= \mathbb{E}[h(X') - h(X)] - \mathbb{E}[(h(X') - h(X)) \cdot \mathbf{1}\{D > k\}] \\
&\geq a(\mu'+\eta) - a(\mu') - \mathbb{P}(D > k).
\end{aligned}$$

To conclude the proof, we average this final inequality over the uniform prior $\mathcal{U}$. This gives us

$$\begin{aligned}
\mathbb{E}_{\mu' \sim \mathcal{U}} \mathbb{E}_{X \sim P_{\mu'}}[S_\eta^h(X)] &\geq \mathbb{E}_{\mu' \sim \mathcal{U}}[a(\mu'+\eta) - a(\mu')] - \mathbb{P}(D > k) \\
&\geq \frac{\eta}{3} - \frac{\eta}{6} = \frac{\eta}{6}.
\end{aligned}$$

This implies $\sup_{\mu' \in [0,1]} \mathbb{E}_{X \sim P_{\mu'}}[S_\eta^h(X)] \geq \frac{\eta}{6}$, which is what we wanted. By Theorem D.4, we thus have $\sup_{\mu \in \mathbb{R}^d} \mathsf{ES}_{\eta,2}(f; \mu) \geq \frac{\eta}{6}$, which proves Theorem 3.5.

H Proof of Theorem 3.6

The proof of Theorem 3.6 can be split into two parts. First, we show that uniform MSE accuracy forces the estimator to have output variance $\Omega(d/n)$ at some parameter point. Then we convert this clean-sample variance into a corruption strategy via block resampling.

H.1 Existence of large variance parameter

We first prove the statistical part of the variance obstruction. To do this, we give two proofs. The first is the shorter Cramer-Rao proof, which assumes a regularity condition on the mean map. The second is the Hammersley-Chapman-Robbins proof, which may be applied to broader distribution classes, without using similar regularity assumptions.

Throughout, for $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}$, we write

$$\mathrm{Var}_\mu(f) \coloneqq \mathrm{Var}_\mu(f(X)) \coloneqq \mathbb{E}_X \|f(X) - \mathbb{E}_X f(X)\|^2.$$

Proposition H.1. *Let $f \colon (\mathbb{R}^d)^n \to \mathbb{R}^d$ be measurable and suppose that*

$$\sup_{\mu \in \mathbb{R}^d} \mathbb{E}_X \|f(X) - \mu\|^2 \leq C_{\mathsf{MSE}} \frac{d}{n}.$$

Then there exists a universal constant $c_{\mathrm{var}} > 0$ and a parameter $\mu^* \in \mathbb{R}^d$ such that

$$\mathrm{Var}_{\mu^*}(f) \geq c_{\mathrm{var}} \frac{d}{n}.$$

We will first show two different proofs for the scalar version of Theorem H.1, then we will use the same argument to lift this to the d-dimensional version. Indeed, for the scalar version: consider $g = g(Z_1, \ldots, Z_n)$ under $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu', 1)$, where we write $a(\mu') \coloneqq \mathbb{E}_Z[g(Z)]$ and $\mathrm{Var}_{\mu'}(g) \coloneqq \mathbb{E}_Z[(g - a(\mu'))^2]$.

H.1.1 Proof I of Scalar Version of Theorem H.1

This proof is the Cramer-Rao proof, which requires the standard regularity condition that the mean map $a(\mu')$ is absolutely continuous and satisfies the Gaussian score identity.

Define the bias $b(\mu') \coloneqq a(\mu') - \mu'$. By Jensen's inequality and the uniform MSE accuracy bound, we have

$$|b(\mu')| = |\mathbb{E}_Z[g(Z) - \mu']| \le \left(\mathbb{E}_Z[(g(Z) - \mu')^2]\right)^{1/2} \le \sqrt{\frac{C_{\mathsf{MSE}}}{n}}$$

for every μ'. Therefore, for $L \coloneqq 4 \max\left\{\frac{1}{\sqrt{n}}, \sqrt{\frac{C_{\mathsf{MSE}}}{n}}\right\}$, we have

$$a(L) - a(0) = L + b(L) - b(0) \ge L - 2\sqrt{\frac{C_{\mathsf{MSE}}}{n}} \ge \frac{L}{2}.$$

