Journal of Machine Learning Research 21 (2020) 1-54 Submitted 6/19; Revised 7/20; Published 8/20

Regression with Comparisons: Escaping the Curse of Dimensionality with Ordinal Information

Yichong Xu YICHONGX@CS.CMU.EDU
Machine Learning Department

Sivaraman Balakrishnan SIVA@STAT.CMU.EDU
Department of Statistics and Data Science
Machine Learning Department

Aarti Singh AARTI@CS.CMU.EDU
Machine Learning Department

Artur Dubrawski AWD@CS.CMU.EDU
Auton Lab, The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Sanjoy Dasgupta

Abstract

In supervised learning, we typically leverage a fully labeled dataset to design methods for function estimation or prediction. In many practical situations, we are able to obtain alternative feedback, possibly at a low cost. A broad goal is to understand the usefulness of, and to design algorithms to exploit, this alternative feedback. In this paper, we consider a semi-supervised regression setting, where we obtain additional *ordinal* (or *comparison*) information for the unlabeled samples. We consider ordinal feedback of varying qualities where we have either a perfect ordering of the samples, a noisy ordering of the samples or noisy pairwise comparisons between the samples. We provide a precise quantification of the usefulness of these types of ordinal feedback in both nonparametric and linear regression, showing that in many cases it is possible to accurately estimate an underlying function with a very small labeled set, effectively *escaping the curse of dimensionality*. We also present lower bounds, that establish fundamental limits for the task and show that our algorithms are optimal in a variety of settings. Finally, we present extensive experiments on new datasets that demonstrate the efficacy and practicality of our algorithms and investigate their robustness to various sources of noise and model misspecification.

Keywords: Pairwise Comparison, Ranking, Regression, Interactive Learning

1. Introduction

Classical regression is centered around the development and analysis of methods that use labeled observations, $\{(X_1, y_1), \ldots, (X_n, y_n)\}$, where each $(X_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, in various tasks of estimation and inference. Nonparametric and high-dimensional methods are appealing in practice owing to their flexibility, and the relatively weak a-priori structural assumptions that they impose on the unknown regression function. However, the price we pay is that these methods typically require a large amount of labeled data to estimate complex target

functions, scaling exponentially with the dimension for fully nonparametric methods and scaling linearly with the dimension for high-dimensional parametric methods – the so-called curse of dimensionality. This has motivated research on structural constraints – for instance, sparsity or manifold constraints – as well as research on active learning and semi-supervised learning where labeled samples are used judiciously.

We consider a complementary approach, motivated by applications in material science, crowdsourcing, and healthcare, where we are able to supplement a small labeled dataset with a potentially larger dataset of *ordinal* information. Such ordinal information is obtained either in the form of a (noisy) ranking of unlabeled points or in the form of (noisy) pairwise comparisons between function values at unlabeled points. Some illustrative applications of the methods we develop in this paper include:

Example 1: Crowdsourcing. In crowdsourcing we rely on human labeling effort, and in many cases humans are able to provide more accurate ordinal feedback with substantially less effort (see for instance Tsukida and Gupta (2011); Shah et al. (2016a,c)). When crowdsourced workers are asked to give numerical price estimates, they typically have difficulty giving a precise answer, resulting in a high variance between worker responses. On the other hand, when presented two products or listings side by side, workers may be more accurate at comparing them. We conduct experiments on the task of price estimation in Section 4.4.

Example 2: Material Synthesis. In material synthesis, the broad goal is to design complex new materials and machine learning approaches are gaining popularity (Xue et al., 2016; Faber et al., 2016). Typically, given a setting of input parameters (temperature, pressure etc.) we are able to perform a synthesis experiment and measure the quality of resulting synthesized material. Understanding this landscape of material quality is essentially a task of high-dimensional function estimation. Synthesis experiments can be costly and material scientists when presented with pairs of input parameters are often able to cheaply and reasonably accurately provide comparative assessments of quality.

Example 3: Patient Diagnosis. In clinical settings, precise assessment of each individual patient's health status can be difficult, expensive, and risky but comparing the relative status of two patients may be relatively easy and accurate.

In each of these settings, it is important to develop methods for function estimation that combine standard supervision with (potentially) cheaper and abundant ordinal or comparative supervision.

1.1. Our Contributions

We consider both linear and nonparametric regression with both direct (cardinal) and comparison (ordinal) information. In both cases, we consider the standard statistical learning setup, where the samples X are drawn i.i.d. from a distribution $\mathbb{P}_X$ on $\mathbb{R}^d$. The labels y are related to the features X as,

$$y = f(X) + \varepsilon,$$

where $f = \mathbb{E}[y|X]$ is the underlying regression function of interest, and ε is the mean-zero label noise. Our goal is to construct an estimator $\widehat{f}$ of f that has low risk or mean squared error (MSE),

$$R(\widehat{f}, f) = \mathbb{E}(\widehat{f}(X) - f(X))^2,$$

where the expectation is taken over the labeled and unlabeled training samples, as well as a new test point X. We also study the fundamental information-theoretic limits of estimation with classical and ordinal supervision by establishing lower (and upper) bounds on the minimax risk. Letting η denote various problem dependent parameters, which we introduce more formally in the sequel, the minimax risk:

$$\mathfrak{M}(m, n; \eta) = \inf_{\widehat{f}} \sup_{f \in \mathcal{F}_\eta} R(\widehat{f}, f), \tag{1}$$

provides an information-theoretic benchmark to assess the performance of an estimator. Here m, n denotes the amount of cardinal and ordinal information we acquire, and $\mathcal{F}_\eta$ is some relevant function class.

First, focusing on nonparametric regression, we develop a novel Ranking-Regression (R^2) algorithm for nonparametric regression that can leverage ordinal information, in addition to direct labels. We make the following contributions:

- To establish the usefulness of ordinal information in nonparametric regression, in Section 2.2 we consider the idealized setting where we obtain a perfect ordering of the unlabeled set. We show that the risk of the R^2 algorithm can be bounded with high-probability[1] as $\widetilde{O}(m^{-2/3} + n^{-2/d})$, where m denotes the number of labeled samples and n the number of ranked samples. To achieve an MSE of ε, the number of labeled samples required by R^2 is *independent* of the dimensionality of the input features. This result establishes that sufficient ordinal information of high quality can allow us to effectively circumvent the curse of dimensionality.

- In Sections 2.3 and 2.4 we analyze the R^2 algorithm when using either a noisy ranking of the samples or noisy pairwise comparisons between them. For noisy ranking, we show that the MSE is bounded by $\widetilde{O}(m^{-2/3} + \sqrt{\nu} + n^{-2/d})$, where ν is the Kendall-Tau distance between the true and noisy ranking. As a corollary, we combine R^2 with algorithms for ranking from pairwise comparisons (Braverman and Mossel, 2009) to obtain an MSE of $\widetilde{O}(m^{-2/3} + n^{-2/d})$ when $d \geq 4$, when the comparison noise is bounded.

Turning our attention to the setting of linear regression with ordinal information, we develop an algorithm that uses active (or adaptive) comparison queries in order to reduce both the label complexity and total query complexity.

- In Section 3 we develop and analyze an interactive learning algorithm that estimates a linear predictor using both labels and comparison queries. Given a budget of m label queries and n comparison queries, we show that MSE of our algorithm decays

1. We use the standard big-O notation throughout this paper, and use $\widetilde{O}$ when we suppress log-factors, and $\widehat{O}$ when we suppress log-log factors.

at the rate of $\widetilde{O}\left(1/m + \exp(-n/d)\right)$. Once again we see that when sufficiently many comparisons are available, the label complexity of our algorithm is *independent* of the dimension d.

To complement these results we also give information-theoretic lower bounds to characterize the fundamental limits of combining ordinal and standard supervision.

- For nonparametric regression, we show that the R^2 algorithm is optimal up to log factors both when it has access to a perfect ranking, as well as when the comparisons have bounded noise. For linear regression, we show that the rate of $O(1/m)$, and the total number of queries, are not improvable up to log factors.

On the empirical side we comprehensively evaluate the algorithms we propose, on simulated and real datasets.

- We use simulated data, and study the performance of our algorithms as we vary the noise in the labels and in the ranking.
- Second, we consider a practical application of predicting people's ages from photographs. For this dataset we obtain comparisons using people's apparent ages (as opposed to their true biological age).
- Finally, we curate a new dataset using crowdsourced data obtained through Amazon's Mechanical Turk. We provide workers AirBnB listings and attempt to estimate property asking prices. We obtain both direct and comparison queries for the listings, and also study the time taken by workers to provide these different types of feedback. We find that, our algorithms which combine direct and comparison queries are able to achieve significantly better accuracy than standard supervised regression methods, for a fixed time budget.

1.2. Related Work

As a classical way to reduce the labeling effort, active learning has been mostly focused on classification (Hanneke, 2009). For regression, it is known that, in many natural settings, the ability to make active queries does not lead to improved rates over passive baselines. For example, Chaudhuri et al. (2015) shows that when the underlying model is linear, the ability to make active queries can only improve the rate of convergence by a constant factor, and leads to no improvement when the feature distribution is spherical Gaussian. In Willett et al. (2006), the authors show a similar result that in nonparametric settings, active queries do not lead to a faster rate for smooth regression functions except when the regression function is piecewise smooth.

There is considerable work in supervised and unsupervised learning on incorporating additional types of feedback beyond labels. For instance, Zou et al. (2015); Poulis and Dasgupta (2017) study the benefits of different types of "feature feedback" in clustering and supervised learning respectively. Dasgupta and Luby (2017) consider learning with partial corrections, where the user provides corrective feedback to the algorithm when the algorithm makes an incorrect prediction. Ramaswamy et al. (2018) consider multiclass classification where the user can choose to abstain from making predictions at a cost.

There is also a vast literature on models and methods for analyzing pairwise comparison data (Tsukida and Gupta, 2011; Agarwal et al., 2017), like the classical Bradley-Terry (Bradley and Terry, 1952) and Thurstone (Thurstone, 1927) models. In these works, the typical focus is on ranking or quality estimation for a fixed set of objects. In contrast, we focus on function estimation and the resulting models and methods are quite different. We build on the work on "noisy sorting" (Braverman and Mossel, 2009; Shah et al., 2016b) to extract a consensus ranking from noisy pairwise comparisons. We also note in passing that a different type of paired feedback plays a prominent role in distance metric learning (Xing et al., 2003; Schultz and Joachims, 2004; Suárez et al., 2018). In this setting one uses similar/dissimilar pairs of data points to learn a metric over the feature space.

Perhaps closest in spirit to this work are the two recent papers (Kane et al., 2017; Xu et al., 2017) that consider binary classification with ordinal information. These works differ from the proposed approach in their focus on classification, emphasis on active querying strategies and binary-search-based methods.

Given ordinal information of sufficient fidelity, the problem of nonparametric regression is related to the problem of regression with shape constraints, or more specifically isotonic regression (Barlow, 1972; Zhang, 2002). Accordingly, we leverage such algorithms in our work and we comment further on the connections in Section 2.2. Some salient differences between this literature and our work are that we design methods that work in a semisupervised setting, and further that our target is an unknown d-dimensional (smooth) regression function as opposed to a univariate shape-constrained function.

The rest of this paper is organized as follows. In Section 2, we consider the problem of combining direct and comparison-based labels for nonparametric regression, providing upper and lower bounds for both noiseless and noisy ordinal models. In Section 3, we consider the problem of combining adaptively chosen direct and comparison-based labels for linear regression. In Section 4, we turn our attention to an empirical evaluation of our proposed methods on real and synthetic data. Finally, we conclude in Section 5 with a number of additional results and open directions. In the Appendix we present detailed proofs for various technical results and a few additional supplementary experimental results.

2. Nonparametric Regression with Ordinal Information

We now provide analysis for nonparametric regression. First, in Section 2.1 we establish the problem setup and notations. Then, we introduce the R^2 algorithm in Section 2.2 and analyze it under perfect rankings. Next, we analyze its performance for noisy rankings and comparisons in Sections 2.3 and 2.4.

