Multi-Observation Regression

Rafael Frongillo
CU Boulder

Nishant A. Mehta
University of Victoria

Tom Morgan
Harvard University

Bo Waggoner
Microsoft Research

Abstract

Given a data set of (x, y) pairs, a common learning task is to fit a model predicting y (a label or dependent variable) conditioned on x. This paper considers the similar but much less-understood problem of modeling "higher-order" statistics of y's distribution conditioned on x. Such statistics are often challenging to estimate using traditional empirical risk minimization (ERM) approaches. We develop and theoretically analyze an ERM-like approach with multi-observation loss functions. We propose four algorithms formalizing the concept of ERM for this problem, two of which have statistical guarantees in settings allowing both slow and fast convergence rates, but which are out-performed empirically by the other two. Empirical results illustrate potential practicality of these algorithms in low dimensions and significant improvement over standard approaches in some settings.

1 Introduction

In the common learning task of regression, one fits a model to a data set of (x, y) pairs in order to form a prediction about y from x. For each x, we assume y is drawn from an unknown distribution $\mathcal{D}_x$, and the model's prediction is generally some statistic of $\mathcal{D}_x$. The canonical examples, of course, are least-squares regression, where the prediction is the mean of y given x, and logistic regression, where one predicts the probability of y given x. More generally, one also has quantile regression (for example by minimizing absolute error rather than squared error), superquantile (or conditional value at risk) regression [1], mode regression [2], and so on.

In many settings in engineering, social and natural sciences, and finance, one would like to fit a model to some *higher-order statistic* of $\mathcal{D}_x$. Such properties include measures of risk (such as conditional value at risk), uncertainty, inequality, and other statistics depending on the spread of y given the features x. As a stylized example, consider the relationship between regional biodiversity (y) and climatic features, say average temperature and precipitation (x). Here, observations collected from citizen scientists will be of the form (x, y) where $x \in \mathbb{R}^2$ denotes the above features and $y \sim \mathcal{D}_x$ is a categorical variable for the species observed. One common measure of biodiversity is $\|\mathcal{D}_x\|_2 = \sqrt{\sum_y \mathcal{D}_x(y)^2}$, the 2-norm of the distribution of species [3, 4]. So, given a data set of (x, y) pairs, we wish to regress the 2-norm of the species distribution against x.

A natural approach to solve this problem is to adapt the empirical risk minimization (ERM) paradigm, which selects a model or hypothesis $f(x)$ by minimizing some loss function over the data set S:

$$\operatorname*{argmin}_{f \in \mathcal{F}} \sum_{(x,y) \in S} \ell(f(x), y).$$

Standard regression problems follow this approach, typically using squared loss $\ell(f(x), y) = (y - f(x))^2$. For most higher-order properties such as the 2-norm or conditional variance, however, the theory of elicitation tells us that no such loss function can be statistically consistent [5]. This raises the main question of this paper: how, algorithmically, to fit such models to data if the ERM paradigm is apparently unavailable.

One common work-around is to use surrogate losses to model e.g. the entire distribution of y given x, and then use these to derive a model for the statistic of interest. In this paper, both lower bounds and simulation results demonstrate that this traditional approach can often do quite poorly. For intuition, we revisit our biodiversity example. Here $\mathcal{D}_x$ (the distribution of species given environmental features) is likely extremely complicated, requiring a high sample complexity or access to additional features. Yet intuitively, this should not be required to uncover simple relationships, e.g.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

that biodiversity increases with rainfall from deserts to rainforests.

This paper proposes to fit such models directly using *multi-observation losses*: losses of the form $\ell(f(x), y_1, \ldots, y_m)$, introduced in [6]. Unlike standard "single-observation" loss functions, these losses *can* be statistically consistent for higher-order properties like variance and 2-norm, when one has multiple independent samples of y for each x. For example, the two-observation loss $\ell(f(x), y_1, y_2) = (f(x)^2 - \mathbb{1}\{y_1 = y_2\})^2$ is statistically consistent for the 2-norm of the distribution of y given x. We will show how to use such multi-observation losses to perform ERM directly on higher-order statistics.

Contributions. First, we propose a general paradigm for performing ERM using multi-observation loss functions and present four algorithms. The key challenge is that ERM for such losses $\ell(f(x), y_1, \ldots, y_m)$ requires m independent observations y_i for each input example x, but we only have access to pairs (x, y). In our algorithms, many samples $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ are clumped together into *metasamples* of the form $(x, y_1, \ldots, y_m)$, and then ERM is performed directly on a multi-observation loss via

$$\operatorname*{argmin}_{f \in \mathcal{F}} \sum_{(x, y_1 \ldots y_m) \in S} \ell(f(x), y_1, \ldots, y_m).$$

Our first two algorithms are, in a sense, *unbiased* and for these we prove, under a natural Lipschitz "slowly-changing" assumption on $\mathcal{D}_x$, statistical convergence guarantees with either slow or fast rates depending on the setting. Thus, we give the first excess risk bounds for ERM with multi-observation losses, aside from the very preliminary (and much weaker) bounds of [6]. Our key technique is an analysis of ERM with *corrupted samples*, i.e. samples from distributions near but not equal to the same underlying distribution. We consider both labeled and unlabeled sample complexity, an important distinction in our paradigm. Our algorithms are most practical in the low-dimensional regime; we show information-theoretic hardness in high dimensions without further assumptions.

Finally, we demonstrate the advantage of our approach in some settings over traditional single-observation approaches, for problems such as predicting conditional variance. We give both theoretical lower bounds and empirical examples; empirically, our other two (biased) algorithms perform best, though we do not have theoretical results for them. While we often use the 2-norm or variance as expository examples, our results are fully general, encompassing myriad other higher-order properties elicitable via multi-observation losses.

Applications. Aside from augmenting the literature on fundamental properties of ERM, our results may have applications to engineering, social sciences, finance, ecology, and beyond; we briefly describe some of these settings. In engineering, the design of an airfoil, building, truss, etc., is often done by choosing design parameters minimizing some objective (drag, cost, weight, displacement), but which are robust to changes in the environment or to manufacturing defects [7, 8], as quantified by some *risk measure*, such as the 95% quantile of the drag for an airfoil under random initial conditions [9]. A common technique to perform this minimization is surrogate optimization, wherein one first fits a model and then optimizes [10, 9]. We show that two popular risk measures for robust engineering design, the upper confidence bound [11] and MINVAR [12], are easily fit with multi-observation losses. In finance, risk measures are also used, both to make decisions and to regulate financial institutions, where the risk is a higher-order statistic of the distribution of financial losses on any given day [13, 14]. As we cannot observe multiple "i.i.d." monetary losses for the same day, the techniques in this paper would be useful in inference for decision making and statistical tests for regulation. Finally, in social sciences, many statistics of interest capture higher-order properties like diversity. Examples in economics include the Gini coefficient, a measure of inequality, and the Herfindahl–Hirschman Index (HHI), a measure of market concentration equal to squared 2-norm.

2 Empirical risk minimization: from classical samples to metasamples

In the classical supervised learning setup an algorithm is presented with an i.i.d. sample of n labeled points $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the objective of selecting an action $f \in \mathcal{F}$ that obtains low expected loss, or risk, $\mathbb{E}_{(X,Y) \sim \mathcal{D} \times \mathcal{D}_X} [\ell_f(X, Y)]$ with respect to a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. Here, we use the notation $\ell_f(X, Y) = \ell(f(X), Y)$, so that each f is a function mapping from $\mathcal{X}$ to $\mathbb{R}$, and we denote by $f^* \in \mathcal{F}$ the risk minimizer over $\mathcal{F}$. We always assume that the loss ℓ is L-Lipschitz in its first argument. ERM, which returns any $\hat{f} \in \mathcal{F}$ that minimizes the empirical risk $\frac{1}{n} \sum_{i=1}^{n} \ell_f(X_i, Y_i)$, is a natural choice for solving this problem. The performance of ERM is known to be tightly characterized by the notion of Rademacher complexity.

Definition 1. *Let $\mathcal{G}$ be a class of functions mapping from a space $\mathcal{Z}$ to $\mathbb{R}$, and let $Z_1, \ldots, Z_n$ be an i.i.d. sample from distribution P over $\mathcal{Z}$. Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables (distributed uniformly on $\{-1, 1\}$). The* Rademacher complexity *of $\mathcal{G}$ (with respect to P) is*

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}\left[\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(Z_i)\right]\right].$$

In the above, we may for instance take the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and class $\mathcal{G} = \{\ell_f : f \in \mathcal{F}\}$.

The following uniform convergence result is well-known; a proof appears in § A for completeness.

Lemma 1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent samples from distribution P and assume, for all $f \in \mathcal{F}$, that $\ell(f(X), Y) \in [0, B]$ almost surely. Further assume that the loss ℓ is L-Lipschitz in its first argument. With probability at least $1 - \delta$,*

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)\right\} \leq 2L\mathcal{R}_n(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2n}}.$$

In particular, the upper bound holds for the empirical risk minimizer $\hat{f}$. A straightforward argument then leads to a high probability bound on the excess risk of $\hat{f}$ (the risk of $\hat{f}$ minus the risk of f^*).

ERM for multi-observation losses. A direct application of the above analysis to multi-observation loss functions would require samples of the form $(X, Y_1, \ldots, Y_m)$, with each Y_i drawn i.i.d. from the conditional distribution $\mathcal{D}_X$; we refer to such tuples as *metasamples*. Unfortunately, in classical supervised learning we are only provided with samples of the form (X, Y). Nevertheless, if features x and x' are similar (e.g. $\|x - x'\| \leq \varepsilon$), then we often expect that D_x and $D_{x'}$ are also similar (e.g. total variation distance $K\varepsilon$). We will formalize this in § 4 as Assumption (A1). It will allow us to recover the kinds of guarantees satisfied by traditional ERM; at a high-level, the approach is to group samples having features close to X to construct metasamples $(X, Y_1, \ldots, Y_m)$, where each Y_j is sampled from a distribution *approximately* equal to $\mathcal{D}_X$.

3 Algorithms

For simplicity, we take $\mathcal{X} = [0, 1]^d$ in this section and the next. The problem setting we consider is slightly unusual in terms of the *number* and *kinds* of samples used. All of our algorithms "clump" together groups of m different (x_i, y_i) pairs having nearby x values to create a metasample of the form $(x, y_1, \ldots, y_m)$. In this paper, n will always denote the number of metasamples constructed and used by a particular algorithm. This differs from N, the total number of data points drawn by the algorithm. We focus on empirical risk minimization over the metasamples,

$$\operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell_f(x_i, y_{i,1}, \ldots, y_{i,m}).$$

Therefore, the algorithmic questions are (1) how to draw or choose samples, and (2) how to construct metasamples. Once addressed, in § 4 we will present theoretical risk guarantees for some of these algorithms.

Learning paradigms. Our algorithms will apply in two different paradigms. In *supervised learning*, the algorithms draw N data points of the form (x, y) i.i.d., construct n metasamples, and then run ERM on the metasamples. The *sample complexity* is N. In *pool-based active learning*, the algorithms draw N unlabeled x points. They may query up to one label y for each x, drawn independently from $\mathcal{D}_x$. This results in a smaller number of labeled pairs (x, y), from which the algorithms construct the metasamples. For those of our algorithms with theoretical guarantees, the *label complexity* will always be nm, because every label we draw is used in exactly one metasample.[1] We do not consider (fully) *active learning*, where algorithms may repeatedly choose any x, query it to obtain an independent draw $y \sim \mathcal{D}_x$, and repeat. There, traditional algorithms and guarantees will generally carry over to the multi-observation setting, as one can query as many i.i.d. observations from $\mathcal{D}_x$ as desired.

3.1 Unbiased algorithms

We first present algorithms for which we will later be able to prove risk bounds, by ensuring that the x values in the metasamples are i.i.d. samples from $\mathcal{D}$. The Naïve algorithm, Algorithm 1, starts by drawing n i.i.d. data points $X_1^*, \ldots, X_n^*$ and using them as the basis for a metasample. For each X_i^*, it then draws many new data points, so that with high probability, enough points come close enough to form a good metasample. It then moves on to the next X_{i+1}^*. In the sequel, the notations $\tilde{O}$ and $\tilde{\Omega}$ omit log factors, including $\log\frac{1}{\delta}$.

In § B we present a result (Lemma 3) which states that if $N = \tilde{\Omega}(mn^{(d+3)/2}d^{d/2})$, then with probability at least $1-\delta$, most of the points $(X_j^*)_{j\in[n]}$ have their m nearest neighbors all within a proximity of $\frac{1}{\sqrt{n}}$. As we will see shortly, this algorithm can be greatly improved. Nevertheless, this algorithm and its analysis arguably are already interesting. First, note that the guarantee is fully general, holding for *any* probability distribution

[1] This is not always true of other algorithms described below and tested in simulations, where the number of metasamples is larger compared to the number of labels because each y may appear in multiple metasamples.