By Theorem A.3, applied to the Gaussian location family $P_{\mu'} = \mathcal{N}(\mu', 1)^{\otimes n}$ and the statistic $T = g$, the Fisher information is $I(\mu') = n$. Therefore, under the stated regularity assumptions on the mean map $a(\mu') = \mathbb{E}_{\mu'} g$, we have

$$\mathrm{Var}_{\mu'}(g) \ge \frac{(a'(\mu'))^2}{n}.$$

Averaging over $[0, L]$, we obtain

$$\begin{aligned}
\frac{1}{L}\int_0^L \mathrm{Var}_{\mu'}(g)\, d\mu' &\ge \frac{1}{nL}\int_0^L a'(\mu')^2\, d\mu' \\
&\ge \frac{1}{n}\left(\frac{1}{L}\int_0^L a'(\mu')\, d\mu'\right)^2 \\
&= \frac{1}{n}\left(\frac{a(L) - a(0)}{L}\right)^2 \ge \frac{1}{4n}.
\end{aligned}$$

This proves the claim for $d = 1$.

H.1.2 Proof II of Scalar Version of Theorem H.1

We now give the Hammersley-Chapman-Robbins proof. This proof avoids any differentiability assumptions. Let us consider $h \coloneqq \frac{1}{\sqrt{n}}$, $L \coloneqq 4 \max\left\{\sqrt{\frac{C_{\mathsf{MSE}}}{n}}, h\right\}$ and $J \coloneqq \left\lceil \frac{L}{h} \right\rceil$.

Now, we consider the grid points $\mu'_\ell \coloneqq \ell h$ for $\ell = 0, \ldots, J$. Since $\mu'_J = Jh \ge L$ and $|b(\mu')| \le \sqrt{\frac{C_{\mathsf{MSE}}}{n}}$, it follows that

$$a(\mu'_J) - a(\mu'_0) = \mu'_J - \mu'_0 + b(\mu'_J) - b(\mu'_0) \ge L - 2\sqrt{\frac{C_{\mathsf{MSE}}}{n}} \ge \frac{L}{2}.$$

For each $\ell = 0, \ldots, J-1$, we let P_ℓ and Q_ℓ denote the probability distribution $\mathcal{N}(\mu'_\ell, 1)^{\otimes n}$ and $\mathcal{N}(\mu'_{\ell+1}, 1)^{\otimes n}$ respectively. Because $\mu'_{\ell+1} - \mu'_\ell = h = \frac{1}{\sqrt{n}}$, we have

$$\chi^2(Q_\ell \| P_\ell) = \exp(nh^2) - 1 = e - 1.$$

Applying Theorem A.2 with $P = P_\ell = \mathcal{N}(\mu'_\ell, 1)^{\otimes n}, Q = Q_\ell = \mathcal{N}(\mu'_{\ell+1}, 1)^{\otimes n}$ and $T = g$, we obtain

$$\mathrm{Var}_{\mu'_\ell}(g) \ge \frac{(a(\mu'_{\ell+1}) - a(\mu'_\ell))^2}{e - 1}.$$

Summing over $\ell = 0, \dots, J-1$ and apply Cauchy Schwarz gives us

$$\sum_{\ell=0}^{J-1} \mathrm{Var}_{\mu'_\ell}(g) \geq \frac{1}{e-1} \sum_{\ell=0}^{J-1} (a(\mu'_{\ell+1}) - a(\mu'_\ell))^2 \geq \frac{1}{(e-1)J} (a(\mu'_J) - a(\mu'_0))^2 \geq \frac{L^2}{4(e-1)J}.$$

Consequently, the average variance satisfies

$$\frac{1}{J+1} \sum_{\ell=0}^{J} \mathrm{Var}_{\mu'_\ell}(g) \geq \frac{L^2}{4(e-1)J(J+1)} \geq \frac{2}{15(e-1)} \cdot \frac{1}{n}$$

This proves the claim for $d = 1$.

H.1.3 Proof of Theorem H.1

We will now state the following general lemma that allows one to lift our argument from one dimension to higher dimension, as long as one can get an averaging measure over scalar parameters under which the expected output variance is $\Omega(1/n)$.