2.1. Background and Problem Setup

We consider a nonparametric regression model with random design, i.e. we suppose first that we are given access to an unlabeled set $\mathcal{U} = \{X_1, \ldots, X_n\}$, where $X_i \in \mathcal{X} \subset [0,1]^d$, and X_i are drawn i.i.d. from a distribution $\mathbb{P}_X$ and we assume that $\mathbb{P}_X$ has a density p which is upper and lower bounded as $0 < p_{\min} \leq p(x) \leq p_{\max}$ for $x \in \mathcal{X}$. Our goal is to estimate a function $f : \mathcal{X} \mapsto \mathbb{R}$, where following classical work (Györfi et al., 2006; Tsybakov, 2008) we

assume that f is bounded in $[-M, M]$ and belongs to a Hölder ball $\mathcal{F}_{s,L}$, with $0 < s \leq 1$:

$$\mathcal{F}_{s,L} = \{f : |f(x) - f(y)| \leq L\|x - y\|_2^s, \forall\ x, y \in \mathcal{X}\}.$$

For $s = 1$ this is the class of Lipschitz functions. We discuss the estimation of smoother functions (i.e. the case when $s > 1$) in Section 5. We obtain two forms of supervision:

1. **Classical supervision:** For a (uniformly) randomly chosen subset $\mathcal{L} \subseteq \mathcal{U}$ of size m (we assume throughout that $m \leq n$ and focus on settings where $m \ll n$) we make noisy observations of the form:

   $$y_i = f(X_i) + \epsilon_i, \quad i \in \mathcal{L},$$

   where ϵ_i are i.i.d. Gaussian with $\mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = 1$. We denote the indices of the labeled samples as $\{t_1, \ldots, t_m\} \subset \{1, \ldots, n\}$.

2. **Ordinal supervision:** For the given dataset $\{X_1, \ldots, X_n\}$ we let π denote the *true ordering*, i.e. π is a permutation of $\{1, \ldots, n\}$ such that for $i, j \in \{1, \ldots, n\}$, with $\pi(i) \leq \pi(j)$ we have that $f(X_i) \leq f(X_j)$. We assume access to one of the following types of ordinal supervision:

   (1) We assume that we are given access to a noisy ranking $\widehat{\pi}$, i.e. for a parameter $\nu \in [0, 1]$ we assume that the Kendall-Tau distance between $\widehat{\pi}$ and the true-ordering is upper-bounded as[2]:

   $$\sum_{i,j \in [n]} \mathbb{I}[(\pi(i) - \pi(j))(\widehat{\pi}(i) - \widehat{\pi}(j)) < 0] \leq \nu n^2. \tag{2}$$

   We denote the set of rankings with a Kendall-Tau distance at most νn^2 by $\Pi(\nu)$.

   (2) For each pair of samples (X_i, X_j), with $i < j$ we obtain a comparison Z_{ij} where for some constant $\lambda > 0$:

   $$\mathbb{P}(Z_{ij} = \mathbb{I}(f(X_i) > f(X_j))) \geq \frac{1}{2} + \lambda. \tag{3}$$

   As we discuss in Section 2.4, it is straightforward to extend our results to a setting where only a randomly chosen subset of all pairwise comparisons are observed.

Although classic supervised learning learns a regression function with labels only and without ordinal supervision, we note that learning cannot happen in the opposite way: That is, we cannot consistently estimate the underlying function with only ordinal supervision and without labels: the underlying function is only identifiable up to certain monotonic transformations.

As discussed in Section 1.1, our goal is to design an estimator $\widehat{f}$ of f that achieves the minimax mean squared error (1), when $f \in \mathcal{F}_{s,L}$. We conclude this section recalling a well-known fact: given access to only classical supervision the minimax risk $\mathfrak{M}(m; \eta) = \Theta(m^{-\frac{2s}{2s+d}})$, suffers from an exponential curse of dimensionality (Györfi et al., 2006, Theorem 5.2).

2. We use the standard notation $[\kappa]$ to represent the set $\{1, 2, ..., \kappa\}$ for any positive integer κ.

2.2. Nonparametric Regression with Perfect Ranking

To ascertain the value of ordinal information we first consider an idealized setting, where we are given a perfect ranking π of the unlabeled samples in $\mathcal{U}$. We present our Ranking-Regression (R^2) algorithm with performance guarantees in Section 2.2.1, and a lower bound for it in Section 2.2.2 which shows that R^2 is optimal up to log factors.

2.2.1. Upper bounds for the R^2 Algorithm

Algorithm 1 R^2: Ranking-Regression

Input: Unlabeled data $\mathcal{U} = \{X_1, \ldots, X_n\}$, a labeled subset of $\mathcal{U}$ of size m, i.e. samples $\{(X_{t_1}, y_{t_1}), \ldots, (X_{t_m}, y_{t_m})\}$, and a ranking $\widehat{\pi}$ of the points in $\mathcal{U}$.

1: Order elements in $\mathcal{U}$ as $(X_{\widehat{\pi}(1)}, \ldots, X_{\widehat{\pi}(n)})$.

2: Run isotonic regression (see (4)) on $\{y_{t_1}, \ldots, y_{t_m}\}$. Denote the estimated values by $\{\widehat{y}_{t_1}, \ldots, \widehat{y}_{t_m}\}$.

3: For $i = 1, 2, \ldots, n$, let $\beta(i) = t_k$, where $\widehat{\pi}(t_k)$ is the largest value such that $\widehat{\pi}(t_k) \leq \widehat{\pi}(i), k = 1, 2, \ldots, m$, and $\beta(i) = \star$ if no such t_k exists. Set

$$\widehat{y}_i = \begin{cases} \widehat{y}_{\beta(i)} & \text{if } \beta(i) \neq \star \\ 0 & \text{otherwise.} \end{cases}$$

Output: Function $\widehat{f} =$ NearestNeighbor$(\{(X_i, \widehat{y}_i)\}_{i=1}^n)$.

Our nonparametric regression estimator is described in Algorithm 1 and Figure 1. We first rank all the samples in $\mathcal{U}$ according to the (given or estimated) permutation $\widehat{\pi}$. We then run isotonic regression (Barlow, 1972) on the labeled samples in $\mathcal{L}$ to de-noise them and borrow statistical strength. In more detail, we solve the following program to de-noise the labeled samples:

$$\begin{aligned} \min_{\{\widehat{y}_{\widehat{\pi}(t_1)}, \ldots, \widehat{y}_{\widehat{\pi}(t_m)}\}} & \sum_{k=1}^{m} (\widehat{y}_{\widehat{\pi}(t_k)} - y_{\widehat{\pi}(t_k)})^2 \\ \text{s.t.} \quad & \widehat{y}_{t_k} \leq \widehat{y}_{t_l} \ \ \forall \ (k, l) \ \text{ such that } \ \widehat{\pi}(t_k) < \widehat{\pi}(t_l) \\ & -M \leq \{\hat{y}_{\widehat{\pi}(t_1)}, \ldots, \hat{y}_{\widehat{\pi}(t_m)}\} \leq M. \end{aligned} \tag{4}$$

We introduce the bounds $\{M, -M\}$ in the above program to ease our analysis. In our experiments, we simply set M to be a large positive value so that it has no influence on our estimator. We then leverage the ordinal information in $\widehat{\pi}$ to impute regression estimates for the unlabeled samples in $\mathcal{U}$, by assigning each unlabeled sample the value of the nearest (de-noised) labeled sample which has a smaller function value according to $\widehat{\pi}$. Finally, for a new test point, we use the imputed (or estimated) function value of the nearest neighbor in $\mathcal{U}$.

In the setting where we use a perfect ranking the following theorem characterizes the performance of R^2:

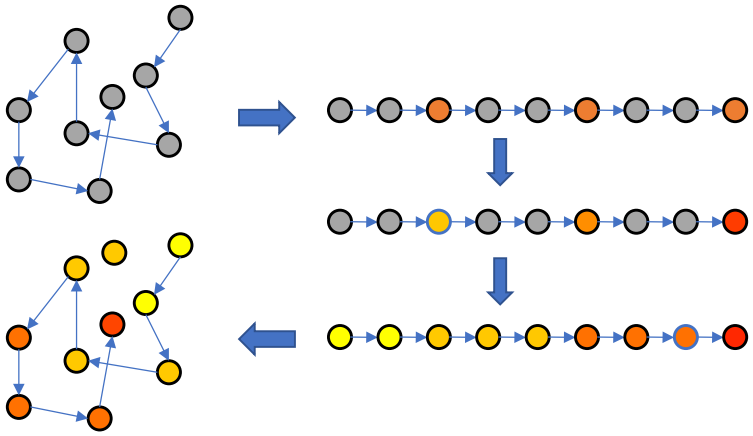


Figure 1: Top Left: A group of unlabeled points are ranked according to function values using ordinal information only. Top Right: We obtain function values of m randomly chosen samples. Middle Right: The values are adjusted using isotonic regression. Bottom Right: Function values of other unlabeled points are inferred. Bottom Left: For a new point, the estimated value is given by the nearest neighbor in $\mathcal{U}$.

Theorem 1 *Suppose that the ranking $\widehat{\pi}$ is perfect. If the target function f belongs to the Hölder ball $\mathcal{F}_{s,L}$ with $0 < s \leq 1$ and constant L, there exist constants $C_1, C_2 > 0$ such that the MSE of $\widehat{f}$ is bounded by,*

$$\mathbb{E}(\widehat{f}(X) - f(X))^2 \leq C_1 m^{-2/3} \log^2 n \log m + C_2 n^{-2s/d}. \tag{5}$$

Before we turn our attention to the proof of this result, we examine some consequences.
Remarks: (1) Theorem 1 shows a surprising dependency on the sizes of the labeled and unlabeled sets (m and n). The MSE of nonparametric regression using only the labeled samples is $\Theta(m^{-\frac{2s}{2s+d}})$ which is exponential in d and makes nonparametric regression impractical in high-dimensions. Focusing on the dependence on m, Theorem 1 improves the rate to $m^{-2/3}\text{polylog}(m, n)$, which is no longer exponential. By using enough ordinal information we can avoid the curse of dimensionality.
(2) On the other hand, the dependence on n (which dictates the amount of ordinal information needed) is still exponential. This illustrates that ordinal information is most beneficial when it is copious. We show in Section 2.2.2 that this is unimprovable in an information-theoretic sense.
(3) Somewhat surprisingly, we also observe that the dependence on n is faster than the $n^{-\frac{2s}{2s+d}}$ rate that would be obtained if all the samples were labeled. Intuitively, this is because of the noise in labels: Given the m labels along with the (perfect) ranking, the difference between two neighboring labels is typically very small (around $1/m$). Therefore, any unlabeled points in $\mathcal{U} \setminus \mathcal{L}$ will be restricted to an interval much narrower than the constant noise in direct labels. In the case where all points are labeled (i.e., $m = n$), the MSE is of order $n^{-2/3} + n^{-2s/d}$, also slightly better rate than when no ordinal information

is available. On the other hand, the improvement is stronger when $m \ll n$.
(4) Finally, we also note in passing that the above theorem provides an upper bound on the minimax risk in (1).

Proof Sketch We provide a brief outline and defer technical details to the Supplementary Material. For a randomly drawn point $X \in \mathcal{X}$, we denote by X_α the nearest neighbor of X in $\mathcal{U}$. We decompose the MSE as

$$\mathbb{E}\left[(\widehat{f}(X) - f(X))^2\right] \leq 2\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] + 2\mathbb{E}\left[(f(X_\alpha) - f(X))^2\right]. \tag{6}$$

The second term corresponds roughly to the finite-sample bias induced by the discrepancy between the function value at X and the closest labeled sample. We use standard sample-spacing arguments (see Györfi et al. (2006)) to bound this term. This term contributes the $n^{-2s/d}$ rate to the final result. For the first term, we show a technical result in the Appendix (Lemma 14). Without loss of generality suppose $f\big(X_{t_1}\big) \leq \cdots f\big(X_{t_m}\big)$. By conditioning on a probable configuration of the points and enumerating over choices of the nearest neighbor we find that roughly (see Lemma 14 for a precise statement):

$$\begin{aligned}&\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] \leq \Big(\frac{\log^2 n \log m}{m}\Big)\times \\ &\mathbb{E}\Big(\sum_{k=1}^{m}\Big(\big(\widehat{f}(X_{t_k}) - f(X_{t_k})\big)^2 + \big(f\big(X_{t_{k+1}}\big) - f\big(X_{t_k}\big)\big)^2\Big)\Big).\end{aligned} \tag{7}$$

Intuitively, these terms are related to the estimation error arising in isotonic regression (first term) and a term that captures the variance of the function values (second term). When the function f is bounded, we show that the dominant term is the isotonic estimation error which is on the order of $m^{1/3}$. Putting these pieces together we obtain the theorem. ■

R^2 with k nearest neighbors. We note that R^2 using k-NN instead of 1-NN in the output step leads to the same rate as in (5), for a constant k. To see this, we can decompose the error (6) as

$$\mathbb{E}\left[(\widehat{f}(X) - f(X))^2\right] \leq 2k\sum_{i=1}^{k}\mathbb{E}\left[(\widehat{f}(X_{\alpha_i}) - f(X_{\alpha_i}))^2\right] + 2k\sum_{i=1}^{k}\mathbb{E}\left[(f(X_{\alpha_i}) - f(X))^2\right].$$

Following the same steps as in the analysis of the 1-NN based estimator, we can bound the second term to be at most $O(kn^{-2s/d})$, and the first term to be at most $O(km^{-2/3})$.