$\mathcal{D}$ over $\mathcal{X} = [0,1]^d$ while simultaneously providing a guarantee that holds with high probability (i.e. paying only $\log \frac{1}{\delta}$ in sampling complexity for failure probability δ). In contrast, similar results from the k-nearest neighbor (k-NN) literature that hold for arbitrary distributions only hold in expectation or asymptotically [15, 16]. There are, however, results from the k-NN literature which hold with high probability, but these works impose additional regularity assumptions [17, 18], a typical one being that the probability density of $\mathcal{D}$ (with respect to Lebesgue measure) is lower bounded by a positive constant on its support.

The key idea in the proof of Lemma 3 is to partition $\mathcal{X}$ into "heavy" and "light" cells according to the probability mass of each and use Hoeffding's inequality to bound the number of X_j^* points in light cells. This idea is later leveraged as one part of the proof of our guarantee for our improved algorithm.

To improve on Naïve Sampling, Algorithm 2 iteratively draws batches of new samples and finds a globally good assignment to the original base sample.[2] The following lemma guarantees this algorithm's performance for $N = \tilde{\Omega}\left(mn^{(d+1)/2}d^{(d+2)/2}\right)$.

Lemma 2 (Improved Nonuniform Sampling Lemma)**.** *Let $d \in \mathcal{N}$ and let $x_1^*, \ldots, x_n^*, x_1, \ldots, x_N$ be drawn independently from an arbitrary distribution $\mathcal{D}$ on $\mathcal{X} = [0,1]^d$. If $N = \tilde{\Omega}\left(mn^{(d+1)/2}d^{(d+2)/2}\right)$, then with probability at least $1-\delta$, there is a set $\mathcal{J} \subseteq [n]$ with $|\mathcal{J}| \geq n - \sqrt{(n \log \frac{2}{\delta})/2}$ and $|\mathcal{J}|m$ distinct indices $\{i(j,k) \in [N] : j \in \mathcal{J}, k \in [m]\}$, such that for all $j \in \mathcal{J}, k \in [m]$ we have $\|x_j^* - x_{i(j,k)}\|_2 \leq \frac{1}{\sqrt{n}}$.*

The full version of this result, Lemma 5, is stated and proved in § B. The idea of the proof is to adopt the high-level idea of the proof of Lemma 3 — the partitioning of $\mathcal{X}$ into heavy and light cells — while using a Poissonization argument to show that heavy cells have enough samples, i.e. at least m x_i's for every x_j^*. This argument is also adapted to obtain a specialized result, Lemma 4 in § B, with considerably better sample complexity for the uniform distribution.

We would like to again stress that Lemma 2 is fully general in that it holds for *any* probability distribution $\mathcal{D}$ over $\mathcal{X} = [0,1]^d$ and yet also provides a high probability guarantee. Leveraging results from the k-NN literature would require either giving up on the high probability nature of our guarantees or restricting the class of distributions for which the result holds.

3.2 Biased algorithms

We now briefly consider algorithms that we feel are likely to perform well in practice and indeed do so in our simulations. These algorithms construct a larger number of metasamples by reusing labels, thereby giving them access to more information but making theoretical guarantees very difficult. For simplicity, we describe both algorithms for single dimensional $\mathcal{X}$, but they can be generalized to higher dimensions at the expense of computational complexity. The first algorithm, Sliding Window, simply iterates from left to right over the x-values on the real line and creates a metasample from each group of m adjacent points.The second, dubbed "ε-Nearby", sets a fixed upper distance limit ε and constructs a metasample from all m-tuples of data points whose x-values lie in an interval of diameter ε. Pseudocode for both algorithms is in § C; we demonstrate their performance in § 5.2. We found that ε-Nearby performed slightly better, but Sliding Window is free of tuning parameters.

4 Risk bounds

Let $\mathcal{D}$ be a probability distribution over $\mathcal{X} \subset \mathbb{R}^d$, and, for each $x \in \mathcal{X}$, let $\mathcal{D}_x$ be the conditional distribution over $\mathcal{Y}$ given x. We take $\mathcal{X} = [0,1]^d$ for simplicity. In this section, we assume that loss values lie in $[0, B]$. As discussed above, to make headway we will relate the conditional distributions D_x for nearby x; formally, we make a Lipschitz assumption on their total variation distance,

$$D_{TV}(\mathcal{D}_x, \mathcal{D}_{x'}) \leq K\|x - x'\|_2. \tag{A1}$$

Intuitively, this means that samples from nearby conditional distributions are almost interchangeable.

4.1 Excess risk bounds for general situations

Imagine an ideal setting where we have i.i.d. points $X_1^*, \ldots, X_n^*$ and, for each i, we are given m i.i.d. labels $Y_{i,1}, \ldots, Y_{i,m}$ sampled from $\mathcal{D}_{X_i^*}$. Then, for purposes of analysis we could treat these labels as a single "mega-label" $\mathbb{Y}_i = (Y_{i,1}, \ldots, Y_{i,m})$. We would have a set of i.i.d. data points of the form $(X_i^*, \mathbb{Y}_i)$ together with a loss $\ell(f(X_i^*), \mathbb{Y}_i)$, and so we could directly apply existing analyses of ERM. The idea of our analysis is to relate the performance of our algorithms, which must construct their own "noisy" metasamples from imperfect data, to this ideal. We give results under both the standard notion of sample complexity from supervised learning and the (much smaller) label complexity from pool-based active learning.

A key idea in the analysis is to view each metasample as being drawn in this idealized fashion (i.e. each $Y_{i,j} \sim$

[2] Note that metasamples constructed conditional on X still ensure that Y is drawn independently from $\mathcal{D}_X$.

Algorithm 1: Naïve Sampling

```
Input: $n, m, N \in \mathbb{N}$
Sample $n$ points $X_1^*, \ldots, X_n^*$ indep. from $\mathcal{D}$
for $i = 1$ to $n$ do
    Sample $k := N/n$ points $X_1^{(i)}, \ldots, X_k^{(i)}$ independently from $\mathcal{D}$
    for $j = 1$ to $m$ do
        Set $X_{i,j}$ to be the $j^{\text{th}}$ nearest neighbor of $X_i^*$ among $(X_j^{(i)})_{j \in [k]}$, with ties broken arbitrarily
        Sample a label $\tilde{Y}_{i,j} \sim \mathcal{D}_{X_{i,j}}$
return $\hat{f} = \mathrm{ERM}_{\mathcal{F},\ell}\left((X_i^*, (\tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,m}))_{i \in [n]}\right)$
```

Algorithm 2: Improved Sampling

```
Input: $n, m, N \in \mathbb{N}, \varepsilon \in (0,1)$
Sample $n$ points $X_1^*, \ldots, X_n^*$ independently from $\mathcal{D}$
for $j = 1$ to $m$ do
    Sample $k := N/m$ points $X_1^{(j)}, \ldots, X_k^{(j)}$ i.i.d. from $\mathcal{D}$.
    Find a maximum matching $M^{(j)}$ between $X_1^*, \ldots, X_n^*$ and $X_1^{(j)}, \ldots, X_k^{(j)}$ where $X_i^*$ and $X_{i'}^{(j)}$ are adjacent iff $|X_i^* - X_{i'}^{(j)}| \leq \varepsilon$
    If $|M^{(j)}| < n$, arbitrarily match the remaining $X_i^*$'s (ignoring distance constraints)
    for $i = 1$ to $n$ do
        Let $X_{i,j}$ denote the match of $X_i^*$ in $M^{(j)}$
        Sample a label $\tilde{Y}_{i,j} \sim \mathcal{D}_{X_{i,j}}$
return $\hat{f} = \mathrm{ERM}_{\mathcal{F},\ell}\left((X_i^*, (\tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,m}))_{i \in [n]}\right)$
```

$\mathcal{D}_{X_i^*}$), but with some chance of corruption. We show this is possible by viewing any nearby distribution $\mathcal{D}_{X'}$ as a mixture of $\mathcal{D}_{X_i^*}$ with an arbitrary corruption. We can then analyze ERM on a set of metasamples, most of which are ideal, but some of which are corrupted.

Theorem 1 (Excess risk with corrupted samples)**.** *Assume that* (A1) *holds. Let $N = \tilde{\Omega}(md^{(d+2)/2}n^{(d+1)/2})$, and let $\hat{f}$ be the hypothesis returned by Algorithm 2 on $(n, m, N, 1/\sqrt{n})$. Then, for $n \geq 2 \log \frac{8}{\delta}$, with probability at least $1 - \delta$,*

$$\mathbb{E}[\ell_{\hat{f}}(X, \mathbf{Y})] - \mathbb{E}[\ell_{f^*}(X, \mathbf{Y})] \leq 2L\mathcal{R}_n(\mathcal{F}) + 2B\big(2\sqrt{\log \tfrac{4}{\delta}} + mK\big)\tfrac{1}{\sqrt{n}} \ ,$$

where X is drawn from $\mathcal{D}$, and, conditionally on X, $\mathbf{Y} = (Y_1, \ldots, Y_m)$ is drawn from $(\mathcal{D}_X)^m$.

The full versions of this result and Theorem 2, along with proofs, can be found in § D.

Proof Sketch First, we appeal to Algorithm 2 to obtain n metasamples $\{(X_i, \tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,m})\}_{i=1}^n$ with each X_i an i.i.d. draw from $\mathcal{D}$ and each $\tilde{Y}_{i,j}$ an independent draw from some $\mathcal{D}_{X_i'}$ with $\|X_i' - X_i\|_2 \leq 1/\sqrt{n}$. This holds except for $O(\sqrt{n})$ metasamples.

Now, from Assumption (A1) we have, for each X_i, m independent samples $\tilde{Y}_{i,j}$ from distributions that are close to $\mathcal{D}_{X_i}$. The key idea is to show that each metasample i's labels can be viewed as coming from $\mathcal{D}_{X_i}^m$ with high probability and from an arbitrary distribution otherwise. This argument is first made for each $\tilde{Y}_{i,j}$: We can view it as a sample from a mixture that puts high probability on $\mathcal{D}_{X_i}$ and small probability on some other distribution. Under this view, with high probability, every $\tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,m}$ comes from the $\mathcal{D}_{X_i}$ component of its mixture. Of course, this fails to be true for some of the metasamples, which we show again number only $O(\sqrt{n})$ with high probability.

The final component is an analysis of *ERM with corrupted samples*. Consider (even in the classical setting) running ERM on a set of n samples, of which $O(\sqrt{n})$ have been corrupted arbitrarily but the rest are drawn i.i.d. from the underlying distribution. In this case, we show that standard generalization bounds continue to hold with an error loss of only $O(1/\sqrt{n})$. □

Recall that nm is the label complexity (pool-based active learning paradigm) and N is the sample complexity (supervised learning paradigm).[3] To illustrate exactly how our results translate, let us adopt the parametric setting where the Rademacher complexity term $\mathcal{R}_n(\mathcal{F})$ decays at the rate of $O(1/\sqrt{n})$ with n (meta)samples. Because the sample complexity is $N = \tilde{\Omega}(md^{(d+2)/2}n^{(d+1)/2})$, the excess risk decays as the rate $\tilde{O}(1/N^{1/(d+1)})$. Similarly, in pool-based active learning, we can write $n' = nm$ for the label complexity and get excess risk decaying at a rate $O(\sqrt{1/n'})$. (We fix m here as it is inherent to the loss function.)

4.2 Faster rates under strong convexity and the uniform distribution

Let $\mathcal{F}$ be a class of linear predictors, so that $\mathcal{F}$ can be identified with a set $\mathcal{W} \subset \mathbb{R}^d$. For $w \in \mathcal{W}$, and fixed $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}^m$, we assume that the loss has the generalized linear form $\ell \colon w \mapsto c(\langle w, \phi(x) \rangle, \mathbf{y})$, for some functions c and ϕ. The risk functional R is then

$$R \colon w \mapsto \mathbb{E}\left[c\big(\langle w, \phi(X) \rangle, \mathbf{Y}\big)\right] ,$$

[3] The algorithm is the same in both paradigms, except the timing of label draws: in supervised learning, labels are drawn with each X, but in pool-based active learning, labels are only queried when X is added to a metasample.

where $X \sim \mathcal{D}$ and, conditionally on X, $\mathbf{Y} = (Y_1, \dots, Y_m) \sim (\mathcal{D}_X)^m$.

Theorem 2. *Assume that (A1) holds. Let $\mathcal{F}$ be a class of linear functionals as above, with the loss taking the generalized linear form. Suppose that $\|\phi(x)\| \leq B$ (the same B as for the upper bound on the loss). Let $\varepsilon = (mKn)^{-1}$ and $N = \tilde{\Omega}(m(n+d(\frac{\sqrt{d}}{\epsilon})^d))$, and let $\tilde{f}$ be the hypothesis returned by Algorithm 2 on (n, m, N, ϵ). If $\mathcal{D}$ is the uniform distribution over $[0,1]^d$ and the risk functional R is σ-strongly convex, then, for any $\delta \leq 3e^{-4}$ and $n \geq 2\log\frac{8}{\delta}$, with probability at least $1-\delta$*

$$\begin{aligned}\mathbb{E}[\ell_{\tilde{f}}(X, \mathbf{Y})] &- \mathbb{E}[\ell_{f^*}(X, \mathbf{Y})] \\ &\leq \tfrac{1}{n} 3B \log\tfrac{3}{\delta} + \tfrac{1}{\sigma n} 8L^2B^2\left(32 + \log\tfrac{3}{\delta}\right)\end{aligned}$$

where X is drawn from $\mathcal{D}$, and, conditionally on X, $\mathbf{Y} = (Y_1, \dots, Y_m)$ is drawn from $(\mathcal{D}_X)^m$.