Lemma H.2. *Fix $C_{\mathsf{MSE}} > 0$ and let $C_0 \coloneqq C_{\mathsf{MSE}} d$. Suppose there exists a probability measure ν_{C_0} on $\mathbb{R}$ and a universal constant $c_{\mathrm{sc}} > 0$ with the following property: for every probability space carrying an auxiliary random element W whose law does not depend on θ, and for every real-valued statistic $g = g(Z_1, \dots, Z_n, W)$ where $Z_1, \dots, Z_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta, 1)$ independently of W, if $\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(g-\theta)^2] \leq C_0/n$, then*

$$\int \mathrm{Var}_\theta(g)\, d\nu_{C_0}(\theta) \geq c_{\mathrm{sc}}/n,$$

where the variance is taken over both the data and W.

Then, any measurable estimator $f : (\mathbb{R}^d)^n \to \mathbb{R}^d$ satisfying $\sup_{\mu \in \mathbb{R}^d} \mathbb{E}_\mu \|f(X) - \mu\|_2^2 \leq C_{\mathsf{MSE}} \frac{d}{n}$ has a parameter $\mu^* \in \mathbb{R}^d$ such that $\mathrm{Var}_{\mu^*}(f) \geq c_{\mathrm{sc}} \frac{d}{n}$, where $\mathrm{Var}_\mu(f) \coloneqq \mathbb{E}_\mu \|f(X) - \mathbb{E}_\mu f(X)\|_2^2 = \sum_{j=1}^d \mathrm{Var}_\mu(f_j(X))$.

Proof. Let $C_0 \coloneqq C_{\mathsf{MSE}} d$, and let $\nu \coloneqq \nu_{C_0}$ be the scalar averaging measure supplied by the assumption. We average the parameters $\mu \in \mathbb{R}^d$ under the product measure $\nu^{\otimes d}$.

Fix a coordinate $j \in [d]$, and fix the remaining coordinates $\mu_{-j} = (\mu_1, \dots, \mu_{j-1}, \mu_{j+1}, \dots, \mu_d) \in \mathbb{R}^{d-1}$. Consider the scalar submodel $\mu(\theta) \coloneqq (\mu_1, \dots, \mu_{j-1}, \theta, \mu_{j+1}, \dots, \mu_d)$ for $\theta \in \mathbb{R}$. For $X \sim \mathcal{N}(\mu(\theta), I_d)^{\otimes n}$, let $Z_i \coloneqq X_{i,j}$ for $i \in [n]$, and let $W \coloneqq (X_{i,\ell})_{1 \leq i \leq n,\, \ell \neq j}$ denote all other coordinates of the sample. Then $Z_1, \dots, Z_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta, 1)$ and W is independent of Z. Crucially, because μ_{-j} is fixed, the law of W does not depend on θ.

Define the scalar statistic $g_{j,\mu_{-j}}(Z_1, \dots, Z_n, W) \coloneqq f_j(X)$. For every $\theta \in \mathbb{R}$, the uniform MSE assumption implies that the error of this single coordinate is bounded by the total vector error: $\mathbb{E}_{\mu(\theta)}[(g_{j,\mu_{-j}} - \theta)^2] \leq \mathbb{E}_{\mu(\theta)} \|f(X) - \mu(\theta)\|_2^2 \leq C_{\mathsf{MSE}} \frac{d}{n} = \frac{C_0}{n}$. Therefore, the scalar averaging principle applies, yielding $\int \mathrm{Var}_{\mu(\theta)}(f_j(X))\, d\nu(\theta) \geq c_{\mathrm{sc}}/n$ for every fixed μ_{-j}. Averaging this inequality over $\mu_{-j} \sim \nu^{\otimes(d-1)}$ via Fubini's theorem yields $\mathbb{E}_{\mu \sim \nu^{\otimes d}}[\mathrm{Var}_\mu(f_j(X))] \geq c_{\mathrm{sc}}/n$.