2.2.2. Lower bounds with Ordinal Data

To understand the fundamental limits on the usefulness of ordinal information, as well as to study the optimality of the R^2 algorithm we now turn our attention to establishing lower bounds on the minimax risk. In our lower bounds we choose $\mathbb{P}_X$ to be uniform on $[0,1]^d$. Our estimators $\widehat{f}$ are functions of the labeled samples: $\{(X_{t_1}, y_{t_1}), \ldots, (X_{t_m}, y_{t_m})\}$, the set $\mathcal{U} = \{X_1, \ldots, X_n\}$ and the true ranking π. We have the following result:

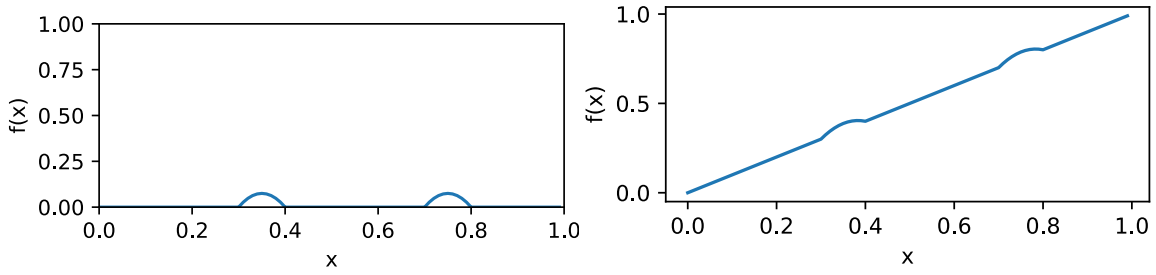


Figure 2: Original construction for nonparametric regression in 1-d (left), and our construction (right).

Theorem 2 *Let $\widehat{f}$ be any estimator with access to m labels and a (perfect) ranking on $n \geq m$ samples. For a universal constant $C > 0$,*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{s,L}} \mathbb{E}\left[(f(X) - \widehat{f}(X))^2\right] \geq C(m^{-2/3} + n^{-2s/d}).$$

Comparing with the result in Theorem 1 we conclude that the R^2 algorithm is optimal up to log factors, when the ranking is noiseless.

Proof Sketch We establish each term in the lower bound separately. Intuitively, for the $n^{-2s/d}$ lower bound we consider the case when all the n points are labeled perfectly (in which case the ranking is redundant) and show that even in this setting the MSE of any estimator is at least $n^{-2s/d}$ due to the finite resolution of the sample.

To prove the $m^{-2/3}$ lower bound we construct a novel packing set of functions in the class $\mathcal{F}_{s,L}$, and use information-theoretic techniques (Fano's inequality) to establish the lower bound. The functions we construct are all increasing functions, and as a result the ranking π provides no additional information for these functions, easing the analysis. Figure 2 contrasts the classical construction for lower bounds in nonparametric regression (where tiny bumps are introduced to a reference function) with our construction where we additionally ensure the perturbed functions are all increasing. To complete the proof, we provide bounds on the cardinality of the packing set we create, as well as bounds on the Kullback-Leibler divergence between the induced distributions on the labeled samples. We provide the technical details in Section B.2.

■

2.3. Nonparametric Regression using Noisy Ranking

In this section, we study the setting where the ordinal information is noisy. We focus here on the setting where as in equation (2) we obtain a ranking $\widehat{\pi}$ whose Kendall-Tau distance from the true ranking π is at most νn^2. We show that the R^2 algorithm is quite robust to ranking errors and achieves an MSE of $\widetilde{O}(m^{-2/3} + \sqrt{\nu} + n^{-2s/d})$. We establish a complementary lower bound of $\widetilde{O}(m^{-2/3} + \nu^2 + n^{-2s/d})$ in Section 2.3.2.

2.3.1. Upper Bounds for the R^2 Algorithm

We characterize the robustness of R^2 to ranking errors, i.e. when $\widehat{\pi}$ satisfies the condition in (2), in the following theorem, which includes Theorem 1 as a special case:

Theorem 3 *Suppose the ranking $\widehat{\pi}$ has an error of at most ν. If the target function f belongs to the Hölder ball $\mathcal{F}_{s,L}$ with $0 < s \leq 1$ and constant L, there exist constants $C_1, C_2 > 0$ such that the MSE of the R^2 estimate $\widehat{f}$ is bounded by,*

$$\mathbb{E}[(\widehat{f}(X) - f(X))^2] \leq C_1 \left(\log^2 n \log m \left(m^{-2/3} + \sqrt{\nu}\right)\right) + C_2 n^{-2s/d}.$$

Remarks: (1) Once again we observe that in the regime where sufficient ordinal information is available, i.e. when n is large, the rate no longer has an exponential dependence on the dimension d.
(2) This result also shows that the R^2 algorithm is inherently robust to noise in the ranking, and the mean squared error degrades gracefully as a function of the noise parameter ν. We investigate the optimality of the $\sqrt{\nu}$-dependence in the next section.

We now turn our attention to the proof of this result.

Proof Sketch When using an estimated permutation $\widehat{\pi}$ the true function of interest f is no longer an increasing (isotonic) function with respect to $\widehat{\pi}$, and this results in a model-misspecification *bias*. The core technical novelty of our proof is in relating the upper bound on the error in $\widehat{\pi}$ to an upper bound on this bias. Concretely, in the Appendix we show the following lemma:

Lemma 4 *For any permutation $\widehat{\pi}$ satisfying the condition in* (2)*,*

$$\sum_{i=1}^{n} (f(X_{\pi^{-1}(i)}) - f(X_{\widehat{\pi}^{-1}(i)}))^2 \leq 8M^2\sqrt{2\nu}n.$$

Using this result we bound the minimal error of approximating an increasing sequence according to π by an increasing sequence according to the estimated (and misspecified) ranking $\widehat{\pi}$. We denote this error on m labeled points by Δ, and using Lemma 4 we show that in expectation (over the random choice of the labeled set)

$$\mathbb{E}[\Delta] \leq 8M^2\sqrt{2\nu}m.$$

With this technical result in place we follow the same decomposition and subsequent steps before we arrive at the expression in equation (7). In this case, the first term for some constant $C > 0$ is bounded as:

$$\mathbb{E}\Big(\sum_{k=1}^{m} \big(\widehat{f}(X_{t_k}) - f(X_{t_k})\big)^2\Big) \leq 2\mathbb{E}[\Delta] + Cm^{1/3},$$

where the first term corresponds to the model-misspecification bias and the second corresponds to the usual isotonic regression rate. Putting these terms together in the decomposition in (7) we obtain the theorem.

■

In settings where ν is large R^2 can be led astray by the ordinal information, and a standard nonparametric regressor can achieve the (possibly faster) $O\big(m^{-\frac{2s}{2s+d}}\big)$ rate by ignoring the ordinal information. In this case, a simple and standard cross-validation procedure can combine the benefits of both methods: we estimate the regression function twice, once using R^2 and once using k nearest neighbors, and choose the regression function that performs better on a held-out validation set. The following theorem shows guarantees for this method and an upper bound for the minimax risk (1):

Theorem 5 *Under the same assumptions as Theorem 3, there exists an estimator $\widehat{f}$ such that*

$$\mathbb{E}[(\widehat{f}(X) - f(X))^2] = \widetilde{O}\left(m^{-2/3} + \min\{\sqrt{\nu}, m^{-\frac{2s}{2s+d}}\} + n^{-2s/d}\right).$$

The main technical difficulty in analyzing the model-selection procedure we propose in this context is that a naïve analysis of the procedure, using classical tail bounds to control the deviation between the empirical risk and the population risk, results in an excess risk of $\widetilde{O}(1/\sqrt{m})$. However, this rate would overwhelm the $\widetilde{O}(m^{-2/3})$ bound that arises from isotonic regression. We instead follow a more careful analysis outlined in the work of Barron (1991) which exploits properties of the squared-loss to obtain an excess risk bound of $\widetilde{O}(1/m)$. We provide a detailed proof in the Appendix for convenience and note that in our setting, the y values are not assumed to be bounded and this necessitates some minor modifications to the original proof (Barron, 1991).

2.3.2. Lower bounds with Noisy Ordinal Data

In this section we turn our attention to lower bounds in the setting with noisy ordinal information. In particular, we construct a permutation $\widehat{\pi}$ such that for a pair (X_i, X_j) of points randomly chosen from $\mathbb{P}_X$:

$$\mathbb{P}[(\pi(i) - \pi(j))(\widehat{\pi}(i) - \widehat{\pi}(j)) < 0] \leq \nu.$$

We analyze the minimax risk of an estimator which has access to this noisy permutation $\widehat{\pi}$, in addition to the labeled and unlabeled sets (as in Section 2.2.2).

Theorem 6 *Let $\widehat{f}$ be any estimator with access to m labels and a ranking on $n \geq m$ samples, then there is a constant $C > 0$ such that,*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{s,L}, \widehat{\pi} \in \Pi(\nu)} \mathbb{E}\big(f(X) - \widehat{f}(X)\big)^2 \geq C(m^{-\frac{2}{3}} + \min\{\nu^2, m^{-\frac{2}{d+2}}\} + n^{-2s/d}).$$

Comparing this result with our result in Remark 3 following Theorem 3, our upper and lower bounds differ by the gap between $\sqrt{\nu}$ and ν^2, in the case of Lipschitz functions ($s = 1$).

Proof Sketch We focus on the dependence on ν, as the other parts are identical to Theorem 2. We construct a packing set of Lipschitz functions, and we subsequently construct a noisy comparison oracle $\widehat{\pi}$ which provides no additional information beyond the labeled samples. The construction of our packing set is inspired by the construction of standard lower bounds in nonparametric regression (see Figure 2), but we modify this construction to ensure that

$\widehat{\pi}$ is uninformative. In the classical construction we divide $[0,1]^d$ into u^d grid points, with $u = m^{1/(d+2)}$ and add a "bump" at a carefully chosen subset of the grid points. Here we instead divide $[0,t]^d$ into a grid with u^d points, and add an increasing function along the first dimension, where t is a parameter we choose in the sequel.

We now describe the ranking oracle which generates the permutation $\widehat{\pi}$: we simply rank sample points according to their first coordinate. This comparison oracle only makes an error when both x, x' lies in $[0,t]^d$, and both x_1, x_1' lie in the same grid segment $[tk/u, t(k+1)/u]$ for some $k \in [u]$. So the Kendall-Tau error of the comparison oracle is $(t^d)^2 \times ((1/u)^2 \times u) = ut^{2d}$. We choose t such that this value is less than ν. Once again we complete the proof by lower bounding the cardinality of the packing-set for our stated choice of t, upper bounding the Kullback-Leibler divergence between the induced distributions and appealing to Fano's inequality. ■

2.4. Regression with Noisy Pairwise Comparisons

In this section we focus on the setting where the ordinal information is obtained in the form of noisy pairwise comparisons, following equation (3). We investigate a natural strategy of aggregating the pairwise comparisons to form a consensus ranking $\widehat{\pi}$ and then applying the R^2 algorithm with this estimated ranking. We build on results from theoretical computer science, where such aggregation algorithms are studied for their connections to sorting with noisy comparators. In particular, Braverman and Mossel (2009) study noisy sorting algorithms under the noise model described in (3) and establish the following result:

Theorem 7 (Braverman and Mossel (2009)) *Let $\alpha > 0$. There exists a polynomial-time algorithm using noisy pairwise comparisons between n samples, that with probability $1 - n^{-\alpha}$, returns a ranking $\widehat{\pi}$ such that for a constant $c(\alpha, \lambda) > 0$ we have that:*

$$\sum_{i,j \in [n]} \mathbb{I}[(\pi(i) - \pi(j))(\widehat{\pi}(i) - \widehat{\pi}(j)) < 0] \leq c(\alpha, \lambda) n.$$

Furthermore, if allowed a sequential (active) choice of comparisons, the algorithm queries at most $O(n \log n)$ pairs of samples.