This result implies that the excess risk decays at the rate $O(1/n)$, where nm is the label complexity. Since the sample complexity is $N = \tilde{\Omega}(m(n+d(mKn\sqrt{d})^d))$, this implies that the excess risk decay rate in terms of N is $\tilde{O}(1/N^{1/d})$. The proof of this result vitally leverages Lemma 4 in § B, our specialized sampling result for the uniform distribution.

Example 1 (Strongly convex risk functional)**.** *Take the example of the variance with $\mathcal{X} \subset \mathbb{R}^d$. For fixed (x, y_1, y_2), the loss and risk functional are, respectively,*

$$\begin{aligned}\ell &\colon w \mapsto \left(\langle w, x\rangle - \tfrac{1}{2}(y_1 - y_2)^2\right)^2 \ , \\ R &\colon w \mapsto \mathop{\mathbb{E}}_{\substack{X\sim\mathcal{D} \\ Y_1,Y_2\sim\mathcal{D}_X}} \left[\left(\langle w, X\rangle - \tfrac{1}{2}(Y_1 - Y_2)^2\right)^2\right] \ .\end{aligned}$$

Then $\nabla_w^2 R(w) = 2\,\mathbb{E}[XX^T]$, and so if $\mathbb{E}[XX^T] \succeq \sigma I$, then (2σ)-strong convexity holds. In the special case of $d = 1$, provided that X is non-trivial we clearly have strong convexity of the risk.

4.3 Sample complexity and dimension

While our sample complexity results above apply for any dimension d of $\mathcal{X}$, they scale exponentially in d. We are motivated by $d = O(1)$ in this paper, where the multi-observation approach can yield significant practical improvements, as we show in the next section. Nevertheless, we briefly note that an exponential dependence on d is information-theoretically necessary for *any algorithm* to model such higher-order statistics, even for the simple problem of estimating the average $\mathrm{Var}(y \mid x)$ over the distribution. We thus leave investigation of algorithms targeting higher dimension (for example, active learning approaches) to future work. For example, in § E we show:

Theorem 3. *If $\mathcal{X}$ is in the d-dimensional hypercube and the Lipschitz constant is $K = 1$, no algorithm for regression on variance of y can have nontrivial accuracy with $o(2^{d/2})$ samples.*

Intuitively, the obstacle is that a subexponential number of samples can (and, under e.g. a uniform distribution, does) have all x values separated from each other by a constant distance. Thus, any given region will only have one (x, y) pair sampled, and, information-theoretically, no knowledge can be gleaned about the variance in that region. § E also provides a similar lower bound for the case of the uniform distribution over $[0,1]^d$, where $K = d$.

5 Comparison to single-observation losses

The traditional ERM approach in our setting utilizes single-observation losses to fit $\bar{d}$ surrogate properties and then computes the higher-order property of interest from these. The theory of elicitation complexity [19, 20] gives a lower bound on the dimensionality $\bar{d}$ of any statistically consistent such procedure; e.g., for variance, $\bar{d} = 2$ which can be achieved by fitting models to the first and second moments, while for 2-norm, $\bar{d}$ is the support size of the distribution minus one. Intuitively, this approach can be problematic for sample complexity in two ways: one must model $\bar{d}$ relationships instead of just one, and these relationships may be much more complex and require many more samples than the original relationship of x with the high-order property.

We now compare multi-observation regression to the typical single-observation approach, theoretically and empirically, in the simple setting of fitting a parametric model to the variance $\mathrm{Var}[Y|X]$. The experiments also compare our algorithms from § 3 and explore other statistics of interest.

5.1 Lower bounds for single-observation losses

One method for regressing the variance is a "two-estimator" approach: use separate single-observation estimators to regress $\mathbb{E}[Y|X]^2$ and $\mathbb{E}[Y^2|X]$ respectively; the variance can then be predicted in terms of the two learned hypotheses. Intuitively, although the conditional variance might have a simple parametric form, this approach will fail if either of the two conditional moments do not. If we estimate the conditional moments from small classes, we suffer large approximation error; conversely, if we conservatively estimate them from rich classes to match our Lipschitz assumption (A1), we may overfit. Indeed, even when $Y|X$ is Bernoulli, so that $\mathbb{E}[Y^2|X] = \mathbb{E}[Y|X]$, a minimax lower bound of Stone [21] implies the following negative result for a two-estimator approach.

Corollary 1. *There is a family of problems where $\mathcal{X} = [0,1]^d$ and $x \mapsto \mathbb{E}[Y|X = x]$ is $\frac{1}{2}$-Lipschitz (as per (A1)) with range $[0,1]$ for which any two-estimator approach $\hat{f}$ has risk (under squared loss) for estimating the variance $\mathbb{E}\big[(\hat{f}(X) - \mathrm{Var}[Y|X])^2\big]$ decaying at a rate no faster than $n^{-2/(2+d)}$.*

Since the above bound is a minimax lower bound, it holds for all estimators, not just those that predict according to K-Lipschitz functions. Moreover, the same lower bound holds in a similar setting even for active learning strategies [22, Theorem 1].

To compare to our results, suppose that the variance is captured by a generalized linear form which lies within our model $\mathcal{F}$, so that $f^*\colon x \mapsto \mathrm{Var}[Y|X = x]$. Then it is easy to show that the excess risk $\mathbb{E}[\ell_{\tilde{f}}(X, \mathbf{Y})] - \mathbb{E}[\ell_{f^*}(X, \mathbf{Y})]$ in Theorem 2 takes the form $\mathbb{E}\big[(\tilde{f}(X) - \mathrm{Var}[Y|X])^2\big]$, where $\tilde{f}$ is our direct, metasample-based estimator. When $d = 1$, the rate of Theorem 2 is n^{-1} and hence better than the rate of $n^{-2/3}$ of the two-estimator approach. Note that there is no contradiction with the minimax lower bound of [21] because we assume that the variance takes a parametric form. Were the variance to be an arbitrary K-Lipschitz function, our rates would degrade; we believe in this case that our rate for $d = 1$ would also be $n^{-2/3}$, based on existing fast rates results for classes whose metric entropy grows as ε^{-1} (the class of K-Lipschitz functions exhibits such growth).

In our experiments, we explore the performance of the two-estimator approach using ERM, which can be significantly worse than a direct multi-observation regression of the variance.

5.2 Experiments

In our experiments we opted for synthetic data over real data. Multi-observation loss functions help in learning higher order statistics about $Y|X = x$; unfortunately, with real data one generally does not know what the true value of those statistics are and thus has no objective way of comparing different algorithms. By using synthetic data we can choose the underlying values of these statistics and evaluate algorithms by how closely they approximate them.

We consider three statistics: the variance, 2-norm, and upper confidence bound (UCB). As observed in prior work [6], the variance can be elicited by the two-observation loss function $\ell(r, y_1, y_2) = \big(r - \frac{1}{2}(y_1 - y_2)^2\big)^2$, and the 2-norm by the two-observation loss $\ell(r, y_1, y_2) = \big(r^2 - \mathbb{1}\{y_1 = y_2\}\big)^2$. The UCB, often used in surrogate optimization for robust engineering design, is defined by $\mathrm{ucb}_\lambda(Y) = \mathbb{E}[Y] + \lambda\sqrt{\mathrm{Var}[Y]}$ for a fixed λ. We show in § G that it can be elicited by a two-observation loss function under some restrictions on the distribution of Y.

In all of our experiments we are trying to learn a statistic of $Y|X = x$, where $X \sim \mathrm{Unif}([0,1])$. For the variance, we tried different distributions $Y|X = x$ of the form $f(x) + N(0,1)$. We present here our results when $f(x)$ is either a sine wave or a line. $\mathrm{Var}(Y|X = x) = 1$ in all cases. For the 2-norm, we constructed a distribution $\alpha_{|\mathcal{Y}|}(x) = Y|X = x$ designed to capture our motivating biodiversity example. The 2-norm of our distribution is very simple — in this case it is constant — yet which species achieve that biodiversity varies with X. In particular, we construct $\alpha_{|\mathcal{Y}|}(x)$ so that the support size of the distribution is always at most 3, but which outcomes are in that support varies with x. For the $\mathrm{ucb}_\lambda(Y)$ experiments, we chose $\lambda = 8$ and $Y|X = x \sim \Gamma(k(x), \theta(x))$ where $k(x)$ and $\theta(x)$ were chosen such that $\mathbb{E}[Y|X = x] = 2 + \sin(4\pi x)$ and $\mathrm{ucb}_\lambda(Y|X = x) = x + 10$.

As a baseline, we compare our algorithms in each case to ERM with single observation loss functions, wherein we learn models for surrogate statistics and combine them to create a model for the desired statistic. In the case of variance and ucb_λ, we fit to $\mathbb{E}[Y|X = x]$ and $\mathbb{E}[Y^2|X = x]$ and then combine those models to estimate $\mathrm{Var}[Y|X = x]$ or $\mathrm{ucb}_\lambda(Y|X = x)$. For the 2-norm, for each $y \in \mathcal{Y}$ we fit a model to $\Pr(Y = y \mid X = x)$, and then combine those models to estimate $||Y \mid X = x||_2$. For ucb_λ we also compared our algorithms to the "Monte Carlo" approach which has the power to draw multiple i.i.d. labels for each random x draw, compute the empirical statistic for that x, and then fit a line to the results. In all cases, unless otherwise specified, the hypothesis class being used is the class of linear functions. See § H for more details.

Our results are depicted in Figure 1. Observe that in all of our experiments, the ε-Nearby algorithm performed the best, closely followed by Sliding Window. Fitting lines to the moments or full distribution never performed well, which is not surprising as in all cases at least one moment was non-linear. However, even when fitting quadratics to moments which are quadratics (in the case of the variance when $f(x) = 2x - 1$), our two observation algorithms still outperformed the two moment approach. This demonstrates that for the two observation approach to be beneficial it is only necessary that the statistic is simpler than the underlying distribution, not that the underlying distribution is from an entirely unknown class. Our algorithms only show very slight improvement over the Monte Carlo approach for ucb_λ, but even being competitive is valuable since the Monte Carlo approach can sample multiple y values for a given x, while our algorithms cannot.

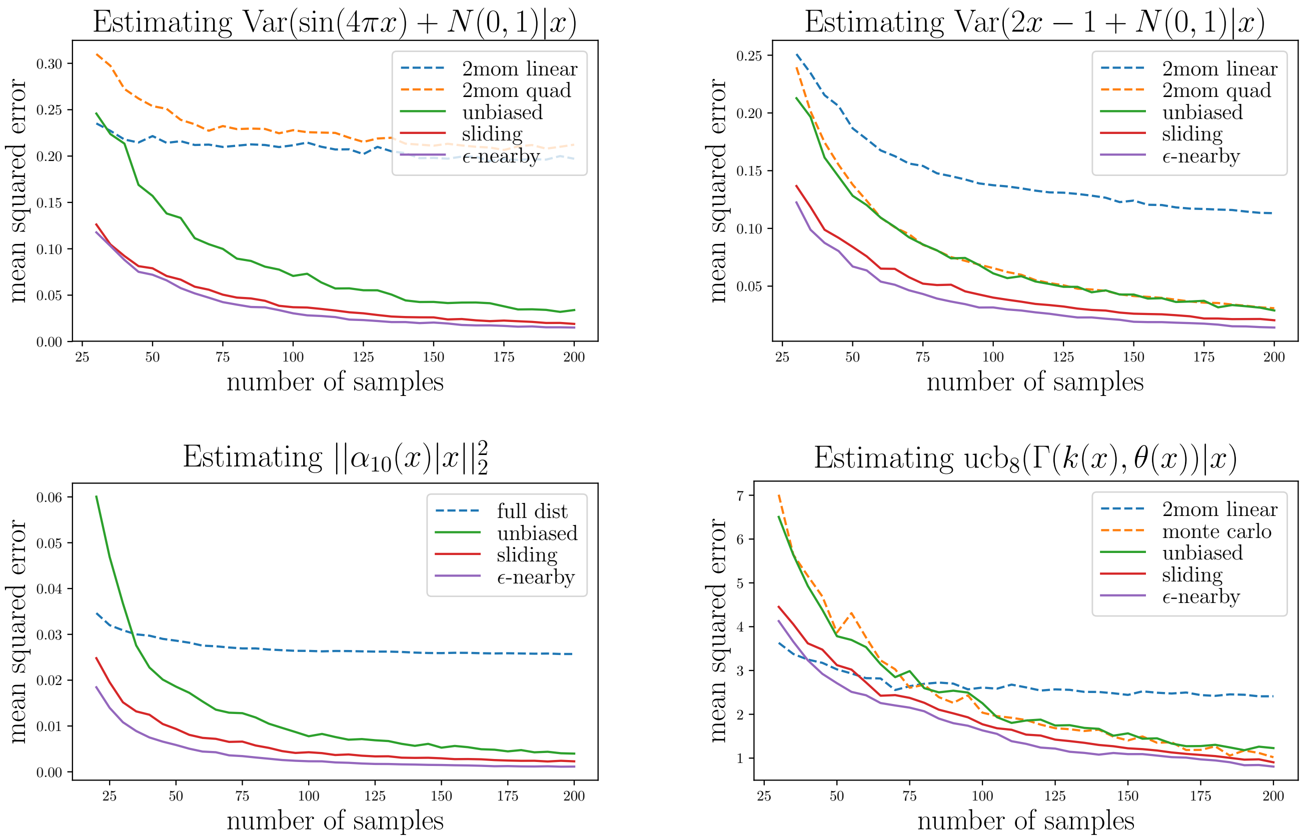

Figure 1: A comparison of ERM strategies. "2mom linear" and "2mom quad" fit lines and quadratics respectively to the first two moments of $Y|X=x$. "full dist" fits lines to the probability of each possible value of Y. "unbiased" is Algorithm 2. "sliding" and "ε-nearby" are the biased algorithms described in § 3.2. "monte carlo" fits a line to empirical estimates of the given statistic. Methods are evaluated by mean squared error to the true value of the statistic.