Summing over all d coordinates gives the total expected variance:

$$\mathbb{E}_{\mu \sim \nu^{\otimes d}}[\mathrm{Var}_\mu(f)] = \sum_{j=1}^d \mathbb{E}_{\mu \sim \nu^{\otimes d}}[\mathrm{Var}_\mu(f_j(X))] \geq c_{\mathrm{sc}} \frac{d}{n}.$$

Since the expected variance under $\mu \sim \nu^{\otimes d}$ is at least $c_{\mathrm{sc}} d/n$, there must exist at least one parameter $\mu^* \in \mathbb{R}^d$ such that $\mathrm{Var}_{\mu^*}(f) \geq c_{\mathrm{sc}} d/n$. □

To finish this, we may instantiate the above lifting principle theorem with two different scalar proof that we have established. The Cramer-Rao scalar proof proves the scalar averaging principle using the continuous measure $\nu_{C_0} = \mathrm{Unif}([0, L])$, $c_{\mathrm{sc}} = \frac{1}{4}$, where $L = 4 \max\{\sqrt{C_0/n}, 1/\sqrt{n}\}$, provided that the scalar mean map satisfies the Cramer Rao regularity assumptions.

Alternatively, the Hammersley–Chapman–Robbins scalar proof establishes the same principle using the discrete measure $\nu_{C_0} = \mathrm{Unif}\{0, h, 2h, \dots, Jh\}$ and $c_{\mathrm{sc}} = \frac{2}{15(e-1)}$, where $h \coloneqq 1/\sqrt{n}$, $L = 4\max\{\sqrt{C_0/n}, h\}$, and $J \coloneqq \lceil L/h \rceil$.

This fully establishes Theorem H.1.

H.2 Conversion into Corruption Strategy

Fix a parameter μ^* from Theorem H.1, and let $X = (X_1, \dots, X_n) \sim \mathcal{N}(\mu^*, \mathbb{I}_d)^{\otimes n}$. Let $k = \lfloor \eta n \rfloor \geq 1$. Partition $[n]$ into $M \coloneqq \lceil n/k \rceil$ disjoint nonempty blocks $B_1, \dots, B_M$ of sizes at most k. Let $Y_i = (X_\ell)_{\ell \in B_i}$ denote the data in block B_i. Equivalently, we may view X as the collection $(Y_1, \dots, Y_M)$. Let Y_i' be an independent fresh copy of Y_i, independent of all other variables, and let $X^{(i)}$ denote the dataset obtained from X by replacing block Y_i by Y_i'.

We note that by definition, each $X^{(i)}$ differs from X in at most k sample positions. Therefore, pointwise, we have

$$S_\eta^f(X)^2 \geq \|f(X) - f(X^{(i)})\|^2, \quad \forall i = 1, 2, \dots, M.$$

By vector Efron-Stein inequality in Theorem A.1: for independent blocks $Y_1, \dots, Y_M$ and independent replacements Y_i', we have

$$\mathrm{Var}_{\mu^*}(f(X)) \leq \frac{1}{2} \sum_{i=1}^{M} \mathbb{E}\|f(X) - f(X^{(i)})\|^2.$$

Hence, we see that

$$\max_{1 \leq i \leq M} \mathbb{E}\|f(X) - f(X^{(i)})\|^2 \geq \frac{2}{M} \mathrm{Var}_{\mu^*}(f(X)).$$

Combining this with Theorem H.1, we see that there exists some constant $c_0 > 0$ such that

$$\mathbb{E}_X[S_\eta^f(X)^2] \geq c_0 \frac{\eta d}{n}.$$

Therefore, we have

$$\sup_{\mu \in \mathbb{R}^d} \mathsf{ES}_{\eta,2}(f; \mu) = \sup_{\mu \in \mathbb{R}^d} \left(\mathbb{E}_X[S_\eta^f(X)^2]\right)^{1/2} \geq \sqrt{c_0} \sqrt{\frac{\eta d}{n}},$$

and taking $C_5 \coloneqq \sqrt{c_0}$ proves Theorem 3.6.

I Proof of Theorem 4.1

Throughout this section, for $t \in \{0, \dots, n\}$, define the Hamming-weight layer $\mathsf{Layer}(t) \coloneqq \{x \in \{0,1\}^n : |x| = t\}$. Let U_t denote a uniformly random element of $\mathsf{Layer}(t)$.