Combining this result with our result on the robustness of R^2 we obtain an algorithm for nonparametric regression with access to noisy pairwise comparisons with the following guarantee on its performance:

Corollary 8 *For constants $C_1, C_2 > 0$, R^2 with $\widehat{\pi}$ estimated as described above produces an estimator $\widehat{f}$ with MSE*

$$\mathbb{E}\big(\widehat{f}(X) - f(X)\big)^2 \leq C_1 m^{-2/3} \log^2 n \log m + C_2 \max\{n^{-2s/d}, n^{-1/2} \log^2 n \log m\}.$$

Remarks:

1. From a technical standpoint this result is an immediate corollary of Theorems 3 and 7, but the extension is important from a practical standpoint. The ranking error of $O(1/n)$ from the noisy sorting algorithm leads to an additional $\widetilde{O}(1/\sqrt{n})$ term in the MSE.

This error is dominated by the $n^{-2s/d}$ term if $d \geq 4s$, and in this setting the result in Corollary 8 is also optimal up to log factors (following the lower bound in Section 2.2.2).

2. We also note that the analysis in Braverman and Mossel (2009) extends in a straightforward way to a setting where only a randomly chosen subset of the pairwise comparisons are obtained.

3. Linear Regression with Comparisons

In this section we investigate another popular setting for regression, that of fitting a linear predictor to data. We show that when we have enough comparisons, it is possible to estimate a linear function even when $m \ll d$, without making *any* sparsity or structural assumptions.

For linear regression we follow a different approach when compared to the nonparametric setting we have studied thus far. By exploiting the assumption of linearity, we see that each comparison now translates to a constraint on the parameters of the linear model, and as a consequence we are able to use comparisons to obtain a good initial estimate. However, the unknown linear regression parameters are not fully identified by comparison. For instance, we observe that the two regression vectors w and $2 \times w$ induce the same comparison results for any pairs (X_1, X_2). This motivates using direct measurements to estimate a global scaling of the unknown regression parameters, i.e the norm of regression vector w^*. Essentially by using comparisons instead of direct measurements to estimate the weights, we convert the regression problem to a *classification* problem, and therefore can leverage existing algorithms and analyses from the passive/active classification literature.

We present our assumptions and notation for the linear setup in the next subsection. Then we give the algorithm along with its analysis in Section 3.2. We also present information-theoretic lower bounds on the minimax rate in Section 3.3.

3.1. Background and Problem Setup

Following some of the previous literature on estimating a linear classifier (Awasthi et al., 2014, 2016), we assume that the distribution $\mathbb{P}_X$ is isotropic and log-concave. In more detail, we assume that coordinates of X are independent, centered around 0, have covariance I_d; and that the log of the density function of X is concave. This assumption is satisfied by many standard distributions, for instance the uniform and Gaussian distributions (Lovász and Vempala, 2007). We let $B(v, r)$ denote the ball of radius r around vector v. Our goal is to estimate a linear function $f(X) = \langle w^*, X \rangle$, and for convenience we denote:

$$r^* = \|w^*\|_2 \quad \text{and} \quad v^* = \frac{w^*}{\|w^*\|_2}.$$

Similar to the nonparametric case, we suppose that we have access to two kinds of supervision using labels and comparisons respectively. We represent these using the following two oracles:

- **Label Oracle:** We assume access to a label oracle $\mathcal{O}_l$, which takes a sample $X \in \mathbb{R}^d$ and outputs a label $y = \langle w^*, X \rangle + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

- **Comparison Oracle:** We also have access to a (potentially cheaper) comparison oracle $\mathcal{O}_c$. For each query, the oracle $\mathcal{O}_c$ receives a pair of samples $(X, X') \sim \mathbb{P}_X \times \mathbb{P}_X$, and returns a random variable $Z \in \{-1, +1\}$, where $Z = 1$ indicates that the oracle believes that $f(X) > f(X')$, and $Z = -1$ otherwise. We assume that the oracle has agnostic noise[3] ν, i.e.:
$$\mathbb{P}(Z \neq \text{sign}(\langle w^*,\, X - X' \rangle)) \leq \nu.$$
That is, for a randomly sampled triplet (X, X', Z) the oracle is wrong with probability at most ν. Note that the error of the oracle for a particular example $(X, X') = (x, x')$ can be arbitrary.

Given a set of unlabeled instances $\mathcal{U} = \{X_1, X_2, \ldots\}$ drawn from $\mathbb{P}_X$, we aim to estimate w^* by querying $\mathcal{O}_l$ and $\mathcal{O}_c$ with samples in $\mathcal{U}$, using a label and comparison budget of m and n respectively. Our algorithm can be either passive, active or semi-supervised; we denote the output of algorithm $\mathcal{A}$ by $\widehat{w} = \mathcal{A}(\mathcal{U}, \mathcal{O}_l, \mathcal{O}_c)$. For an algorithm $\mathcal{A}$, we study the minimax risk (1), which in this case can be written as

$$\mathfrak{M}(m, n) = \inf_{\mathcal{A}} \sup_{w^*} \mathbb{E}\left[\langle w^* - \widehat{w},\, X \rangle^2\right]. \tag{8}$$

Our algorithm relies on a linear classification algorithm $\mathcal{A}^c$, which we assume to be a proper classification algorithm (i.e. the output of $\mathcal{A}^c$ is a linear classifier which we denote by $\widehat{v}$). We let $\mathcal{A}^c(\widetilde{\mathcal{U}}, \widetilde{\mathcal{O}}, \widetilde{m})$ denote the output (linear) classifier of $\mathcal{A}^c$ when it uses the unlabeled data pool $\widetilde{\mathcal{U}}$, the label oracle $\widetilde{\mathcal{O}}$ and acquires $\widetilde{m}$ labels. $\mathcal{A}^c$ can be either passive or active; in the former case the $\widetilde{m}$ labels are randomly selected from $\widetilde{\mathcal{U}}$, whereas in the latter case $\mathcal{A}^c$ decides which labels to query. We use $\varepsilon_{\mathcal{A}^c}(\widetilde{m}, \delta)$ to denote (an upper bound on) the 0/1 error of the algorithm $\mathcal{A}^c$ when using $\widetilde{m}$ labels, with $1 - \delta$ probability, i.e.:

$$\mathbb{P}[\text{err}(\mathcal{A}^c(\widetilde{\mathcal{U}}, \widetilde{\mathcal{O}}, \widetilde{m})) \leq \varepsilon_{\mathcal{A}^c}(\widetilde{m}, \delta)] \geq 1 - \delta.$$

We note that by leveraging the log-concave assumption on P_X, it is straightforward to translate guarantees on the 0/1 error to corresponding guarantees on the ℓ_2 error $\|\widehat{v} - v^*\|_2$.

We conclude this section by noting that the classical minimax rate for ordinary least squares (OLS) is of order $O(d/\widetilde{m})$, where $\widetilde{m}$ is the number of label queries. This rate cannot be improved by active label queries (see for instance Chaudhuri et al. (2015)).

3.2. Algorithm and Analysis

Our algorithm, Comparison Linear Regression (CLR), is described in Algorithm 2. We first use the comparisons to construct a classification problem with samples $(X - X')$ and oracle $\mathcal{O}_c$. Here we slightly overload the notation of $\mathcal{O}_c$ that $\mathcal{O}_c(X_i - X_j) = \mathcal{O}_c(X_i, X_j)$. Given these samples we use an active linear classification algorithm to estimate a normalized weight vector $\widehat{v}$. After classification, we use the estimated $\widehat{v}$ along with actual label queries to estimate the norm of the weight vector $\widehat{r}$. Combining these results we obtain our final estimate $\widehat{w} = \widehat{r} \cdot \widehat{v}$.

3. As we discuss in the sequel our model can also be adapted to the bounded noise model case of eqn. (3) using a different algorithm from active learning; see Section 5 for details.

Algorithm 2 Comparison Linear Regression (CLR)

Input: comparison budget n, label budget m, unlabeled data pool $\mathcal{U}$, algorithm $\mathcal{A}^c$.

1: Construct the pairwise pool $\mathcal{U}' = \{X_1 - X_2, X_3 - X_4, \ldots\}$ from $\mathcal{U}$.
2: Run $\mathcal{A}^c(\mathcal{U}', \mathcal{O}_c, n)$ and obtain a classifier with corresponding weight vector $\widehat{v}$.
3: Query random samples $\{(X_i, y_i)\}_{i=1}^m$.
4: Let $\widehat{r} = \dfrac{\sum_{i=1}^m \langle \widehat{v},\, X_i\rangle y_i}{\sum_{i=1}^m \langle \widehat{v},\, X_i\rangle^2}$.

Output: $\widehat{w} = \widehat{r} \cdot \widehat{v}$.

We have the following main result which relates the error of $\widehat{w}$ to the error of the classification algorithm $\mathcal{A}^c$.

Theorem 9 *There exists some constants C, M such that if $m > M$, the MSE of Algorithm 2 satisfies*

$$\mathbb{E}[\langle w^* - \widehat{w},\, X\rangle^2] \leq \widehat{O}\left(\frac{1}{m} + \log^2 m \cdot \varepsilon_{\mathcal{A}^c}\left(n, \frac{1}{m}\right) + \nu^2\right).$$

We defer a detailed proof to the Appendix and provide a concise proof sketch here.

Proof Sketch

We recall that, by leveraging properties of isotropic log-concave distributions, we can obtain an estimator $\widehat{v}$ such that $\|\widehat{v} - v^*\|_2 \leq \varepsilon = O\left(\varepsilon_{\mathcal{A}^c}(n, \delta)\right)$. Now let $T_i = \langle \widehat{v},\, X_i\rangle$, and we have

$$\widehat{r} = \frac{\sum_{i=1}^m T_i y_i}{\sum_{i=1}^m T_i^2} = r^* + \frac{\sum_{i=1}^m T_i r^* \langle v^* - \widehat{v},\, X_i\rangle + \varepsilon_i}{\sum_{i=1}^m T_i^2}.$$

And thus

$$\langle w^*,\, X\rangle - \langle \widehat{w},\, X\rangle = r^*\langle v^* - \widehat{v},\, X\rangle - \frac{\sum_{i=1}^m T_i r^* \langle v^* - \widehat{v},\, X_i\rangle}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle + \frac{\sum_{i=1}^m T_i \varepsilon_i}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle.$$

The first term can be bounded using $\|\widehat{v} - v^*\|_2 \leq \varepsilon$; for the latter two terms, using Hoeffding bounds we show that $\sum_{i=1}^m T_i^2 = O(m)$. Then by decomposing the sums in the latter two terms, we can bound the MSE. ■

Leveraging this result we can now use existing results to derive concrete corollaries for particular instantiations of the classification algorithm $\mathcal{A}^c$. For example, when $\nu = 0$ and we use passive learning, standard VC theory shows that the empirical risk minimizer has error $\varepsilon_{\text{ERM}} = O(d \log(1/\delta)/n)$. This leads to the following corollary:

Corollary 10 *Suppose that $\nu = 0$, and that we use the passive ERM classifier as $\mathcal{A}^c$ in Algorithm 2, then the output $\widehat{w}$ has MSE bounded as:*

$$\mathbb{E}[\langle w^* - \widehat{w},\, X\rangle^2] \leq \widehat{O}\left(\frac{1}{m} + \frac{d \log^3 m}{n}\right).$$

When $\nu > 0$, we can use other existing algorithms for either the passive/active case. We give a summary of existing results in Table 1, and note that (as above) each of these results can be translated in a straightforward way to a guarantee on the MSE when combining direct and comparison queries. We also note that previous results using isotropic log-concave distribution can be directly exploited in our setting since $X - X'$ follows isotropic log-concave distribution if X does (Lovász and Vempala, 2007). Each of these results provide upper bounds on minimax risk (8) under certain restrictions.