6 Conclusion and future work

ERM with multi-observation loss functions presents challenges and complications as compared to traditional ERM, but also interesting opportunities. With initial theoretical guarantees in hand, one next step is to explore "risk" or "variance" regression problems encountered in practice for which the multi-observation approach may be useful. Active learning settings may be among the most fruitful, as they are well-suited to collecting multiple labels at or near the same feature. This investigation will interact with *elicitation* and the design of loss functions that are consistent for a desired property of the conditional distribution, such as the variance. In particular, an open problem is to discover a loss function for the UCB property that does not require restrictions on the distribution of Y.

Another direction lies in the dimension d of $\mathcal{X}$. Our lower bounds show that, without additional assumptions, the sample complexity of estimating e.g. the variance is exponential in d. To make headway, one could assume a low intrinsic dimension, embedded in some higher-dimensional space, or appeal to active learning, which could sidestep these lower bounds.

With respect to adapting to the intrinsic dimension, a natural area for insight is the analysis of the k-nearest neighbor method. To our knowledge, all works that develop finite-sample high probability guarantees for the k-NN method do so by invoking assumptions of a stronger nature than those imposed in the present paper. At a high level, these assumptions amount to the density of $\mathcal{D}$ being lower bounded by a positive constant on the support of $\mathcal{D}$. Adapting these results to our setting, without stronger assumptions, may be a major undertaking. Alternatively, some classical works [15, 16] avoid stronger assumptions but only provide guarantees that hold in expectation or asymptotically, which we view as a significant practical weakness. Thus, a new analysis which adapts to the intrinsic dimension is an important, but nontrivial, direction for future work.

References

[1] R. T. Rockafellar, J. O. Royset, and S. I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234:140–154, 2014.

[2] Myoung-jae Lee. Mode regression. *Journal of Econometrics*, 42(3):337–349, November 1989.

[3] Lou Jost. Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10):2427–2439, 2007.

[4] Hanna Tuomisto. A consistent terminology for quantifying species diversity? yes, it does exist. *Oecologia*, 164(4):853–860, 2010.

[5] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[6] Sebastian Casalaina-Martin, Rafael Frongillo, Tom Morgan, and Bo Waggoner. Multi-observation elicitation. In *Proceedings of the 30th Conference on Learning Theory*, pages 1–18, 2017.

[7] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33):3190–3218, 2007.

[8] Johannes O Royset, Luca Bonfiglio, Giuliano Vernengo, and Stefano Brizzolara. Set-Based Approach to Design under Uncertainty and Applications to Shaping a Hydrofoil. *Preprint*, 2016.

[9] J.-C. Jouhaud, Pierre Sagaut, Marc Montagnac, and Julien Laurenceau. A surrogate-model based multidisciplinary shape optimization method with application to a 2d subsonic airfoil. *Computers & Fluids*, 36(3):520–529, 2007.

[10] Muhammad Shahbaz, Zhong-Hua Han, W. P. Song, and M. Nadeem Aizud. Surrogate-based robust design optimization of airfoil using inexpensive Monte Carlo method. In *Applied Sciences and Technology (IBCAST), 2016 13th International Bhurban Conference on*, pages 497–504. IEEE, 2016.

[11] Ioannis Doltsinis and Zhan Kang. Robust design of structures using optimization methods. *Computer Methods in Applied Mechanics and Engineering*, 193(23):2221–2237, 2004.

[12] Yew-Soon Ong, Prasanth B. Nair, and Kai Yew Lum. Max-min surrogate-assisted evolutionary algorithm for robust design. *IEEE Transactions on Evolutionary Computation*, 10(4):392–404, 2006.

[13] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[14] Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, January 2004.

[15] Charles J Stone et al. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

[16] Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

[17] Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

[18] Steve Hanneke. Nonparametric active learning, part 1: Smooth regression functions. http://www.stevehanneke.com/docs/2016/nonparametric-part-1.pdf, 2017.

[19] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

[20] Rafael Frongillo and Ian A. Kash. On elicitation complexity. In *Advances in Neural Information Processing Systems 29*, 2015.

[21] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[22] Rui M Castro, Rebecca Willett, and Robert Nowak. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems*, pages 179–186, 2006.

[23] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[24] Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

[25] Govinda M Kamath, Eren Şaşoğlu, and David Tse. Optimal haplotype assembly from high-throughput mate-pair reads. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 914–918. IEEE, 2015.

[26] Peter W Glynn. Upper bounds on Poisson tail probabilities. *Operations research letters*, 6(1):9–14, 1987.

[27] Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pages 1545–1552, 2009.

[28] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

[29] Charles J Stone et al. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

[30] N.S. Lambert. Elicitation and Evaluation of Statistical Forecasts. *Preprint*, 2011.

[31] Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Proceedings of the 27th Conference on Learning Theory*, pages 482–526, 2014.

A Proofs from background

Proof of Lemma 1 Let P be the probability measure operator with respect to P, and let P_n be the empirical measure operator with respect to the sample $(X_1, Y_1), \dots, (X_n, Y_n)$; that is, for each $f \in \mathcal{F}$, we have $\mathsf{P}\,\ell_f = \mathbb{E}[\ell_f(X,Y)]$ and $\mathsf{P}_n\,\ell_f = \frac{1}{n}\sum_{i=1}^n \ell_f(X_i, Y_i)$.

From the bounded differences inequality (Lemma 1.2 of [23]) with bounded differences $c_i = \frac{B}{n}$,

$$\Pr\left(\sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)\ell_f > \mathbb{E}\left[\sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)\ell_f \right] + t \right) \leq e^{-2nt^2/B^2}.$$

Next, let $(X_i', Y_i')_{i \in [n]}$ be an independent copy of $(X_i, Y_i)_{i \in [n]}$. Jensen's inequality implies that

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)\ell_f \right] &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}[\ell_f(X_i', Y_i')] - \ell_f(X_i, Y_i) \right) \right] \\ &\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i) \right) \right], \end{aligned}$$

which, for independent Rademacher random variables $\epsilon_1, \dots, \epsilon_n$ is in turn equal to

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}_{\epsilon_1, \dots, \epsilon_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \left(\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i) \right) \right] \right] &\leq 2\,\mathbb{E}\left[\mathbb{E}_{\epsilon_1, \dots, \epsilon_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell_f(X_i, Y_i) \right] \right] \\ &= 2\mathcal{R}_n(\{\ell_f \colon f \in \mathcal{F}\}). \end{aligned}$$

The result follows from Theorem 7 of [24], which shows that under the L-Lipschitzness of the loss as a function of the prediction $\hat{y} = f(x)$, we have the comparison inequality $\mathcal{R}_n(\{\ell_f \colon f \in \mathcal{F}\}) \leq L\mathcal{R}_n(\mathcal{F})$. □

B Proofs of sampling methods

B.1 Naïve Sampling Lemma

Lemma 3 (Naïve Sampling Lemma)**.** *Let $d \in \mathcal{N}$ and let $x_1^*, \dots, x_n^*, x_1, \dots, x_N$ be drawn independently from a distribution $\mathcal{D}$ on $\mathcal{X} = [0,1]^d$. If $N \geq mn^{(d+3)/2} d^{d/2} \log \frac{2mn}{\delta}$, then with probability at least $1 - \delta$, there is a set $\mathcal{J}$ of cardinality at least $n - \sqrt{(n \log \frac{2}{\delta})/2}$ for which, for each x_j^* with $j \in \mathcal{J}$, there are at least m points $x_{i_{j,1}}, \dots, x_{i_{j,m}}$ satisfying $\|x_j^* - x_{i_{j,m}}\|_2 \leq \frac{1}{\sqrt{n}}$, and all the $i_{1,1}, \dots, i_{1,m}, \dots, i_{|\mathcal{J}|,1}, \dots, i_{|\mathcal{J}|,m} \in [N]$ are distinct.*

Proof. Take $C_1, \dots, C_r$ to be a partition of $\mathcal{X}$ for which every cell C_j has diameter $\sup_{x,x' \in C_j} \|x - x'\|_2$ at most ε. Observe that we may always take $r \leq \mathcal{N}(\mathcal{X}, \varepsilon/2)$, where $\mathcal{N}(\mathcal{X}, \varepsilon)$ is the minimum number of radius-ε balls in the Euclidean norm $\|\cdot\|_2$ whose union contains $\mathcal{X}$. Note that $r \leq (\sqrt{d}/\varepsilon)^d$.

We take $\varepsilon = \frac{1}{\sqrt{n}}$ and partition the set of cells of $\mathcal{X}$ into light cells and heavy cells, where any light cell C_j satisfies $\Pr(C_j) \leq \frac{1}{r\sqrt{n}}$. Since there are at most r cells, the aggregate probability measure among all the light cells is at most $\frac{1}{\sqrt{n}}$. Hoeffding's inequality implies that only with probability at most $\delta/2$ will more than $\sqrt{(n \log \frac{2}{\delta})/2}$ samples of $x_1^*, \dots, x_n^*$ fall into light cells. The remainder of the points therefore fall into heavy cells, each of which has probability measure at least $\frac{1}{r\sqrt{n}}$. Now, for some fixed x_j^* in a heavy cell, if we sample $r\sqrt{n} \log \frac{1}{\delta}$ points, then with probability at least $1 - \delta$ at least one of these latter points would fall into the same cell as x_j^*. Thus, if we sample

$$N = mn^{3/2} r \log \frac{2mn}{\delta} \leq mn^{(d+3)/2} d^{d/2} \log \frac{2mn}{\delta}$$

points $x_1, \dots, x_N$, then with probability at least $1 - \delta/2$ every x_j^* in a heavy cell will have at least m samples falling into its cell. □

B.2 Improved Uniform Sampling Lemma

Lemma 4 (Improved Uniform Sampling Lemma)**.** *Let $\varepsilon \in (0,1)$, $d \in \mathcal{N}$ and let $x_1^*, \ldots, x_n^*, x_1, \ldots, x_N$ be drawn independently from $\mathcal{D} = \mathrm{Unif}([0,1]^d)$. If $N \geq Cm\big(n + (\frac{\sqrt{d}}{\varepsilon})^d (\log \frac{m}{\delta} + d \log \frac{d}{\varepsilon})\big)$ for a universal constant C, then with probability at least $1-\delta$, for each $j \in [n]$, there are at least m points $x_{i_{j,1}}, \ldots, x_{i_{j,m}}$ satisfying $\|x_j^* - x_{i_{j,m}}\|_2 \leq \varepsilon$, and all the $i_{1,1}, \ldots, i_{1,m}, \ldots, i_{n,1}, \ldots, i_{n,m} \in [N]$ are distinct.*

Proof. We will prove this for the case when $m = 1$, from which the general result follows by setting $\delta = \delta/m$ and repeating m times.

We partition $[0,1]^d$ into $r = (\sqrt{d}/\varepsilon)^d$ hypercubes of width $w = \varepsilon/\sqrt{d}$, $C_1, \ldots, C_r$. $S_i = |\{i | x_i^* \in C_i\}|$ is the number of x_i^*'s that lie in the i^{th} hypercube. Similarly, let $T_i = |\{i | x_i \in C_i\}|$ be the number of x_i's that fall in the i^{th} hypercube. We will now show that for all $i \in [r]$, $T_i \geq S_i$ with probability at least $1-\delta$. The lemma follows from this as we can then simply match up points within each hypercube and the maximum distance between two points in a hypercube is $w\sqrt{d} = \varepsilon$.

In order to bound $\Pr[\forall i \; T_i \geq S_i]$, we will first consider a slight alteration of our setting. Instead of having fixed numbers of samples n and N, they will be random variables $\tilde{n} \sim \mathrm{Pois}(2n)$ and $\tilde{N} \sim \mathrm{Pois}(N/2)$. In this setting, we let $\tilde{S}_i$ and $\tilde{T}_i$ be the number of samples of each category falling into the i^{th} interval. The key property we will use here is that

$$\tilde{S}_i \sim \mathrm{Bin}(\tilde{n}, w^d) = \mathrm{Pois}(2nw^d),$$

and analogously $\tilde{T}_i \sim \mathrm{Pois}(Nw^d/2)$.