Lemma I.1. *Let $P \sim \mathsf{Unif}[0,1]$ and, conditional on $P = p$, sample $X \sim \mathrm{Bern}(p)^{\otimes n}$. Let $T \coloneqq |X| = \sum_{i=1}^n X_i$. Then $\mathbb{P}(T = t) = \frac{1}{n+1}$ for all $t \in \{0, \dots, n\}$. Therefore, conditional on $T = t$, X is uniform on $\mathsf{Layer}(t)$.*

Proof. For $t \in \{0, \dots, n\}$, we have

$$\mathbb{P}(T = t) = \int_0^1 \binom{n}{t} p^t (1-p)^{n-t} \, dp = \binom{n}{t} \frac{t!(n-t)!}{(n+1)!} = \frac{1}{n+1}.$$

□

Lemma I.2. *Fix integers t, ℓ with $0 \leq t \leq n - \ell$. Sample U_t uniformly from $\mathsf{Layer}(t)$. Conditional on $U_t = x$, choose an ℓ-subset I uniformly among the zero coordinates of x and flip those coordinates to 1. Let V be the resulting vector. Then $V \sim \mathsf{Unif}(\mathsf{Layer}(t + \ell))$ and $d_H(U_t, V) = \ell$.*

Proof. The Hamming-distance claim is immediate because exactly ℓ zero coordinates are flipped to one. It remains to prove uniformity. Fix $y \in \mathsf{Layer}(t+\ell)$. A pair (x, I) maps to y exactly when I is an ℓ-subset of the one coordinates of y, and x is obtained from y by changing the coordinates in I back to zero. Thus, every $y \in \mathsf{Layer}(t+\ell)$ has exactly $\binom{t+\ell}{\ell}$ preimages. Since the sampling probability of each valid pair (x, I) is the same, V is thus uniform on $\mathsf{Layer}(t+\ell)$. □

We are now ready to prove Theorem 4.1. Indeed, we may replace $f = \mathrm{clip}_{[0,1]} \circ f$ by Theorem D.3.

Let $\varepsilon \coloneqq \frac{C_{\mathsf{Bern}}}{\sqrt{n}}$. Choose a constant $\alpha = \alpha(C_{\mathsf{Bern}}) \in (0, 1/4]$ small enough that $2\alpha + \frac{2C_{\mathsf{Bern}}}{c_*}\sqrt{\alpha} \le \frac{1}{2}$. Let $k \coloneqq \lfloor \eta n \rfloor$ and define $\ell \coloneqq \min\{k, \lfloor \alpha n \rfloor\}$. For all sufficiently large n, depending only on α, the assumptions $k \ge 1$ and n sufficiently large imply $1 \le \ell \le k$ and $\ell \le \frac{n}{2}$.

For each $t \in \{0, \dots, n\}$, define $a_t \coloneqq \mathbb{E}[f(U_t)]$. Fix $t \in \{0, \dots, n-\ell\}$. By Theorem I.2, we can couple U_t and $U_{t+\ell}$ so that $d_H(U_t, U_{t+\ell}) = \ell \le k$. Hence $\mathbb{E}[S_\eta^f(U_t)] \ge \mathbb{E}[|f(U_{t+\ell}) - f(U_t)|] \ge |a_{t+\ell} - a_t|$. Therefore, we have

$$\sum_{t=0}^{n-\ell} \mathbb{E}[S_\eta^f(U_t)] \ge \sum_{t=0}^{n-\ell} |a_{t+\ell} - a_t|.$$

Decompose RHS into arithmetic progressions modulo ℓ. For each residue $r \in \{0, \dots, \ell-1\}$, write $r, r+\ell, \dots, r+T_r\ell$, where $T_r \coloneqq \lfloor \frac{n-r}{\ell} \rfloor$ and $j_r \coloneqq r + T_r\ell$. Then $j_r \in \{n-\ell+1, \dots, n\}$. By the triangle inequality along this arithmetic progression, $\sum_{q=0}^{T_r-1} |a_{r+(q+1)\ell} - a_{r+q\ell}| \ge |a_{j_r} - a_r|$. Summing over $r = 0, \dots, \ell-1$ yields

$$\sum_{t=0}^{n-\ell} |a_{t+\ell} - a_t| \ge \sum_{r=0}^{\ell-1} |a_{j_r} - a_r|.$$

We now lower-bound the boundary gaps. Fix $r \in \{0, \dots, \ell-1\}$ and set $p = \frac{r}{n}$. If $X \sim \mathrm{Bern}(p)^{\otimes n}$ and $T = |X|$, then conditional on $T = r$, X is uniform on $\mathsf{Layer}(r)$. Therefore,

$$\varepsilon \ge \mathbb{E}[|f(X) - p|] \ge \mathbb{P}(T = r)\mathbb{E}[|f(U_r) - p|] \ge \mathbb{P}(T = r)|a_r - p|.$$