Table 1: Summary of existing results for passive/active classification for isotropic log-concave X distributions. C denotes some fixed constant; $\varepsilon = \varepsilon_{\mathcal{A}^c}(\widetilde{m}, \delta)$. Using ERM does not need $\mathbb{P}_X$ to be log-concave; Yan and Zhang (2017) additionally requires that $X - X'$ follows a uniform distribution. Here "Efficient" means running time polynomial in $\widetilde{m}, 1/\varepsilon, d, 1/\delta$.

Algorithm	Oracle	Requirement	Rate of ε	Efficient?
ERM (Hanneke, 2009)	Passive	None	$\widetilde{O}\left(d\sqrt{\frac{1}{\widetilde{m}}}\right)$	No
		$\nu = O(\varepsilon)$	$\widetilde{O}\left(\frac{d}{\widetilde{m}}\right)$	
Awasthi et al. (2016)	Passive	$\nu = O(\varepsilon)$	$\widetilde{O}\left(\left(\frac{d}{\widetilde{m}}\right)^{1/3}\right)$	Yes
Awasthi et al. (2014)	Active	$\nu = O(\varepsilon)$	$\exp\left(-\frac{C\widetilde{m}}{d+\log(\widetilde{m}/\delta)}\right)$	Yes
Yan and Zhang (2017)	Passive	$\nu = O\left(\frac{\varepsilon}{\log d + \log\log\frac{1}{\varepsilon}}\right)$	$\widetilde{O}\left(\frac{d}{m}\right)$	Yes
	Active		$\exp\left(-\frac{C\widetilde{m}}{d+\log(\widetilde{m}/\delta)}\right)$	

3.3. Lower Bounds

Now we turn to information-theoretic lower bounds on the minimax risk (1). We consider *any* active estimator $\widehat{w}$ with access to the two oracles $\mathcal{O}_c, \mathcal{O}_l$, using n comparisons and m labels. In this section, we show that the $1/m$ rate in Theorem 9 is optimal; we also show a lower bound in the active setting on the total number of queries in the appendix.

Theorem 11 *Suppose that X is uniform on $[-1,1]^d$, and $\varepsilon \sim \mathcal{N}(0,1)$. Then, for any (active) estimator $\widehat{w}$ with access to m labels and unlimited amount of comparisons, there is a universal constant $c > 0$ such that,*

$$\inf_{\widehat{w}} \sup_{w^*} \mathbb{E}\left[\langle w^* - \widehat{w},\, X\rangle^2\right] \geq \frac{c}{m}.$$

Theorem 11 shows that the $O(1/m)$ term in Theorem 9 is necessary. The proof uses classical information-theoretic techniques (Le Cam's method) applied to two increasing functions with $d = 1$, and is included in Appendix B.7.

We note that we cannot establish lower bounds that depend solely on number of comparisons n, since we can of course achieve $O(d/m)$ MSE without using any comparisons. Consequently, we show a lower bound on the *total* number of queries. This bound shows that when using the algorithm in the paper of Awasthi et al. (2014) as $\mathcal{A}^c$ in CLR, the total number of queries is optimal up to log factors.

Theorem 12 *Suppose that X is uniform on $[-1,1]^d$, and $\varepsilon \sim \mathcal{N}(0,1)$. For any (active) estimator $\widehat{w}$ with access to m labels and n comparisons, there exists a ground truth weight $\widetilde{w}$ and a global constant C, such that when $w^* = \widetilde{w}$ and $2n + m < d$,*

$$\mathbb{E}\left[\langle \widehat{w} - w^*, X\rangle^2\right] \geq C.$$

Theorem 12 shows a lower bound on the total number of queries in order to get low error. Combining with Theorem 11, in order to get a MSE of γ for some $\gamma < C$, we need to make at least $O(1/\gamma + d)$ queries (i.e., labels+comparisons). Note that for the upper bound in Theorem 9, we need $m + n = \widehat{O}(1/\gamma + d\log(d/\gamma))$ for Algorithm 2 to reach γ MSE, when using Awasthi et al. (2014) as $\mathcal{A}^c$ (see Table 1). So Algorithm 2 is optimal in terms of total queries, up to log factors.

The proof of Theorem 12 is done by considering an estimator with access to $m + 2n$ *noiseless* labels $\{(x_i, w^* \cdot x_i)\}_{i=1}^{m+2n}$, which can be used to generate n comparisons and m labels. We sample w^* from a prior distribution in $B(0,1)$, and show that the expectation of MSE in this case is at least a constant. Thus there exists a weight vector $\widetilde{w}$ that leads to constant error.

4. Experiments

To verify our theoretical results and test our algorithms in practice, we perform three sets of experiments. First, we use simulated data, where the noise in the labels and ranking can be controlled separately. Second, we consider an application of predicting people's ages from photographs, where we synthesize comparisons from data on people's apparent ages (as opposed to their true biological ages). Finally, we crowdsource data using Amazon's Mechanical Turk to obtain comparisons and direct measurements for the task of estimating rental property asking prices. We then evaluate various algorithms for predicting rental property prices, both in terms of their accuracy, as well as in terms of their time cost.

Baselines. In the nonparametric setting, we compare R^2 with k-NN algorithms in all experiments. We choose k-NN methods because they are near-optimal theoretically for Lipschitz functions, and are widely used in practice. Also, the R^2 method is a nonparametric method built on a nearest neighbor regressor. It may be possible to use the ranking-regression method in conjunction with other nonparametric regressors but we leave this for future work. We choose from a range of different constant values of k in our experiments.

In the linear regression setting, for our CLR algorithm, we consider both a passive and an active setting for comparison queries. For the passive comparison query setting, we simply use a Support Vector Machine (SVM) as $\mathcal{A}$ in Algorithm 2. For the active comparison query setting, we use an algorithm minimizing hinge loss as described in Awasthi et al. (2014). We compare CLR to the LASSO and to Linear Support Vector Regression (LinearSVR), where the relevant hyperparameters are chosen based on validation set performance. We choose LASSO and LinearSVR as they are the most prevalent linear regression methods. Unless otherwise noted, we repeat each experiment 20 times and report the average MSE[4].

Cost Ratio. Our algorithms aim at reducing the overall cost of estimating a regression function when comparisons can be more easily available than direct labels. In practice, the

4. Our plots are best viewed in color.

cost of obtaining comparisons can vary greatly depending on the task, and we consider two practical setups:

1. In many applications, both direct labels and comparisons can be obtained, but labels cost more than comparisons. Our price estimation task corresponds to this case. The cost, in this case, depends on the ratio between the cost of comparisons and labels. We suppose that comparisons cost 1, and that labels cost c for some constant $c > 1$ and that we have a total budget of C. We call c the *cost ratio*. Minimizing the risk of our algorithms requires minimizing $\mathfrak{M}(m, n; \eta)$ as defined in (1) subject to $cm + n \leq C$; for most cases, we need a small m and large n. In experiments with a finite cost ratio, we fix the number of direct measurements to be a small constant m and vary the number of comparisons that we use.

2. Direct labels might be substantially harder to acquire because of privacy issues or because of inherent restrictions in the data collection process, whereas comparisons are easier to obtain. Our age prediction task corresponds to this case, where it is conceivable that only some of the biological ages are available due to privacy issues. In this setting, the cost is dominated by the cost of the direct labels and we measure the cost of estimation by the number of labeled samples used.

4.1. Modifications to Our Algorithms

While R^2 and CLR are near optimal from a theoretical standpoint, we adopt the following techniques to improve their empirical performance:

R^2 with k-NN. Our analysis considers the case when we use 1-NN after isotonic regression. However, we empirically find that using more than 1 nearest neighbor can also improve the performance. So in our experiments, we use k-NN in the final step of R^2, where k is a small fixed constant. We note in passing that our theoretical results remain valid in this slightly more general setting.

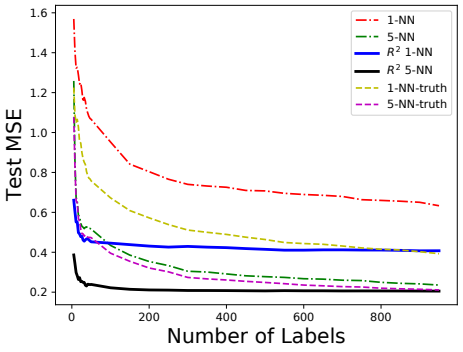
R^2 with comparisons. When R^2 uses passively collected comparisons, we would need $O(n^2)$ pairs to have a ranking with $O(1/n)$ error in the Kendall-Tau metric if we use the algorithm from Braverman and Mossel (2009). We instead choose to take advantage of the feature space structure when we use R^2 with comparisons. Specifically, we build a nonparametric rankSVM (Joachims, 2002) to score each sample using pairwise comparisons. We then rank samples according to their scores given by the rankSVM. We discuss another potential method, which uses nearest neighbors based on Borda counts, in Appendix A.

CLR with feature augmentation. Using the directly labeled data only to estimate the norm of the weights corresponds to using linear regression with the direct labels with a *single* feature $\langle \widehat{v}, x \rangle$ from Algorithm 2. Empirically, we find that using all the features together with the estimated $\langle \widehat{v}, x \rangle$ results in better performance. Concretely, we use a linear SVR with input features $(x; \langle \widehat{v}, x \rangle)$, and use the output as our prediction.

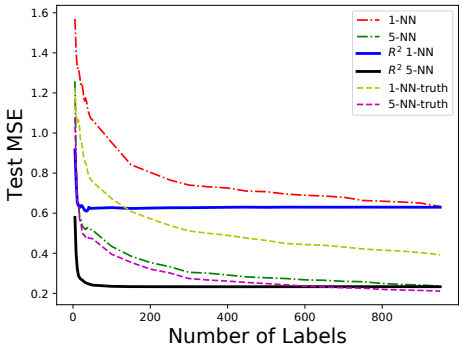
4.2. Simulated Data

We generate different synthetic datasets for nonparametric and linear regression settings in order to verify our theory.

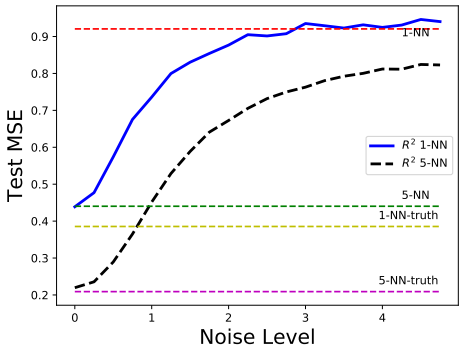
4.2.1. Simulated Data for R^2



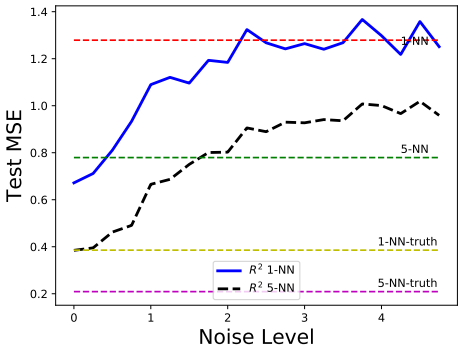
(a) Perfect ranking



(b) Noisy ranking



(c) Effect of ranking noise, $m = 100$



(d) Effect of ranking noise, $m = 10$

Figure 3: Experiments on simulated data for R^2. 1-NN and 5-NN represent algorithms using noisy label data only; R^2 1-NN and R^2 5-NN use noisy labels as well as rankings; 1-NN-truth and 5-NN-truth use perfect label data only.