Using this we have

$$\begin{aligned}\Pr[\tilde{T}_i < \tilde{S}_i] &= \Pr[\mathrm{Pois}(Nw^d/2) < \mathrm{Pois}(2nw^d)] \\ &\leq e^{-\left(\sqrt{Nw^d/2} - \sqrt{2nw^d}\right)^2}\end{aligned}$$

where the final inequality is a standard bound on Poisson races which follows from a Chernoff bound [25, Appendix A]. By a union bound, the probability that $\tilde{T}_i \geq \tilde{S}_i$ for all $i \in [r]$ is lower bounded by

$$1 - r \cdot e^{-\left(\sqrt{Nw^d/2} - \sqrt{2nw^d}\right)^2}.$$

Now observe that

$$\begin{aligned}\Pr[\forall i \; T_i \geq S_i] &= \Pr[\forall i \; \tilde{T}_i \geq \tilde{S}_i \mid \tilde{N} = N \wedge \tilde{n} = n] \\ &\geq \Pr[\forall i \; \tilde{T}_i \geq \tilde{S}_i \mid \tilde{N} \leq N \wedge \tilde{n} \geq n]\end{aligned}$$

because increasing N or decreasing n only increases our chances of finding a matching. Let I be the event that $\tilde{N} \leq N \wedge \tilde{n} \geq n$, and let $\bar{I}$ be its negation.

$$\begin{aligned}\Pr[\forall i \; \tilde{T}_i \geq \tilde{S}_i \mid I] &= \frac{\Pr[\forall i \; \tilde{T}_i \geq \tilde{S}_i] - \Pr[\forall i \; \tilde{T}_i \geq \tilde{S}_i \mid \bar{I}] \Pr[\bar{I}]}{\Pr[I]} \\ &\geq 1 - r \cdot e^{-\left(\sqrt{Nw^d/2} - \sqrt{2nw^d}\right)^2} - \Pr[\bar{I}] \\ &\geq 1 - r \cdot e^{-\left(\sqrt{Nw^d/2} - \sqrt{2nw^d}\right)^2} - e^{-\Theta(n+N)}.\end{aligned}$$

The final inequality follows from standard Poisson tail bounds [26, Proposition 1]. Plugging in $N = C\left(n + \left(\frac{\sqrt{d}}{\varepsilon}\right)^d \left(\log \frac{1}{\delta} + d \log \frac{d}{\varepsilon}\right)\right)$, for a sufficiently large constant C, gives us the lemma. □

B.3 Proof of Lemma 2 (Improved Nonuniform Sampling Lemma)

The next result is the full version of Lemma 2.

Algorithm 3: Sliding Window

```
Input: $n, m \in \mathbb{N}$
Sample $n$ pairs $(X_1, Y_1), \dots, (X_n, Y_n)$
Sort the pairs so that $X_1 \le X_2 \dots \le X_n$
for $i = 1$ to $n - m + 1$ do
    Set $\tilde{X}_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} X_{i+j-1}$
    for $j = 1$ to $m$ do
        Set $\tilde{Y}_{i,j} \leftarrow Y_{i+j-1}$
$\hat{f} = \mathrm{ERM}_{\mathcal{F},\ell}((\tilde{X}_i, (\tilde{Y}_{i,1}, \dots, \tilde{Y}_{i,m}))_{i=1}^{n-m+1})$
return $\hat{f}$
```

Algorithm 4: ε-Nearby

```
Input: $n, m \in \mathbb{N}, \varepsilon \in (0,1)$
Sample $n$ pairs $(X_1, Y_1), \dots, (X_n, Y_n)$
Sort the pairs so that $X_1 \le X_2 \dots \le X_n$
Set $t \leftarrow 1$
for $i_1 = 1$ to $n - m + 1$ do
    Set $k \leftarrow \max\{j \mid i_1 < j \le n \wedge X_j - X_{i_1} \le \varepsilon\}$
    for $\{i_2, \dots, i_m\} \subseteq \{i_1 + 1, \dots, k\}$ do
        Set $\tilde{X}_t \leftarrow \frac{1}{m} \sum_{j=1}^{m} X_{i_j}$
        Set $\tilde{Y}_{t,j} \leftarrow Y_{i_j}$
        for $j = 1$ to $m$ do  Set $t \leftarrow t + 1$
return $\hat{f} = \mathrm{ERM}_{\mathcal{F},\ell}((\tilde{X}_i, (\tilde{Y}_{i,1}, \dots, \tilde{Y}_{i,m}))_{i \in [t]})$
```

Lemma 5 (Improved Nonuniform Sampling Lemma)**.** *Let $d \in \mathcal{N}$ and let $x_1^*, \dots, x_n^*, x_1, \dots, x_N$ be drawn independently from a distribution $\mathcal{D}$ on $\mathcal{X} = [0,1]^d$. If $N \ge Cmn^{(d+1)/2}d^{d/2}\left(\log\frac{m}{\delta} + d\log(nd)\right)$ for a sufficiently large constant C, then with probability at least $1-\delta$, there is a set $\mathcal{J} \subset [n]$ of cardinality at least $n - \sqrt{(n\log\frac{2}{\delta})/2}$ for which, for each x_j^* with $j \in \mathcal{J}$, there are at least m points $x_{i_{j,1}}, \dots, x_{i_{j,m}}$ satisfying $\|x_j^* - x_{i_{j,m}}\|_2 \le \frac{1}{\sqrt{n}}$, and all the $i_{1,1}, \dots, i_{1,m}, \dots, i_{|\mathcal{J}|,1}, \dots, i_{|\mathcal{J}|,m} \in [N]$ are distinct.*

Proof. This follows from a combination of the proofs of Lemma 3 and Lemma 4. As in Lemma 3, we partition $[0,1]^d$ into heavy and light cells of diameter $1/\sqrt{n}$ and see that with probability $\delta/2$ at most $\sqrt{(n\log\frac{2}{\delta})/2}$ x_i^*'s fall in light cells. We then apply the argument of Lemma 4 to these heavy cells, using $\frac{1}{r\sqrt{n}}$ (for $r = (nd)^{d/2}$) as their probability mass instead of w^d. Thus, it suffices that

$$N \ge Cmn^{(d+1)/2}d^{d/2}\left(\log\frac{m}{\delta} + d\log(nd)\right) \ge C'm\left(n + r\sqrt{n}\log\frac{mr}{\delta}\right)$$

for a sufficiently large constant C. □

C Details of biased algorithms

Algorithms 3 and 4 provide the details of our biased algorithms for single dimensional $\mathcal{X}$.

D Proofs of excess risk bounds

D.1 Proof of Theorem 1

The next result is the full version of Theorem 1.

Theorem 4 (Excess risk with corrupted samples)**.** *Assume that (A1) holds. Let $N = Cmn^{(d+1)/2}d^{d/2}\log\frac{m(nd)^d}{\delta}$ for some universal constant C, and let $\tilde{f}$ be the hypothesis returned by Algorithm 2 on $(n, m, N, 1/\sqrt{n})$. Then, for $n \ge 2\log\frac{8}{\delta}$, with probability at least $1-\delta$,*

$$\mathbb{E}[\ell_{\tilde{f}}(X, \mathbf{Y})] - \mathbb{E}[\ell_{f^*}(X, \mathbf{Y})] \le 2L\mathcal{R}_n(\mathcal{F}) + 2B\big(2\sqrt{\log\tfrac{4}{\delta}} + mK\big)\frac{1}{\sqrt{n}} \ ,$$

where X is drawn from $\mathcal{D}$, and, conditionally on X, $\mathbf{Y} = (Y_1, \dots, Y_m)$ is drawn from $(\mathcal{D}_X)^m$.

Proof of Theorem 4 For each $i \in [n]$ and $j \in [m]$, draw $Y_{i,j}$ independently according to distribution $\mathcal{D}_{X_i^*}$. This "clean" sample is simply a theoretical device for the analysis.

We first set up some convenient notation. For each $i \in [n]$, define $\mathbf{Y}_i := (Y_{i,1}, \dots, Y_{i,m})$ and $\tilde{\mathbf{Y}}_i := (\tilde{Y}_{i,1}, \dots, \tilde{Y}_{i,m})$. Let P be a probability measure operator, defined according to $\mathsf{P}\ \ell_f = \mathbb{E}[\ell_f(X, \mathbf{Y})]$; here, X is drawn from $\mathcal{D}$, and, conditionally on X, $\mathbf{Y} = (Y_1, \ldots, Y_m)$ is drawn from $(\mathcal{D}_X)^m$. For a fixed f, P takes ℓ_f to its expected value on a new draw from the distribution X and an m-tuple $\mathbf{Y}$ from $\mathcal{D}_X$. We also define the empirical probability measure operators P_n and $\tilde{\mathsf{P}}_n$ via

$$\mathsf{P}_n\, \ell_f = \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_j^*, \mathbf{Y}_i) \qquad \text{and} \qquad \tilde{\mathsf{P}}_n\, \ell_f = \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_i^*, \tilde{\mathbf{Y}}_i).$$

Now according to Lemma 6 below, we have for any positive t_1, t_2, t_3,

$$\begin{aligned} &\Pr\Big(\mathsf{P}\,\ell_{\tilde{f}} > \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2 + t_3\Big) \\ &\le \Pr\left(\sup_{f\in\mathcal{F}}(\mathsf{P} - \mathsf{P}_n)\ell_f > t_1\right) + \Pr\left(\sup_{f\in\mathcal{F}}|(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| > t_2\right) + \Pr\big((\mathsf{P}_n - \mathsf{P})\ell_{f^*} > t_3\big). \end{aligned}$$

From Lemma 1, the first probability is at most $\delta/4$ when $t_1 = 2L\mathcal{R}_n(\mathcal{F}) + B\sqrt{\frac{\log(4/\delta)}{2n}}$. From Hoeffding's inequality, the third probability is at most $\delta/4$ when $t_3 = B\sqrt{\frac{\log(4/\delta)}{2n}}$ (note that f^* is fixed). The remainder of the proof controls the second probability. As we will see, we will be able to take $t_2 = O\left(B\sqrt{\frac{\log(1/\delta)}{n}}\right)$ when the probability is at most $\delta/2$.

First, under Lemma 5, with probability at least $1 - \delta/4$, there is a subset $\mathcal{I}_G \subset [n]$ of cardinality at least $n_G := n - \sqrt{(n\log\frac{8}{\delta})/2}$ for which, for each X_i^* with $i \in \mathcal{I}_G$, there are at least m points $X_{k_{i,1}}, \ldots, X_{k_{i,m}}$ within distance ε of x_i^*, and all the $k_{1,1}, \ldots, k_{1,m}, \ldots, k_{n,1}, \ldots, k_{n,m} \in [N]$ are distinct.

Next, we make the observation that the *observed* sample can be obtained by the following corruption modifications to $(\mathbf{Y}_i)_{i\in[n]}$.

1. For $i \in [n] \setminus \mathcal{I}_G$, draw $\tilde{Y}_{i,j}$ from distribution $\mathcal{D}_{X_{i,j}}$.

2. For $i \in \mathcal{I}_G$, observe that Assumption (A1) implies that, without loss of generality, we can view each $Y_{i,j}$ as drawn in the following way. First, set $\tilde{Y}_{i,j}$ to $Y_{i,j}$. Next, draw a Bernoulli random variable $Z_{i,j}$ with success probability $\tau := K\varepsilon$, and if $Z_{i,j} = 1$, we corrupt $\tilde{Y}_{i,j}$ by setting it (again) to a new draw from some distribution $Q_{i,j}$ that can depend on both X_i^* and $X_{i,j}$.

For each i, if $Z_{i,j} = 0$, we say that (i, j) are GOOD, and if $(i, 1), \ldots, (i, m)$ all are GOOD, we say that i is GOOD. If some i is not GOOD, then it is BAD. Clearly, for each i separately, with probability at least $1 - mK/\sqrt{n}$ over $(Z_{i,j})_{j\in[m]}$ it holds that i is GOOD (recall that $m\tau = mK\varepsilon = mK/\sqrt{n}$). Thus, from Hoeffding's inequality the probability (over $(Z_{i,j})_{i\in[n],j\in[m]}$) that at least $(C+1)\frac{mKn_G}{\sqrt{n}}$ of the i's are BAD is at most $e^{-2(n_G/n)(mKC)^2} \le e^{-C^2}$ (recall that $n \ge 2\log\frac{8}{\delta}$) , so if $C = \sqrt{\log\frac{4}{\delta}}$ then this probability is at most $\delta/4$ (and our total probability of failure thus far is $\delta/2$). We denote the (further diminished) good set of indices by $\mathcal{I}'_G := \{i \in \mathcal{I}_G \colon i \text{ is GOOD}\}$; this set has cardinality at least $n'_G := n_G\left(1 - \frac{\sqrt{\log\frac{4}{\delta}} + mK}{\sqrt{n}}\right)$ with probability at least $1 - \delta/2$.