We know that since $T \sim \mathsf{Bin}(n, p)$, Theorem A.6 gives us $\mathbb{P}(T = r) \ge \frac{c_*}{\sqrt{r+1}} \ge \frac{c_*}{\sqrt{\ell}}$ for some positive constant c_*. Therefore, $|a_r - r/n| \le \frac{\varepsilon}{c_*}\sqrt{\ell}$. Since $r \le \ell - 1$, we have $a_r \le \frac{\ell}{n} + \frac{\varepsilon}{c_*}\sqrt{\ell}$. Similarly, fix $j \in \{n-\ell+1, \dots, n\}$ and write $j = n - s$ with $0 \le s \le \ell - 1$. Set $p = 1 - \frac{s}{n}$. By the same argument, we have $|a_j - j/n| \le \frac{\varepsilon}{c_*}\sqrt{\ell}$, and thus we get $a_j \ge 1 - \frac{\ell}{n} - \frac{\varepsilon}{c_*}\sqrt{\ell}$.

Combining both of these, we see that for every $r \in \{0, \dots, \ell-1\}$, $|a_{j_r} - a_r| \ge 1 - \frac{2\ell}{n} - \frac{2\varepsilon}{c_*}\sqrt{\ell}$. Because $\ell \le \alpha n$ and $\varepsilon = \frac{C_{\mathsf{Bern}}}{\sqrt{n}}$, our choice of α guarantees that $|a_{j_r} - a_r| \ge \frac{1}{2}$ for all residues r. Substituting this back into our telescoping sums give us

$$\sum_{t=0}^{n-\ell} \mathbb{E}[S_\eta^f(U_t)] \ge \frac{\ell}{2}.$$

To conclude, let $P \sim \mathsf{Unif}[0, 1]$ and, conditional on $P = p$, sample $X \sim \mathrm{Bern}(p)^{\otimes n}$. By Theorem I.1,

$$\int_0^1 \mathbb{E}_{X \sim \mathrm{Bern}(p)^{\otimes n}}[S_\eta^f(X)]\, dp = \frac{1}{n+1}\sum_{t=0}^{n} \mathbb{E}[S_\eta^f(U_t)] \ge \frac{\ell}{2(n+1)}.$$

Therefore this also gives us

$$\sup_{p \in [0,1]} \mathbb{E}_{X \sim \mathrm{Bern}(p)^{\otimes n}}[S_\eta^f(X)] \ge \frac{\ell}{2(n+1)}.$$

It remains to compare $\frac{\ell}{n+1}$ to η. Indeed, suppose $k \le \lfloor \alpha n \rfloor$. Then $\ell = k$. In this case, $\frac{k}{n} \ge \frac{\eta}{2}$, and thus $\frac{\ell}{2(n+1)} \ge \frac{\eta}{8}$. Otherwise, in the case $k > \lfloor \alpha n \rfloor$, then $\ell = \lfloor \alpha n \rfloor$. For all sufficiently large n, we have $\ell \ge \frac{\alpha n}{2}$ which therefore gives us the bound $\frac{\ell}{2(n+1)} \ge \frac{\alpha}{8} \ge \frac{\alpha}{8}\eta$. This shows us

$$\sup_{p \in [0,1]} \mathbb{E}_{X \sim \mathrm{Bern}(p)^{\otimes n}}[S_\eta^f(X)] \ge \frac{\alpha}{8}\eta.$$

Therefore, this theorem holds with $C_6 \coloneqq \frac{\alpha}{8}$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We define the notion of empirical sensitivity, and show lower bounds for Gaussian mean estimation, and an upper bound matching the lower bound up to logarithmic factors via robust empirical mean recovery.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Our results are theoretical, and we clearly state the assumptions and conditions they apply.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions are explicitly stated in the formulation of the formal results themselves. Proof sketches are given and all formal proofs are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not include experiments.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: None of the listed concerns are applicable.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no societal impact of the work performed. The work described in this paper is purely theoretical and not tied to any particular applications.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: There is no experiments performed in this paper.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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