Data Generation. We generate simulated data following Härdle et al. (2012). We take $d = 8$, and sample X uniformly random from $[0, 1]^d$. Our target function is $f(x) = \sum_{i=1}^{d} f^{(i \bmod 4)}(x_i)$, where x_i is x's i-th dimension, and

$$f^{(1)}(x) = px - 1/2, \qquad f^{(2)}(x) = px^3 - 1/3,$$
$$f^{(3)}(x) = -2\sin(-px), \qquad f^{(4)}(x) = e^{-px} + e^{-1} - 1$$

with p sampled uniformly at random in $[0, 10]$. We generate a training and a test set each of size $n = 1000$ samples respectively. We rescale f so that it has 0 mean and unit variance over the training set. This makes it easy to control the noise that we add relative to the function value. For training data, we generate the labels as $y_i = f(X_i) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 0.5^2)$, for the training data $\{X_1, \ldots, X_n\}$. At test time, we compute the MSE $\frac{1}{n}\sum_{i=1}^{n}(f(X_i^{\text{test}}) - \widehat{f}(X_i^{\text{test}}))^2$ for the test data $\{X_1^{\text{test}}, \ldots, X_n^{\text{test}}\}$.

Setup and Baselines. We test R^2 with either 1-NN or 5-NN for our simulated data, denoted as R^2 1-NN and R^2 5-NN respectively. We compare them with the following baselines: i) 1-NN and 5-NN using noisy labels (X, y) only. Since R^2 uses ordinal data in addition to labels, it should have lower MSE than 1-NN and 5-NN. ii) 1-NN and 5-NN using perfect labels $(X, f(X))$. Since these algorithms use perfect labels, when all sample points are labeled they serve as a benchmark for our algorithms.

R^2 with rankings. Our first set of experiments suppose that R^2 has access to the ranking over all 1000 training samples, while the k-NN baseline algorithms only have access to the labeled samples. We measure the cost as the number of labels in this case. Figure 3a compares R^2 with baselines when the ranking is perfect. R^2 1-NN and R^2 5-NN exhibited better performance than their counterparts using only labels, whether using noisy or perfect labels; in fact, we find that R^2 1-NN and R^2 5-NN perform nearly as well as 1-NN or 5-NN using all 1000 perfect labels, while only requiring around 50 labeled samples.

In Figure 3b, R^2 uses an input ranking obtained from noisy labels $\{y_1, ..., y_n\}$. In this case, the noise in the ranking is dependent on the label noise, since the ranking is directly derived from the noisy labels. This makes the obtained labels consistent with the ranking, and thus eliminates the need for isotonic regression in Algorithm 1. Nevertheless, we find that the ranking still provides useful information for the unlabeled samples. In this setting, R^2 outperformed its 1-NN and 5-NN counterparts using noisy labels. However, R^2 was outperformed by algorithms using perfect labels when $n = m$. As expected, R^2 and k-NN with noisy labels achieved identical MSE when $n = m$, as ranking noise is derived from noise in labels.

We consider the effect of *independent* ranking noise in Figure 3c. We fixed the number of labeled/ranked samples to 100/1000 and varied the noise level of ranking. For a noise level of σ, the ranking is generated from

$$y' = f(X) + \varepsilon' \tag{9}$$

where $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$. We also plot the performance of 1-NN and 5-NN using 100 noisy labels and 1,000 perfect labels for comparison. We varied σ from 0 to 5 and plotted the MSE. We repeat these experiments 50 times.

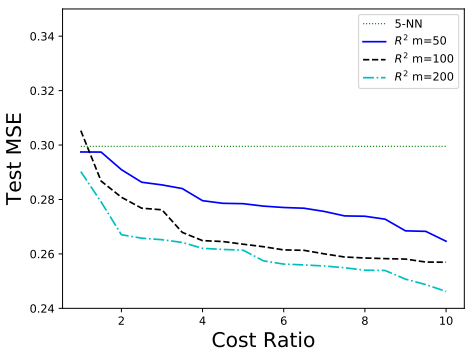
For both R^2 1-NN and 5-NN – despite the fact that they use noisy labels – their performance is close to the NN methods using noiseless labels. As σ' increases, both methods start to deteriorate, with R^2 5-NN hitting the naive 5-NN method at around $\sigma' = 1$ and R^2 1-NN at around $\sigma' = 2.5$. This shows that R^2 is robust to ranking noise of comparable magnitude to the label noise. We show in Figure 3d the curve when we use 10 labels and 1000 ranked samples, where a larger amount of ranking noise can be tolerated[5].

R^2 with comparisons. We also investigate the performance of R^2 when we have pairwise comparisons instead of a total ranking. We train a rankSVM with an RBF kernel with a bandwidth of 1, i.e. $k(x, x') = \exp(-\|x - x'\|_2^2)$, as described in Section 4.1. Our rankSVM has a ranking error of $\nu = 11.8\%$ on the training set and $\nu = 13.8\%$ on the validation set. For simplicity, we only compare with 5-NN here since it gives best performance amongst the label-only algorithms. The results are depicted in Figure 4. When comparisons are perfect, we first investigate the effect of the cost ratio in Figure 4a. We fixed the budget to

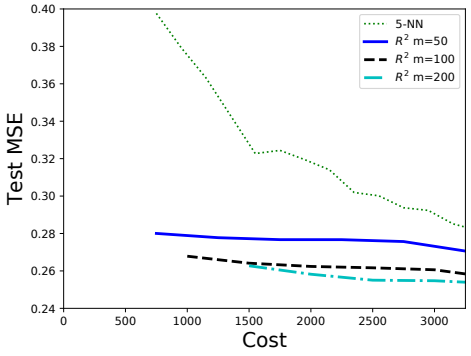
5. For this set of experiments, we repeat them 50 times to obtain stable results.

equal to $C = 500c$ (i.e., we would have 500 labels available if we only used labels), and each curve corresponds to a value of $m \in \{50, 100, 200\}$ and a varied n such that the total cost is $C = 500c$. We can see for almost all choices of m and cost ratio, R^2 provides a performance boost. In Figure 4b we fix $c = 5$, and vary the total budget C from 500 to 4,000. We find that R^2 outperforms the label-only algorithms in most setups.

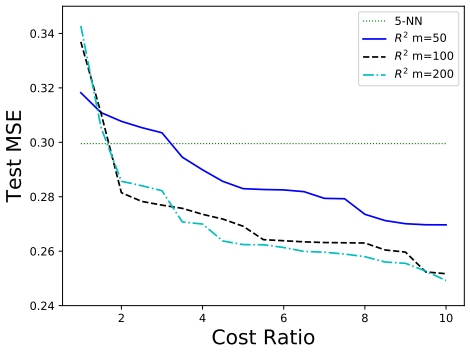
In Figures 4c and 4d, we consider the same setup of experiments, but with comparisons generated from (9), where $\varepsilon' \sim \mathcal{N}(0, 0.5^2)$. Note that here the noise ε' is of the same magnitude but independent from the label noise. Although R^2 gave a less significant performance boost in this case, it still outperformed label-only algorithms when $c \geq 2$.



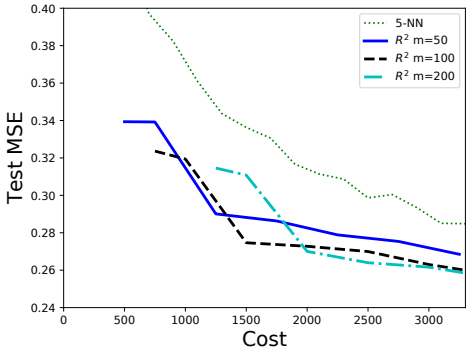
(a) Perfect comparisons, $C = 500c$



(b) Perfect comparisons, $c = 5$



(c) Noisy comparisons, $C = 500c$

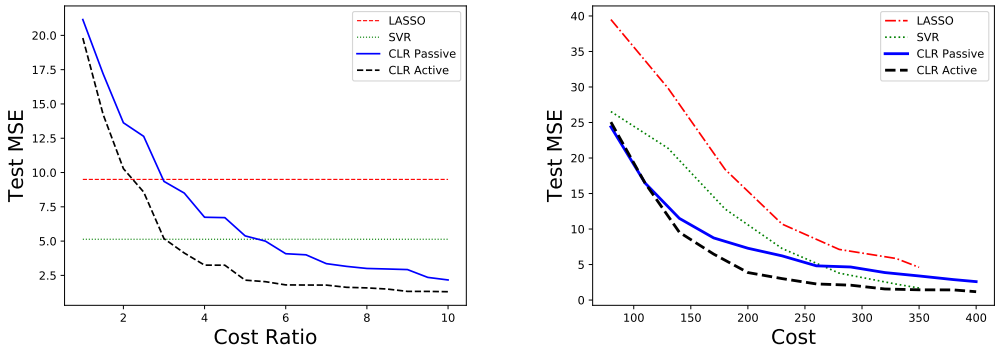


(d) Noisy comparisons, $c = 5$

Figure 4: Experimental results on synthetic dataset for R^2 with comparisons. Each curve corresponds to a fixed m, and we vary n as the cost ratios or total budget change. Note that curves start at different locations because of different m values.

4.2.2. Simulated Data for CLR

For CLR we only consider the case with a cost ratio, because we find that a small number of comparisons already suffice to obtain a low error. We set $d = 50$ and generate both X and w^* from the standard normal distribution $\mathcal{N}(0, I_d)$. We generate noisy labels with



(a) Varying cost ratio c with $C = 50c$ (b) Varying budget C with $c = 5$

Figure 5: Experimental results on synthetic dataset for CLR.

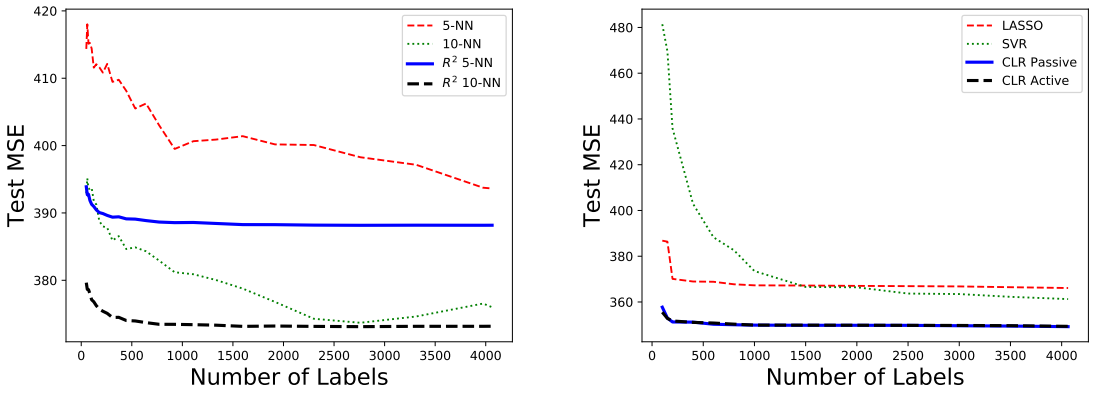
distribution $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. The comparison oracle generates response using the same noise model: $Z = \text{sign}((\langle w^*, x_1\rangle + \varepsilon_1) - (\langle w^*, x_2\rangle - \varepsilon_2))$ for input (x_1, x_2), with $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, 0.5^2)$ independent of the label noise.

Model performances are compared in Figure 5. We first investigate the effect of cost ratio in Figure 5a, where we fixed the budget to $C = 50c$, i.e. if we only used labels we would have a budget of 50 labels. Both passive and active versions of CLR use 10 labeled samples and the remaining budget to obtain comparisons. The passive comparison query version of CLR requires roughly $c > 5$ to outperform baselines, and the active comparison query version requires $c > 3$. We also experiment with a fixed cost ratio $c = 5$ and varying budget C in Figure 5b (and we retain $m = 10$ for CLR). The active version outperformed all baselines in most scenarios, whereas the passive version gave a performance boost when the budget was less than 250 (i.e. number of labels is restricted to less than 250 in label only setting). We note that the active and passive versions of CLR only differ in their collection of comparisons; both algorithms (along with the baselines) are given a random set of labeled samples, making it a fair competition.

4.3. Predicting Ages from Photographs

To further validate R^2 in practice, we consider the task of estimating people's ages from photographs. We use the APPA-REAL dataset (Agustsson et al., 2017), which contains 7,591 images, and each image is associated with a biological age and an apparent age. The biological age is the person's actual age, whereas the apparent ages are collected by asking crowdsourced workers to estimate their apparent ages. Estimates from (on average 38 different) labelers are averaged to obtain the apparent age. APPA-REAL also provides the standard deviation of the apparent age estimates. The images are divided into 4,063 train, 1,488 validation and 1,962 test samples, and we choose the best hyperparameters using the validation samples.