From the above argument, we see that with probability at least $1 - \delta/2$, at most $n'_B := n - n'_G$ corruption modifications occurred, and hence

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_f(X_i^*, \tilde{\mathbf{Y}}_i) - \frac{1}{n}\sum_{i=1}^{n}\ell_f(X_i^*, \mathbf{Y}_i)\right| \le \frac{B(n - n'_G)}{n}.$$

Observe that

$$\begin{aligned} n - n'_G &= n - \left(n - \sqrt{(n \log \frac{8}{\delta})/2}\right)\left(1 - \frac{\sqrt{\log \frac{4}{\delta} + mK}}{\sqrt{n}}\right) \\ &= \sqrt{n}\left(\sqrt{\log \frac{4}{\delta} + mK} + \sqrt{\frac{\log \frac{8}{\delta}}{2}}\right) - \sqrt{\frac{\log \frac{8}{\delta}}{2}}\left(\sqrt{\log \frac{4}{\delta} + mK}\right) \\ &\leq \sqrt{n}\left(2\sqrt{\log \frac{4}{\delta} + mK}\right), \end{aligned}$$

and thus we may take $t_2 = \frac{B\left(2\sqrt{\log \frac{4}{\delta} + mK}\right)}{\sqrt{n}}$. □

Lemma 6. *Under the hypotheses of Theorem 4 we have the following, with probability taken over the random sample (i.e. $\tilde{f}$, P_n, and $\tilde{\mathsf{P}}_n$ are functions of the random sample):*

$$\begin{aligned} &\Pr\left(\mathsf{P}\,\ell_{\tilde{f}} > \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2 + t_3\right) \\ &\leq \Pr\left(\sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)\ell_f > t_1\right) + \Pr\left(\sup_{f \in \mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| > t_2\right) + \Pr\big((\mathsf{P}_n - \mathsf{P})\ell_{f^*} > t_3\big). \end{aligned}$$

Proof. First, observe that (using $[\![E]\!]$ for the 0-1 indicator function of a random event)

$$\begin{aligned} &\left[\!\!\left[\left(\sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)\ell_f \leq t_1\right) \wedge \left(\sup_{f \in \mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| \leq t_2\right) \wedge \big((\mathsf{P}_n - \mathsf{P})\ell_{f^*} \leq t_3\big)\right]\!\!\right] \\ &\leq \left[\!\!\left[\mathsf{P}\,\ell_{\tilde{f}} \leq \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2 + t_3\right]\!\!\right]. \end{aligned}$$

To see this,

$$\begin{aligned} \mathsf{P}\,\ell_{\tilde{f}} \leq \mathsf{P}_n\,\ell_{\tilde{f}} + t_1 &\leq \mathsf{P}_n\,\ell_{\hat{f}} + \mathsf{P}_n\,(\ell_{\tilde{f}} - \ell_{\hat{f}}) + t_1 \\ &\overset{\text{(a)}}{\leq} \mathsf{P}_n\,\ell_{\hat{f}} + 2t_2 + t_1 \\ &\overset{\text{(b)}}{\leq} \mathsf{P}_n\,\ell_{f^*} + 2t_2 + t_1 \\ &\leq \mathsf{P}\,\ell_{f^*} + t_3 + 2t_2 + t_1, \end{aligned}$$

where (a) is from Lemma 7 and (b) is from the optimality of ERM under P_n.

By subtracting each side from one and rearranging, we get an implication on the negation of these events

$$\begin{aligned} &\left[\!\!\left[\left(\sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)\ell_f > t_1\right) \vee \left(\sup_{f \in \mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| > t_2\right) \vee \big((\mathsf{P}_n - \mathsf{P})\ell_{f^*} > t_3\big)\right]\!\!\right] \\ &\geq \left[\!\!\left[\mathsf{P}\,\ell_{\tilde{f}} > \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2 + t_3\right]\!\!\right] \end{aligned}$$

and we can use the union bound. □

Lemma 7. *The following statement is true:*

$$\mathsf{P}_n\,\ell_{\tilde{f}} - \mathsf{P}_n\,\ell_{\hat{f}} \leq 2 \sup_{f \in \mathcal{F}} \left|(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f\right|.$$

Proof. Observe that

$$\begin{aligned} \mathsf{P}_n\,\ell_{\tilde{f}} - \mathsf{P}_n\,\ell_{\hat{f}} &= \left(\mathsf{P}_n\,\ell_{\tilde{f}} - \tilde{\mathsf{P}}_n\,\ell_{\tilde{f}}\right) + \left(\tilde{\mathsf{P}}_n\,\ell_{\tilde{f}} - \mathsf{P}_n\,\ell_{\hat{f}}\right) \\ &\leq \left(\mathsf{P}_n\,\ell_{\tilde{f}} - \tilde{\mathsf{P}}_n\,\ell_{\tilde{f}}\right) + \left(\tilde{\mathsf{P}}_n\,\ell_{\hat{f}} - \mathsf{P}_n\,\ell_{\hat{f}}\right) \\ &\leq 2 \sup_{f \in \mathcal{F}} \left|(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f\right|, \end{aligned}$$

where the first inequality is from the optimality of $\tilde{f}$ under $\tilde{\mathsf{P}}_n$. □

D.2 Proof of Theorem 2

The next result is the full version of Theorem 2.

Theorem 5. *Assume that (A1) holds. Let $\mathcal{F}$ be a class of linear functionals as above, with the loss taking the generalized linear form. Suppose that $\|\phi(x)\| \leq B$ (the same B as for the upper bound on the loss). Let $\varepsilon = (mKn)^{-1}$ and $N = Cm\big(n + (\frac{\sqrt{d}}{\varepsilon})^d \log \frac{3md^d}{\delta\varepsilon^d}\big)$ for a universal constant C, and let $\tilde{f}$ be the hypothesis returned by Algorithm 2 on (n, m, N, ε). If $\mathcal{D}$ is the uniform distribution over $[0,1]^d$ and the risk functional R is σ-strongly convex, then, for any $\delta \leq 3\mathrm{e}^{-4}$ and $n \geq 2\log\frac{8}{\delta}$, with probability at least $1-\delta$*

$$\mathbb{E}[\ell_{\tilde{f}}(X, \mathbf{Y})] - \mathbb{E}[\ell_{f^*}(X, \mathbf{Y})] \leq \tfrac{1}{n} 3B \log \tfrac{3}{\delta} + \tfrac{1}{\sigma n} 8L^2B^2 \left(32 + \log \tfrac{3}{\delta}\right)$$

where X is drawn from $\mathcal{D}$, and, conditionally on X, $\mathbf{Y} = (Y_1, \ldots, Y_m)$ is drawn from $(\mathcal{D}_X)^m$.

Proof of Theorem 5 For each $i \in [n]$ and $j \in [m]$, draw $Y_{i,j}$ independently according to distribution $\mathcal{D}_{X_i^*}$. This “clean” sample is simply a theoretical device for the analysis.

We first set up some convenient notation. For each $i \in [n]$, define $\mathbf{Y}_i := (Y_{i,1}, \ldots, Y_{i,m})$ and $\tilde{\mathbf{Y}}_i := (\tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,m})$. Let P be a probability measure operator, defined according to $\mathsf{P}\,\ell_f = \mathbb{E}[\ell_f(X, \mathbf{Y})]$; here, X is drawn from $\mathcal{D}$, and, conditionally on X, $\mathbf{Y} = (Y_1, \ldots, Y_m)$ is drawn from $(\mathcal{D}_X)^m$. For a fixed f, P takes ℓ_f to its expected value on a new draw from the distribution X and an m-tuple $\mathbf{Y}$ from $\mathcal{D}_X$. We also define the empirical probability measure operators P_n and $\tilde{\mathsf{P}}_n$ via

$$\mathsf{P}_n\,\ell_f = \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_j^*, \mathbf{Y}_i) \qquad \text{and} \qquad \tilde{\mathsf{P}}_n\,\ell_f = \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_i^*, \tilde{\mathbf{Y}}_i).$$

Now according to Lemma 8 below, we have for any positive t_1, t_2,

$$\begin{aligned}
&\Pr\Big(\mathsf{P}\,\ell_{\tilde{f}} > \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2\Big) \\
&\leq \Pr\left(\sup_{f\in\mathcal{F}}\Big\{\mathsf{P}\,(\ell_f - \ell_{f^*}) - \mathsf{P}_n\,(\ell_f - \ell_{\hat{f}})\Big\} > t_1\right) + \Pr\left(\sup_{f\in\mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| > t_2\right).
\end{aligned}$$

Now, since the risk is σ-strongly convex, the first probability is at most $\delta/3$ from Theorem 1 of [27] with $a = 1$ and $\lambda = 2\sigma$, yielding the choice $t_1 = 8L^2B^2\,(32 + \log(3/\delta))\,/(\sigma n)$.

The remainder of the proof controls the second probability. As we will see, we will be able to take $t_2 = O\left(\frac{B\log(1/\delta)}{n}\right)$ when the probability is at most $2\delta/3$.

First recall that by using Algorithm 2 with our choice of parameters, Lemma 4 gives us that with probability at least $1 - \delta/3$, for all $i \in [n]$, there are at least m points $X_{k_{i,1}}, \ldots, X_{k_{i,m}}$ within distance ε of X_i^*, and all the $k_{1,1}, \ldots, k_{1,m}, \ldots, k_{n,1}, \ldots, k_{n,m} \in [N]$ are distinct.

Next, we make the observation that the *observed* sample can be obtained by the following “corruption” modifications to $(\mathbf{Y}_i)_{i\in[n]}$.

1. For $i \in [n] \setminus \mathcal{I}_G$, draw $\tilde{Y}_{i,j}$ from distribution $\mathcal{D}_{X_{i,j}}$.
2. For $i \in \mathcal{I}_G$, observe that Assumption (A1) implies that, without loss of generality, we can view each $Y_{i,j}$ as drawn in the following way. First, set $\tilde{Y}_{i,j}$ to $Y_{i,j}$. Next, draw a Bernoulli random variable $Z_{i,j}$ with success probability $\tau := mK\varepsilon$, and if $Z_{i,j} = 1$, we “corrupt” $\tilde{Y}_{i,j}$ by setting it (again) to a new draw from some distribution $Q_{i,j}$ that can depend on both X_i^* and $X_{i,j}$.

For each i, if $Z_{i,j} = 1$, we say that (i, j) are GOOD, and if $(i, 1), \ldots, (i, m)$ all are GOOD, we say that i is GOOD. If some i is not GOOD, then it is BAD. Clearly, for each i separately, with probability at least $1 - 1/n$ over

$(Z_{i,j})_{j\in[m]}$ it holds that i is GOOD (recall that $m\tau = mK\varepsilon = 1/n$). Thus, from a multiplicative Chernoff bound[4] the probability (over $(Z_{i,j})_{i\in[n],j\in[m]}$) that at least $\frac{3}{2}\log\frac{3}{\delta}$ of the i's are BAD is at most $\delta/3$ (for $\delta < 3e^{-4}$) (and our total probability of failure thus far is $2\delta/3$). We denote the good set of indices by $\mathcal{I}'_G := \{i \in [n] \colon i \text{ is GOOD}\}$; this set has cardinality at least $n'_G = n - \frac{3}{2}\log\frac{3}{\delta}$ with probability at least $1 - 2\delta/3$.

From the above argument, we see that with probability $1 - 2\delta/3$, at most $n'_B := n - n'_G$ corruption modifications occur, and hence

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_i^*, \tilde{\mathbf{Y}}_i) - \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_i^*, \mathbf{Y}_i) \right| \leq \frac{B(n - n'_G)}{n}.$$

Thus, we may take $t_2 = \frac{\frac{3}{2}B\log\frac{3}{\delta}}{n}$. □

Lemma 8. *Under the hypotheses of Theorem 5 we have the following, with probability taken over the random sample (i.e. $\tilde{f}$, P_n, and $\tilde{\mathsf{P}}_n$ are functions of the random sample):*

$$\begin{aligned} &\Pr\left(\mathsf{P}\,\ell_{\tilde{f}} > \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2\right) \\ &\leq \Pr\left(\sup_{f\in\mathcal{F}}\left\{\mathsf{P}\,(\ell_f - \ell_{f^*}) - \mathsf{P}_n\,(\ell_f - \ell_{\hat{f}})\right\} > t_1\right) + \Pr\left(\sup_{f\in\mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| > t_2\right). \end{aligned}$$

Proof. First, observe that (using $[\![E]\!]$ for the 0-1 indicator function of a random event)

$$\begin{aligned} &\left[\!\!\left[\left(\sup_{f\in\mathcal{F}}\left\{\mathsf{P}\,(\ell_f - \ell_{f^*}) - \mathsf{P}_n\,(\ell_f - \ell_{\hat{f}})\right\} \leq t_1\right) \wedge \left(\sup_{f\in\mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| \leq t_2\right)\right]\!\!\right] \\ &\leq \left[\!\!\left[\mathsf{P}\,\ell_{\tilde{f}} \leq \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2\right]\!\!\right]. \end{aligned}$$

To see this,

$$\mathsf{P}\,\ell_{\tilde{f}} \leq \mathsf{P}\,\ell_{f^*} + \mathsf{P}_n\,(\ell_{\tilde{f}} - \ell_{\hat{f}}) + t_1 \leq \mathsf{P}\,\ell_{f^*} + 2\sup_{f\in\mathcal{F}}\left|(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f\right| + t_1,$$

where the second inequality is from Lemma 7.