Task. We consider the task of predicting biological age. Direct labels come from biological age, whereas the ranking is based on apparent ages. This is motivated by the collection process of many modern datasets: for example, we may have the truthful biological age only



(a) R^2 (b) Linear Regression

Figure 6: Experimental results on age prediction.

for a fraction of samples, but wish to collect more through crowdsourcing. In crowdsourcing, people give comparisons based on apparent age instead of biological age. As a consequence, in our experiments we assume additional access to a ranking that comes from the apparent ages. Since collecting crowdsourced data can be much easier than collecting the real biological ages, as discussed earlier, we define the cost as the number of direct labels used in this experiment.

Features and Models. We extract a 128-dimensional feature vector for each image using the last layer of FaceNet (Schroff et al., 2015). We rescale the features so that every $X \in [0,1]^d$ for R^2, or we centralize the feature to have zero mean and unit variance for CLR. We use 5-NN and 10-NN to compare with R^2 in this experiment. Utilizing extra ordinal information, R^2 has additional access to the ranking of apparent ages; since CLR does not use a ranking directly we provide it access to 4,063 comparisons (the same size as the training set) based on apparent ages.

Our results are depicted in Figure 6. The 10-NN version of R^2 gave the best overall performance amongst nonparametric methods. R^2 5-NN and R^2 10-NN both outperformed other algorithms when the number of labeled samples was less than 500. Interestingly, we observe that there is a gap between R^2 and its nearest neighbor counterparts even when $n = m$, i.e. the ordinal information continues to be useful even when all samples are labeled; this might be because the biological ages are "noisy" in the sense that they are also determined by factors not present in image (e.g., lifestyle). Similarly, for linear regression methods, our method also gets the lowest overall error for any budget of direct labels. In this case, we notice that active comparisons only have a small advantage over passive comparisons, since the comparison classifiers both converge with a sufficient number of comparisons.

4.4. Estimating AirBnB Listing Prices

In our third set of experiments, we consider the cost of both comparisons and direct labels. We use data of AirBnB listings and ask Amazon Mechanical Turk (AMT) workers to estimate

their price. To measure the cost, we collect not only the labels and comparisons but also the time taken to answer the question. We use the time as an estimate of the cost.

Data Collection. Our data comes from Kaggle[6] and contains information about AirBnB postings in Seattle. We use 358 listings as the training set and 92 as the test set. We additionally pick 50 listings from the training set as validation data to select hyperparameters. We use the following features for our experiments:

(1) Host response rate, (2) Host acceptance rate, (3) Host listings count,
(4) Number of reviews per month, (5) Number of bathrooms, (6) Number of bedrooms,
(7) Number of beds, (8) Number of reviews, (9) Review scores rating,
(10) Number of people the house accommodates, (11) Bed type,
(12) Property type, (13) Room type, (14) Cancellation policy,
(15) Whether host is a super host, (16) Zipcode of property.

We transform the categorial values (11-16) to binary values to make the problem a regression task.

Each worker from AMT is asked to do either of the following two tasks: i) given the description of an AirBnB listing, estimate the per-night price on a regular day in 2018; ii) given the description of two AirBnB listings, select the listing that has a higher price. We collect 5 direct labels for each data point in the training set and 9 labels for each data point in the test set. For comparisons, we randomly draw 1,841 pairs from the training set and ask 2 workers to compare their prices. We release our data at `https://github.com/xycforgithub/Airbnb-Comparison`.

Tasks. We consider two tasks, motivated by real-world applications.

1. In the first task, the goal is to predict the real listing price. This is motivated by the case where collecting the real price might be difficult to obtain or involves privacy issues. We assume that our training data, including both labels and comparisons, comes from AMT workers.

2. In the second task, the goal is to predict the user-estimated price. This can be of particular interest to AirBnB company and house-owners, for deciding the best price of the listing. We do not use the real prices in this case; we use the average of 9 worker estimated prices for each listing in the test set as the ground truth label, and the training data also comes from our AMT results.

Raw Data Analysis. Before we proceed to the regression task, we analyze the workers' performance for both tasks based on raw data in Table 2. Our first goal is to compare a pairwise comparison to an induced comparison, where the induced comparison is obtained by making two consecutive direct label queries and subtracting them. Similar to Shah et al. (2016a), we observe that comparisons are more accurate than the induced comparisons.

We first convert labels into pairwise comparisons by comparing individual direct labels: namely, for each obtained labeled sample pair $(x_1, y_1), (x_2, y_2)$ where y_1, y_2 are the *raw* labels from workers, we create a pairwise comparison that corresponds to comparing (x_1, x_2) with label being $\text{sign}(y_1 - y_2)$. We then compute the error rate of raw and label-induced

6. `https://www.kaggle.com/AirBnB/seattle/home`

comparisons for both Task 1 and 2. For Task 1, we directly compute the error rate w.r.t. the true listing price. For Task 2, we do not have the ground truth user prices; we instead follow the approach of Shah et al. (2016a) to compute the fraction of disagreement between comparisons. Namely, in either the raw or label-induced setting, for every pair of samples (x_i, x_j) we compute the majority of labels z_{ij} based on all comparisons on (x_i, x_j). The disagreement on (x_i, x_j) is computed as the fraction of comparisons that disagrees with z_{ij}, and we compute the overall disagreement by averaging over all possible (x_i, x_j) pairs.

If an ordinal query is equivalent to two consecutive direct queries and subtracting the labels, we would expect a similar accuracy/disagreement for the two kinds of comparisons. However our results in Table 2 show that this is not the case: direct comparison queries have better accuracy for Task 1, as well as a lower disagreement within collected labels. This shows that a comparison query cannot be replaced by two consecutive direct queries. We do not observe a large difference in the average time to complete a query in Table 2; however the utility of comparisons in predicting price can be higher since they yield information about two labels.

Performance	Comparisons	Labels
Task 1 Error	**31.3%**	41.3%
Task 2 Disagreement	**16.4%**	29.5%
Average Time	64s	**63s**

Table 2: Performance of comparisons versus labels for both tasks.

In Figure 7, we look into the relation between true and user estimated price. Here we show a scatter plot of the true listing prices of AirBnB data with respect to the user estimated prices. Although the true prices are linearly correlated with the user prices (with p-value $6e-20$), the user price is still very different from true price even if we take the average of 5 labelers. The average of all listings' true price is higher than the average of all user prices by 25 dollars, partially explaining the much higher error when we use user prices to estimate true prices.

Results. We plot the experimental results in Figure 8. For nonparametric regression, R^2 had a significant performance boost over the best nearest neighbor regressor under the same total worker time, especially for Task 1. For Task 2, we observe a smaller improvement, but R^2 is still better than pure NN methods for all total time costs.

For linear regression, we find that the performance of CLR varies greatly with m (number of labels), whereas its performance does not vary as significantly with the number of comparisons. In fact, the errors of both CLR passive and active already plateau with a mere 50 comparisons, since the dimension of data is small ($d = 10$). So deviating from our previous experiments, in this setting, we vary the number of labels in Figure 8c and 8d. As in the nonparametric case, CLR also outperforms the baselines in both tasks. For Task 1, the active and passive versions of CLR perform similarly, whereas active queries lead to a moderate performance boost on Task 2. This is probably because the error on Task 2 is much lower than that on Task 1 (see Table 2), and active learning typically has an advantage over passive learning when the noise is not too high.

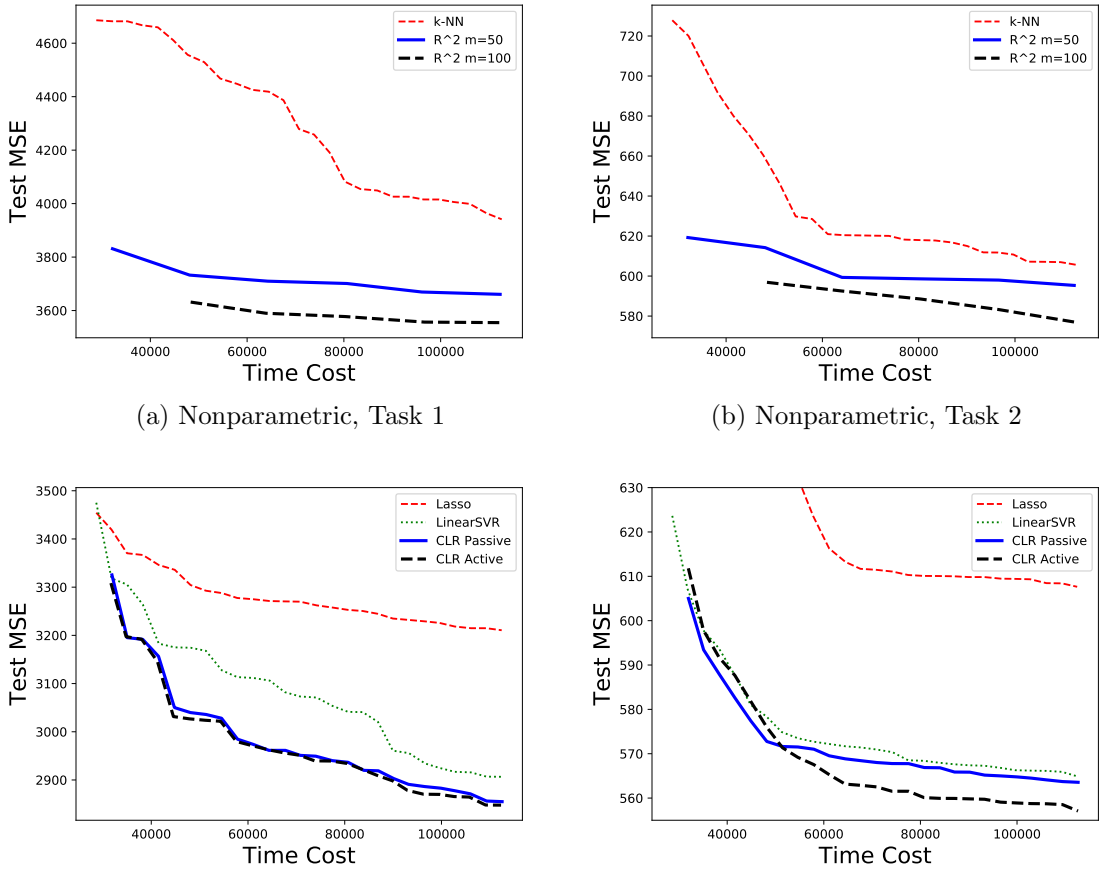
Figure 7: Scatter plot of true prices versus average user estimated prices, along with the ordinary least square result. We only show prices smaller than 200 to make the relation clearer.

5. Discussion

We design (near) minimax-optimal algorithms for nonparametric and linear regression using additional ordinal information. In settings where large amounts of ordinal information are available, we find that limited direct supervision suffices to obtain accurate estimates. We provide complementary minimax lower bounds, and illustrate our proposed algorithms on real and simulated datasets. Since ordinal information is typically easier to obtain than direct labels, one might expect in these favorable settings the R^2 algorithm to have lower effective cost than an algorithm based purely on direct supervision.

Several directions exist for future work. On the nonparametric regression side, it remains to extend our results to the case where the Hölder exponent $s > 1$. In this setting the optimal rate $O\left(m^{-\frac{2s}{2s+d}}\right)$ can be faster than the convergence rate of isotonic regression, which can make our algorithm sub-optimal. Furthermore, using nearest neighbor typically only leads to optimal rates when $0 < s \leq 1$. In addition, it is important to address the setting where both direct and ordinal supervision are actively acquired. For linear regression, an open problem is to consider the bounded noise model for comparisons. Our results can be easily extended to the bounded noise case using the algorithm in Hanneke (2009), however that algorithm is computationally inefficient. The best efficient active learning algorithm in this bounded noise setting (Awasthi et al., 2016) requires $m \geq O\left(d^{O(\frac{1}{(1-2\lambda)^4})}\right)$ comparisons, and a large gap remains between what can be achieved in the presence and absence of computational constraints.

Motivated by practical applications in crowdsourcing, we list some further extensions to our results:



(a) Nonparametric, Task 1

(b) Nonparametric, Task 2

(c) Linear, Task 1

(d) Linear, Task 2

Figure 8: Results for AirBnB price estimation. In (a)(b), each curve has a fixed m with varied n; for (c)(d), each curve uses only 50 comparisons with a varied number of labels. For Figure (d), LASSO performs much worse than a LinearSVR and we only show part of the curve.