By subtracting each side from one and rearranging, we get an implication on the negation of these events

$$\begin{aligned} &\left[\!\!\left[\left(\sup_{f\in\mathcal{F}}\left\{\mathsf{P}\,(\ell_f - \ell_{f^*}) - \mathsf{P}_n\,(\ell_f - \ell_{\hat{f}})\right\} > t_1\right) \vee \left(\sup_{f\in\mathcal{F}} |(\mathsf{P}_n - \tilde{\mathsf{P}}_n)\ell_f| > t_2\right)\right]\!\!\right] \\ &\geq \left[\!\!\left[\mathsf{P}\,\ell_{\tilde{f}} > \mathsf{P}\,\ell_{f^*} + t_1 + 2t_2 + t_3\right]\!\!\right] \end{aligned}$$

and we can use the union bound. □

E Lower Bounds for High Dimensions

Theorem 3. *If $\mathcal{X}$ is in the d-dimensional hypercube and the Lipschitz constant is $K = 1$, no algorithm for regression on variance of y can have nontrivial accuracy with $o(2^{d/2})$ samples.*

Proof. We consider the simplest nontrivial hypothesis class, constant functions (i.e. the set $\{f_c : c \in \mathbb{R}\}$ where each $f_c : x \mapsto c \ \ (\forall x)$. The instances we construct will be realizable, i.e. in each instance, there will exist a constant c such that $\mathrm{Var}(y \mid x) = c \ \ (\forall x)$.

Consider the discrete uniform distribution on the boolean hypercube $\mathcal{X} = \{0,1\}^d$. We have $\mathcal{Y} = \{0,1\}$. In instance A, $\Pr[y = 1 \mid x] = 0.5$ independently for all x, and $\mathrm{Var}(y \mid x) = 0.25$ for all x. In the family of instances

[4] The bound being used is $\Pr(S_n \geq R) \leq 2^{-R}$ for $R \geq 6\,\mathbb{E}[S_n]$, from equation (4.3) of [28], where S_n is the sum of i.i.d. Bernoulli random variables with success probability $1/n$.

$\mathcal{B}$, we construct an instance B by drawing, for each x, $\theta_x \in \{0,1\}$ uniformly and independently at random; then conditioned on x, the distribution of y is given by $\Pr[y = 1 \mid x] = \theta_x$. Notice that for all instances in the family $\mathcal{B}$, $\mathrm{Var}(y \mid x) = 0$ for every x. Also, the Lipschitz constant satisfied by instances in $\mathcal{B}$ is 1, as all distinct x lie at a distance at least 1 from each other.

Informal sketch. Any algorithm that is accurate must, with probability close to 1, produce a different output when given access to A than when given access to a uniform instance from B. However, by the principle of deferred decisions, we can rewrite the algorithm's behavior in the latter case as follows: Each time a uniformly random x is drawn, if the algorithm has already seen a sample from x, then set y consistent with that previous sample; otherwise, draw θ_x uniformly at random and set $y = \theta_x$.

Thus, the input to the algorithm is distributed exactly identically in both cases unless the algorithm obtains multiple samples from the same x. However, with $o(2^{0.5d})$ samples, the probability of this occurring is $o(1)$ (by the "birthday paradox"), so the algorithm has the same distribution of outputs with probability $1 - o(1)$.

Formal proof. Let M be an algorithm and write $M(A)$ for the random variable which is M's hypothesis when run on samples from A, while $M(B)$ is M's hypothesis when run on samples from a uniformly randomly chosen instance B from the family $\mathcal{B}$. Suppose that M satisfies that, with probability at least $\frac{2}{3}$, its hypothesis (which is some constant c) is within ε of the correct variance, i.e. $M(A) \geq 0.25 - \varepsilon$ and $M(B) \leq \varepsilon$ each with probability at least $\frac{2}{3}$.[5] Suppose $\varepsilon < 0.125$ and the number of samples drawn by M is $o(2^{0.5d})$; we show a contradiction.

Use s to denote a set of samples each of the form (x, y), and let NR denote those sample-sets which have "no repeated x's", i.e. $NR = \{s : \text{each } x \text{ in } s \text{ is unique}\}$. Use $\Pr_A[s]$ to denote the probability of drawing a set of samples s given access to A, with $\Pr_B[s]$ the probability of drawing s given access to a uniformly random instance from B, and so on. We have

$$\begin{aligned}
\frac{2}{3} &\leq \Pr[M(A) \geq 0.25 - \varepsilon] \\
&= \sum_s \Pr_A[s] \Pr[M(s) \geq 0.25 - \varepsilon] \\
&= \sum_{s \in NR} \Pr_A[s] \Pr[M(s) \geq 0.25 - \varepsilon] \;+\; \sum_{s \notin NR} \Pr_A[s] \Pr[M(s) \geq 0.15] \\
&\leq \sum_{s \in NR} \Pr_A[s] \Pr[M(s) \geq 0.25 - \varepsilon] \;+\; \sum_{s \notin NR} \Pr_A[s]. \qquad (1)
\end{aligned}$$

Now, for all $s \in NR$, we claim $\Pr_A[s] = \Pr_B[s]$, as each is equal to

$$\prod_{(x,y) \in s} \Pr[x] \Pr[y \mid x] = \prod_{(x,y) \in s} 2^{-d} \left(\frac{1}{2}\right).$$

(In the case of A, this is immediate; in the case of B, by the principle of deferred decisions, we can construct B piece-by-piece; as each sample $(x, y) \in s \cap NR$ is drawn, we draw $\theta_x \in \{0,1\}$ uniformly at random and set $\Pr[y = 1] = \theta_x$, which results in a uniform distribution on y.)

Meanwhile, $\sum_{s \notin NR} \Pr_A[s]$ is the probability of drawing a sample with some repeated x value, which we claim is $o(1)$ with $o(2^{0.5d})$ samples. The distribution is uniform on the 2^d possible x values. The probability of a repeat or "collision", by Markov's inequality, is at most the expected number of collisions; with m samples, there are $\binom{m}{2}$ pairs each with a 2^{-d} chance of collision, so the expected number of collisions is $O\left(\frac{m^2}{2^d}\right)$, which is $o(1)$ for $m = o(2^{0.5d})$.

So with $o(2^{0.5d})$ samples, we have by (1) that

$$\begin{aligned}
\frac{2}{3} &\leq \sum_{s \in NR} \Pr_B[s] \Pr[M(s) \geq 0.25 - \varepsilon] \;+\; o(1) \\
&\leq \Pr[M(B) \geq 0.25 - \varepsilon] + o(1)
\end{aligned}$$

[5] One can also state this condition as an $\approx \varepsilon$ generalization error guarantee for M with appropriate loss function.

which, if $\varepsilon < 0.125$, implies that $\Pr[M(B) \leq \varepsilon] \leq \frac{1}{3} + o(1)$. This contradicts the accuracy assumption that $\Pr[M(B) \leq \varepsilon] \geq \frac{2}{3}$, so with this small number of samples, no such accurate M exists. □

Theorem 6. *With a uniform distribution on the unit hypercube $[0,1]^d$ in d dimensions, with Lipschitz constant $K = d$, there is no algorithm for regression on variance with nontrivial accuracy drawing $o(2^{0.5d})$ samples.*

More precisely, estimating average variance over the hypercube to accuracy $\varepsilon < \frac{1}{32}$ with success rate at least $\frac{2}{3}$ requires $\Omega(2^{0.5d})$ samples.

Proof. The construction is very similar to the Boolean hypercube above. We have $\mathcal{Y} = \{0,1\}$. On instance A, for every x, $\Pr[y = 1 \mid x] = 0.5$. Hence, $\mathrm{Var}(y \mid x) = 0.25$ for all x.

Constructing $\mathcal{B}$. We now construct the family of instances $\mathcal{B}$ and show that each has Lipschitz constant $K \leq d$ and has average variance $\mathbb{E}_x \mathrm{Var}(y \mid x) \leq \frac{3}{16}$.

In each instance, the hypercube is divided into "corners" and "interior regions". Let $\beta = \frac{1}{2}\frac{1}{2d}$ and let $\alpha = \frac{1}{2} - \beta$. Each "corner" C_v is a hypercube of side length α, inscribed in the unit hypercube and sharing the vertex $v \in \{0,1\}^d$. In other words, $C_v = \{x : \|x - v\|_\infty \leq \alpha\}$. The portions of $[0,1]^d$ not contained in any corner are considered the "interior regions".

To construct an instance in the family $\mathcal{B}$, draw $\theta_v \in \{0,1\}$ i.i.d. uniformly for each v. For points x in some corner C_v, we have $\Pr[y = 1 \mid x] = \theta_v$. For points x in the interior regions, let the notation $\|x - C_v\|_2$ denote $\min_{x' \in C_v} \|x' - x\|_2$. If there exists a v such that $\|x - C_v\|_2 < \beta$, then, letting $r = \|x - C_v\|_2$, set $\Pr[y = 1 \mid x] = \left(1 - \frac{r}{\beta}\right)\theta_v + \left(\frac{r}{\beta}\right)\left(\frac{1}{2}\right)$. (Note that this can only be true for at most one v, as this implies $\|x - v\|_\infty < \frac{1}{2}$, i.e. x is contained within the "corner" of side length $\frac{1}{2}$ touching v.) For all other x (those not within β of any C_v), we have $\Pr[y = 1 \mid x] = \frac{1}{2}$.

Now we show that any instance in family B has Lipschitz constant $K = d$. For shorthand, write $p_x := \Pr[y = 1 \mid x]$. Note that total variation distance between the distributions on y at x and at x' is $|p_x - p_{x'}|$. The Lipschitz constant is bounded by the maximal directional derivative of p_x with respect to x in any direction. This is zero if x lies within some C_v or if x is not within distance β of some C_v. Otherwise, if $x' = \mathrm{argmin}_{x'' \in C_v} \|x'' - x\|_2$, then the absolute value of the directional derivative is maximized in the direction $x' - x$, where it is $\frac{1}{2\beta}$. This gives a Lipschitz constant of $\frac{1}{2\beta}$.

Now we bound the average variance of y given x. The volume of each corner C_v is α^d and there are 2^d corners, so the total volume of the corners is

$$
\begin{aligned}
2^d \alpha^d &= 2^d \left(\frac{1}{2} - \frac{1}{2}\frac{1}{d}\right)^d \\
&= \left(1 - \frac{1}{d}\right)^d \\
&\geq \left(1 - \frac{1}{2}\right)^2 \\
&= \frac{1}{4}
\end{aligned}
$$

assuming $d \geq 2$. For any $x \in C_v$ for any v, $\mathrm{Var}(y \mid x) = 0$. For all other x, $\mathrm{Var}(y \mid x) \leq 0.25$; and the volume computation shows that they make up at most $\frac{3}{4}$ of the hypercube. Hence, average variance in this instance is at most $0.25\left(\frac{3}{4}\right) = \frac{3}{16}$.

(We note that, by letting $\beta = \frac{1}{2}\frac{1}{Cd}$ for $C \geq 1$, the same derivation gives Lipschitz constant $K = \frac{C}{d}$ and an average variance bounded by $\frac{4C-1}{16C^2} \leq \frac{1}{4C}$.)

Indistinguishability. From here, the proof is almost identical to the Boolean case. Consider any algorithm M that is with $\varepsilon < \frac{1}{32}$ of the correct average variance with probability at least $\frac{2}{3}$. For each vertex v, let $R_v = \{x : \|x - v\|_\infty < \frac{1}{2}\}$. There are 2^d disjoint regions R_v, each with volume $\frac{1}{2^d}$, and with probability 1, each sample (x, y) has $x \in R_v$ for some v.

Let s denote a set of samples drawn by M and let $NR = \{s : s \text{ has no repeated } xs\}$. If M draws $o(2^d)$ samples, then $\Pr[s \in NR] = o(1)$. If $s \notin NR$, then we claim $\Pr_A[s] = \Pr_B[s]$, where the notation is shorthand for the probability of drawing the samples s given oracle access to A, or to a uniformly chosen member B of $\mathcal{B}$, respectively. The reason is that, by the principle of deferred decisions, we can in the case of B choose θ_v at the moment that a sample (x, y) is drawn with $x \in R_v$, which occurs at most once for each v because $s \notin NR$. Because the distribution on θ_v is uniform $\{0, 1\}$, for any $x \in R_v$, the distribution of $p_x = \Pr[y = 1 \mid x]$ is uniform around $\frac{1}{2}$, or in other words, the unconditional probability that this $y = 1$ is exactly $\frac{1}{2}$.

So,

$$\begin{aligned}
\frac{2}{3} &\leq \Pr[M(A) \geq 0.25 - \varepsilon] \\
&= \sum_{s \in NR} \Pr_A[s] \Pr[M(s) \geq 0.25 - \varepsilon] \ + \ o(1) \\
&= \sum_{s \in NR} \Pr_B[s] \Pr[M(s) \geq 0.25 - \varepsilon] \ + \ o(1) \\
&\leq \Pr[M(B) \geq 0.25 - \varepsilon] \ + \ o(1).
\end{aligned}$$

For $\varepsilon < \frac{1}{32}$, this contradicts the accuracy requirement that $\Pr[M(B) \leq \frac{3}{16} + \varepsilon]$ with probability $\frac{2}{3}$. □

F Supporting arguments for Lipschitz regression lower bound

In this section, we show how the the minimax lower bound of [21] implies Corollary 1, our minimax lower bound for predictors that adopt the two-estimator (single-observation) approach. All notation in this section is from [21].

Proof of Corollary 1 The goal can be equivalently stated as minimizing the $L_2(\mathcal{D})$-estimation error of $\mathrm{Var}[Y \mid X]$, i.e., to find an estimator $\hat{f}$ for which $\mathbb{E}\left[\left(\hat{f}(X) - \mathrm{Var}[Y \mid X]\right)^2\right]$ is as small as possible. In the two-estimator approach, $\hat{f}$ takes the form $\hat{f} = \hat{g} - \hat{h}$ for estimators $\hat{g}$ and $\hat{h}$ of $\mathbb{E}[Y^2 \mid X]$ and $(\mathbb{E}[Y \mid X])^2$ respectively. Then

$$\mathbb{E}\left[\left(\hat{f}(X) - \mathrm{Var}[Y \mid X]\right)^2\right] = \mathbb{E}\left[\left(\left(\hat{g}(X) - \mathbb{E}\left[Y^2 \mid X\right]\right) + \left(\mathbb{E}\left[Y \mid X\right]^2 - \hat{h}(x)\right)\right)^2\right]. \tag{2}$$

The two-estimator approach needs to ensure that $L_2(\mathcal{D})$ norm of $\hat{g}(X) - \mathbb{E}\left[Y^2 \mid X\right]$ is small (and likewise for the second term). Suppose that $Y \mid X = x$ is Bernoulli for any $x \in \mathcal{X}$. Then this latter necessary goal reduces to the familiar regression problem of minimizing $\mathbb{E}\left[(\hat{g}(X) - \mathbb{E}[Y \mid X])^2\right]$. The above minimax lower bound will apply to the above estimation problem in the following setting (for full details see § F): let $\mathcal{D}$ be the uniform distribution over $\mathcal{X} = [0, 1]^d$. Suppose for all $x \in \mathcal{X}$ that the distribution $\mathcal{D}_x$ is a certain subclass of Bernoulli distributions with $x \mapsto \mathbb{E}[Y \mid X = x]$ a K-Lipschitz function. Then for any estimator $\hat{g}$, there exists a law $Y \mid X$ satisfying the aforementioned assumptions such that $\mathbb{E}\left[(\hat{g}(X) - \mathbb{E}[Y \mid X])^2\right] = \Omega\left(n^{-2/(2+d)}\right)$, and there is a matching upper bound, so that this is the minimax optimal rate of convergence. Moreover, *the same lower bound holds in a similar setting even for active learning strategies* [22, Theorem 1].

Take $\mathcal{X} = [0, 1]^d$ and let $\mathcal{D}$ be the uniform distribution over $\mathcal{X}$. For all $x \in \mathcal{X}$, the distribution $\mathcal{D}_x$ is a certain subclass of Bernoulli distributions[6] with $x \mapsto \mathbb{E}[Y \mid X = x]$ a K-Lipschitz function. Then for any estimator $\hat{g}$, there exists a law $Y \mid X$ satisfying the aforementioned assumptions such that

$$\mathbb{E}\left[(\hat{g}(X) - \mathbb{E}[Y \mid X])^2\right] = \Omega\left(n^{-2/(2+d)}\right),$$

and there is a matching upper bound, so that this is the minimax optimal rate of convergence for this problem.

To see how the above result follows from [21], we take $T(\theta) = \theta$ and observe that the Lipschitz condition is reflected in Stone's equation (1.2) by setting K_2 to our K, k to zero, and $\beta = 1$ (so that $p = k + \beta = 1$ and hence $r = 1/(2 + d)$). Since θ is the parameter of a Bernoulli distribution, we need to verify from the lower bound construction that for all choices of $\theta \in \Theta_n$ used in the lower bound, we have $\theta \in (0, 1)$. We now do this verification.

[6] See Condition 2 of [21] for details; in our setting, we have $t = \theta(x) = \mathbb{E}[Y \mid X = x]$, so the Bernoulli distribution can vary with x, as further explained on p. 1350 of [29].

As mentioned in the proof of Lemma 1 of [21], for any binary sequence $\tau_n \in \{0,1\}^{V_n}$, the corresponding[7] $g_n^{(\tau)}$ vanishes at the boundary of $[0,1]^d$. Moreover, by assumption each $g_n^{(\tau)}$ is K-Lipschitz for $K = 1/(2\sqrt{d})$. Let us verify that, for all $\theta \in \Theta_n$ and for all $x \in [0,1]^d$, it holds that $\theta(x) \in (0,1)$. Let $\theta_0 \equiv \frac{1}{2}$, and note that $\theta(x) = \theta_0(x) + g^{(\tau)}(x)$. It suffices to show that $|g_n^{(\tau)}(x)| \leq \frac{1}{2\sqrt{2}}$ for all $\tau \in \{0,1\}^d$ and all $x \in [0,1]^d$. This is easily verified. First, observe that

$$|g_n^{(\tau)}(x)| = \left| g_n^{(\tau)}(x) - g_n^{(\tau)}(\mathbf{0}) \right| = \left| g_n^{(\tau)}(x) - g_n^{(\tau)}(\mathbf{1}) \right|,$$

since $g^{(\tau)}$ vanishing at the boundary of $[0,1]^d$ implies that $g_n^{(\tau)}(\mathbf{0}) = g_n^{(\tau)}(\mathbf{1}) = 0$. Therefore,

$$\begin{aligned} |g_n^{(\tau)}(x)| &= \min\left\{ \left| g_n^{(\tau)}(x) - g_n^{(\tau)}(\mathbf{1}) \right|, \left| g_n^{(\tau)}(x) - g_n^{(\tau)}(\mathbf{0}) \right| \right\} \\ &\leq K \min\left\{ \|x - \mathbf{1}\|, \|x - \mathbf{0}\| \right\} \\ &= K \left\| \frac{1}{2} \cdot \mathbf{1} \right\| \\ &= K\sqrt{d/2} \\ &= \frac{1}{2\sqrt{2}}. \end{aligned}$$

The result then follows by applying Theorem 1 of [21] with $q = 2$.

□

G Eliciting the Upper Confidence Bound

Given a random variable Y, define $\mathrm{ucb}_\lambda(Y) = \mathbb{E}[Y] + \lambda\sigma[Y]$, where $\sigma[Y] = \sqrt{\mathrm{Var}[Y]} = \sqrt{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}$. We show here the motivation behind the loss function used in § 5.2.

The level sets of ucb_λ, as a function of the *law* of Y, i.e., the distribution, are given by

$$r = \mathrm{ucb}_\lambda(Y) = \mathbb{E}[Y] + \lambda\sigma[Y] \ . \tag{3}$$

We can rewrite this as follows:

$$\begin{aligned} r - \mathbb{E}[Y] &= \lambda\sigma[Y] && (4) \\ (r - \mathbb{E}[Y])^2 &= \lambda^2 \mathrm{Var}[Y] && (5) \\ 0 &= \lambda^2 \mathrm{Var}[Y] - \mathbb{E}[Y]^2 + 2r\,\mathbb{E}[Y] - r^2 \ , && (6) \end{aligned}$$

though note that we have introduced another solution: both $r = \mathbb{E}[Y] + \lambda\sigma[Y]$ and $r = \mathbb{E}[Y] - \lambda\sigma[Y]$ now satisfy eq. (6) but only the former satisfies eq. (3). Apart from this spurious solution, the following would be an identification function for ucb_λ, meaning a distribution has zero expectation if and only if it is in the level set for r; see [30, 31].

$$V(r, y_1, y_2) = \frac{\lambda^2}{2}(y_1 - y_2)^2 - y_1 y_2 + (y_1 + y_2)r - r^2 \ , \tag{7}$$

whence

$$\mathbb{E}[V(r, Y_1, Y_2)] = \lambda^2 \mathrm{Var}[Y] - \mathbb{E}[Y]^2 + 2\,\mathbb{E}[Y]r - r^2 \ , \tag{8}$$

where of course $Y_1, Y_2 \sim Y$ are independent.

Despite the fact that V does not completely identify ucb_λ, as for a given r, distributions where $\mathbb{E}[Y] - \sigma[Y] = r$ also satisfy $\mathbb{E}[V(r, Y_1, Y_2)] = 0$, we can still try to integrate $-V$ with respect to r to get a loss function. The result is

$$\ell(r, y_1, y_2) = -\left(\frac{\lambda^2}{2}(y_1 - y_2)^2 - y_1 y_2 \right) r - \frac{1}{2}(y_1 + y_2)r^2 + \frac{1}{3}r^3 \ . \tag{9}$$

[7] We use the notation $g_n^{(\tau)}$ rather than Stone's notation g_n to make explicit the dependence on τ.

As V was not a true identification function, we know that $\mathbb{E}[\ell(r, Y_1, Y_2)]$ has multiple extrema, at $r = \mathbb{E}[Y] \pm \lambda\sigma[Y]$. What's worse, since ℓ is cubic, we see that one can actually achieve arbitrarily negative loss as $r \to -\infty$. Still, we can impose conditions on r and Y so that $r = \text{ucb}_\lambda(Y)$ is the unique minimizer of $\mathbb{E}[\ell]$. In particular, if we restrict r to the range $\mathbb{E}[Y] - 2\lambda\sigma[Y] \leq r \leq \infty$, then the loss will elicit ucb_λ. If we further restrict $r \geq \mathbb{E}[Y] - \lambda\sigma[Y]$, then $\mathbb{E}[\ell]$ is quasi-convex. Finally, if we further restrict $r \geq \mathbb{E}[Y]$, then $\mathbb{E}[\ell]$ is convex.

Another possible loss is obtained by integrating $-rV$, that is, r times $-V$. This gives us,

$$\ell(r, y_1, y_2) = -\frac{1}{2}\left(\frac{\lambda^2}{2}(y_1 - y_2)^2 - y_1 y_2\right) r^2 - \frac{1}{3}(y_1 + y_2) r^3 + \frac{1}{4} r^4 \ .$$

This removes the problem of unbounded negative loss for incorrect reports, but can still have a local optimum at $r = \mathbb{E}[Y] - \lambda\sigma[Y]$.

H Simulation Details

Algorithm implementation. For Algorithm 2 and the ε-Nearby algorithm, we used $\varepsilon = \frac{1}{2\sqrt{n}}$, which we found to generally perform the best. There is some question about how to apply Algorithm 2 in the setting where one is given a fixed set of samples, rather than an oracle for drawing samples. The strategy that we found to work best, which we used here, was given $\hat{n}$ samples, to use the first n for $X_1^*, \ldots, X_n^*$ and the remaining $N = \hat{n} - n$ for $X_1^{(1)}, \ldots, X_N^{(1)}$. We then binary searched to find the largest n which allowed a perfect matching.[8] In all of our experiments, our two-observation methods used the linear functions as their hypothesis class. In all cases, the true statistic is within this class.

Monte Carlo approach. For ucb_λ we compared against the standard strategy used in practice. This strategy is to sample n/k random points $x_1, \ldots, x_{n/k}$ from X, and then for each x_i sample k values $y_{i,1}, \ldots, y_{i,k} \sim Y|X{=}x_i$. For each i, we compute the empirical ucb_λ u_i of $y_{i,1}, \ldots, y_{i,k}$ and then fit a line to $(x_1, u_1), \ldots, (x_{n/k}, u_{n/k})$ via least-squares regression. We found that best results were achieved by letting $k = \sqrt{n}$.

$\alpha_k(x)$ implementation. For the 2-norm, we chose our distribution $\alpha_k(x) = (Y|X{=}x)$ so that $||\alpha_k(x)||_2^2 = 1/2$ for all $x \in [0,1]$ and the support is always at most 3, but what values make up that support shift with x. We constructed $\alpha_k(x)$ as follows. $\mathcal{Y} = \{0, \ldots, k-1\}$. Given $x \in [0,1)$, choose $r \in \{0, \ldots, k-3\}$ such that $r \leq (k-2)x \leq r+1$. Let $a = (x(k-2) - r)/2$ and $b = (2a - 1 + \sqrt{1 + 4a - 12a^2})/4$. Then $\Pr[\alpha_k(x) = r] = 1/2 - a$, $\Pr[\alpha_k(x) = r+1] = 1/2 + b$, $\Pr[\alpha_k(x) = r+2] = a - b$, and all other outcomes have probability 0.

Evaluation framework. As we used simple underlying statistics (1, 1/2, and $x + 10$) and simple hypothesis classes, we were able to compute the mean squared error between an algorithm's reported hypothesis and the true statistic via closed form expressions. The only exception to this was for the two moment method of learning ucb_λ, where the instead we estimated the mean squared error via 1000 sample Monte Carlo integration. Each data point in Figure 1 is the median of 1000 independent trials.

[8] Note that this modification may introduce some bias so our theoretical results about Algorithm 2 no longer directly apply. Nevertheless, we found this modification to be effective in practice.