Partial orders: In this paper, we focus on ordinal information either in the form of a total ranking or pairwise comparisons. In practice, ordinal information might come in the form of partial orders, where we have several subsets of unlabeled data ranked, but the relation between these subsets is unknown. A straightforward extension of our results in the nonparametric case leads to the following result: if we have k (partial) orderings, each with $n_1, \ldots, n_k$ samples, and $m_1, \ldots, m_k$ samples in each ordering are labeled, we can show an upper bound on the MSE of $m_1^{-2/3} + \cdots + m_k^{-2/3} + (n_1 + \cdots + n_k)^{-2s/d}$. It would be interesting to study the optimal rate, as well as to consider other smaller partial orders.

Other models for ordinal information: Beyond the bounded noise model for comparison, we can consider other pairwise comparison models, like Plackett-Luce (Plackett, 1975; Luce, 2005) and Thurstone (Thurstone, 1927). These parametric models can be quite restrictive and

can lead to unnatural results that we can recover the function values even *without* querying any direct labels (see for example Shah et al. (2016a)). One might also consider pairwise comparisons with Tsybakov-like noise Tsybakov (2004) which have been studied in the classification setting Xu et al. (2017); the main obstacle here is the lack of computationally-efficient algorithms that aggregate pairwise comparisons into a complete ranking under this noise model.

Other classes of functions: Several recent papers (Chatterjee et al., 2015; Bellec and Tsybakov, 2015; Bellec, 2018; Han et al., 2017) demonstrate the adaptivity (to "complexity" of the unknown parameter) of the MLE in shape-constrained problems. Understanding precise assumptions on the underlying smooth function which induces a low-complexity isotonic regression problem is interesting future work.

Acknowledgements

We thank Hariank Muthakana for his help on the age prediction experiments. We also thank the reviewers and editors for their comments and suggestions on this paper. This work has been partially supported by the Air Force Research Laboratory (8750-17-2-0212), the National Science Foundation (CIF-1763734 and DMS-1713003), Defense Advanced Research Projects Agency (FA8750-17-2-0130), and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

References

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pages 39–75, 2017.

E Agustsson, R Timofte, S Escalera, X Baro, I Guyon, and R Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE, 2017.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458. ACM, 2014.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory*, pages 152–192, 2016.

R.E. Barlow. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley, 1972.

Andrew R. Barron. *Complexity Regularization with Application to Artificial Neural Networks*, chapter 7, pages 561–576. Springer Netherlands, 1991.

Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

Pierre C Bellec and Alexandre B Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16:1879–1892, 2015.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 08 2015.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343–351, 2010.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015.

Cecil C Craig. On the Tchebychef inequality of Bernstein. *The Annals of Mathematical Statistics*, 4(2):94–102, 1933.

Sanjoy Dasgupta and Michael Luby. Learning from partial correction. *arXiv preprint arXiv:1705.08076*, 2017.

Persi Diaconis and Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268, 1977.

Felix A. Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million Elpasolite(ABC2D6)crystals. *Physical Review Letters*, 117(13), 2016.

Edgar N Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31 (3):504–522, 1952.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *arXiv preprint arXiv:1708.09468*, 2017.

Steve Hanneke. *Theoretical foundations of active learning*. ProQuest, 2009.

Wolfgang Karl Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric models*. Springer Science & Business Media, 2012.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

Daniel M Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries. *arXiv preprint arXiv:1704.03564*, 2017.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

Robin L Plackett. The analysis of permutations. *Applied Statistics*, pages 193–202, 1975.

Stefanos Poulis and Sanjoy Dasgupta. Learning with Feature Feedback: from Theory to Practice. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Harish G Ramaswamy, Ambuj Tewari, Shivani Agarwal, et al. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1): 530–554, 2018.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*, pages 41–48, 2004.

Nihar Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 2016a.

Nihar Shah, Sivaraman Balakrishnan, Aditya Guntuboyina, and Martin Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 2016b.

Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness, 2016c.

Juan Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms and software. *arXiv preprint arXiv:1812.05944*, 2018.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

Kristi Tsukida and Maya R Gupta. How to analyze paired comparison data. Technical report, WASHINGTON UNIV SEATTLE DEPT OF ELECTRICAL ENGINEERING, 2011.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

RR Varshamov. Estimate of the number of signals in error correcting codes. In *Dokl. Akad. Nauk SSSR*, 1957.

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

Eric P. Xing, Michael I. Jordan, Stuart J Russell, and Andrew Y. Ng. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 521–528. MIT Press, 2003.

Yichong Xu, Hongyang Zhang, Kyle Miller, Aarti Singh, and Artur Dubrawski. Noise-tolerant interactive learning from pairwise comparisons with near-minimal label complexity. *arXiv preprint arXiv:1704.05820*, 2017.

Dezhen Xue, Prasanna V. Balachandran, John Hogden, James Theiler, Deqing Xue, and Turab Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nature Communications*, 7, 2016.

Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems*, pages 1056–1066, 2017.

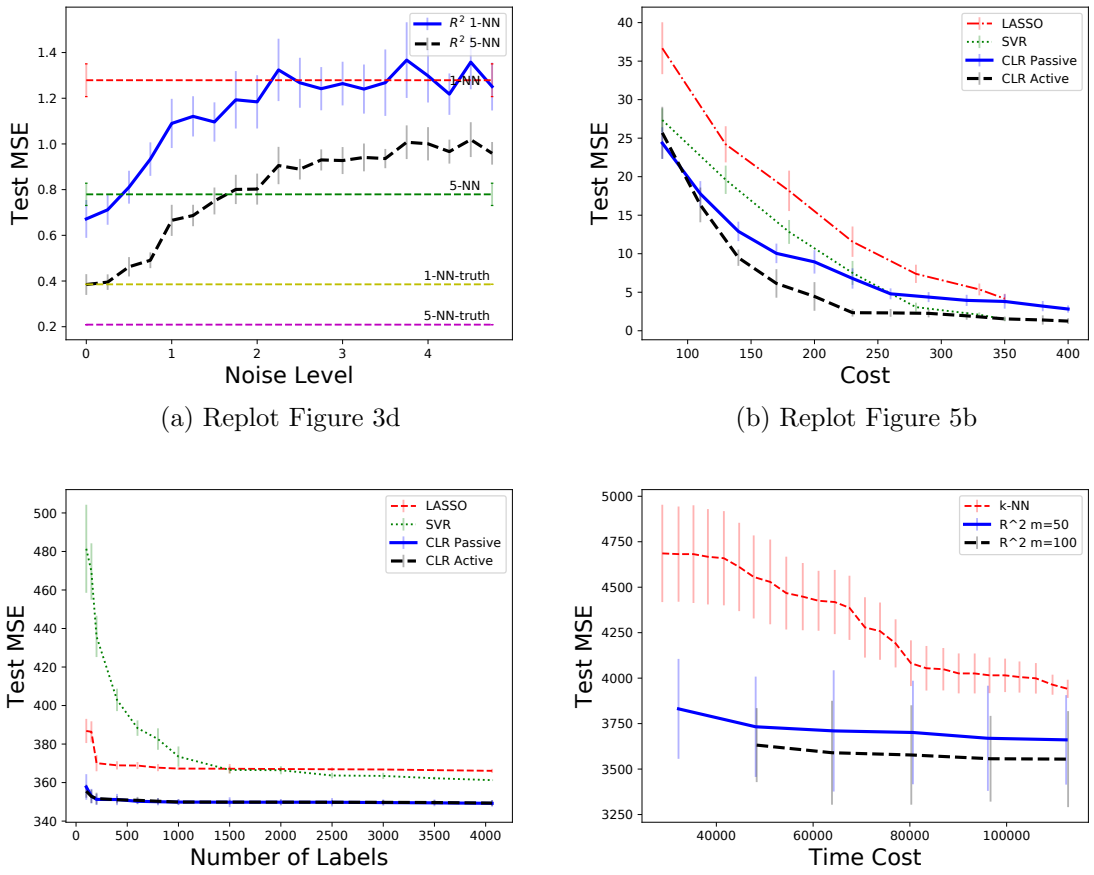
Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.

James Y. Zou, Kamalika Chaudhuri, and Adam Tauman Kalai. Crowdsourcing feature discovery via adaptively chosen comparisons. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California.*, page 198, 2015.

Appendix A. Additional Experimental Results

Ranges of the performance. We do not include the ranges of the performances in our previous experiment plots, because we compare many variants of our algorithms and baselines, and plotting the range can make it hard to see the difference. Our methods usually outperform the baselines by a large margin; we replot some of the figures in Figure 9. In addition to the average performance from 20 (50 for Figure 3c) experiments, we also plot the two-sided confidence interval with a significance level of 0.05. Our algorithms consistently outperform the baselines in these experiments, and our conclusions remain the same.

An alternative to using RankSVM. As an alternative to training RankSVM, we can also use nearest neighbors on Borda counts to take into account the structure of feature



(a) Replot Figure 3d

(b) Replot Figure 5b

(c) Replot Figure 6b

(d) Replot Figure 8a

Figure 9: Replots of previous experiments with confidence intervals.

space: for each sample x, we use the score $s(x) = \frac{1}{k}\sum_{x' \in k\text{-NN}(x)} \text{Borda}(x')$, where k-NN(x) is the k-th nearest neighbor of x in the feature space, including x itself. The scores are then used to produce a ranking. When c is large, this method does provide an improvement over the label-only baselines, but generally does not perform as well as our rankSVM method. The results when cost ratio $c = 10$ and comparisons are perfect are depicted in Figure 10. We use $k = 25$ for deciding the nearest neighbors for Borda counts, and 5-NN as the final prediction step. While using R^2 with Borda counts do provide a gain over label-only methods, the improvement is less prominent than using rankSVM.

For estimating AirBnB price the results are shown in Figure 11. For task 1, NN of Borda counts introduces an improvement similar to (or less than) RankSVM, but for task 2 it is worse than nearest neighbors. We note that for task 1, the best number of nearest neighbors of Borda counts is 50, whereas for task 2 it is 5 (close to raw Borda counts). We suspect that this is due to the larger noise in estimating the true price, however a close examination for this observation remains as future work.

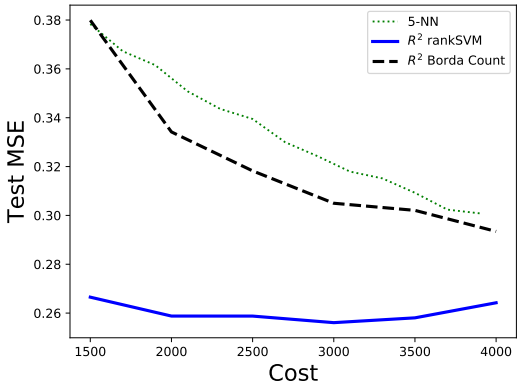
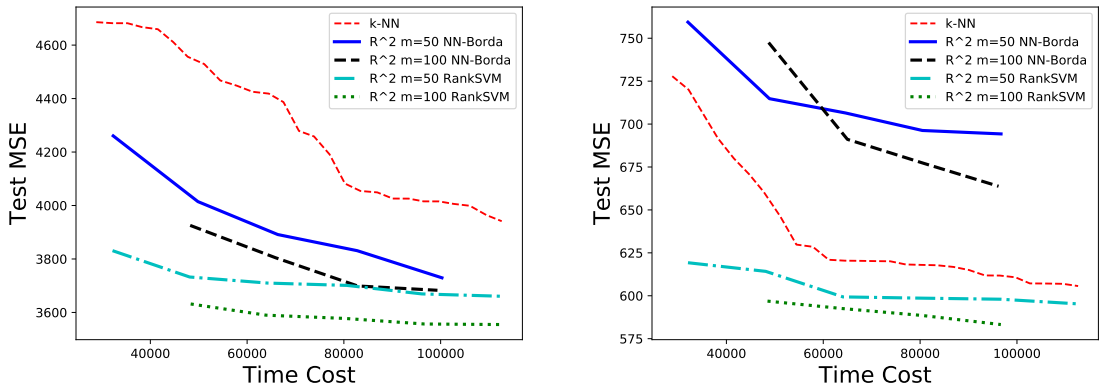


Figure 10: Experiments on a synthetic dataset for R^2 with comparisons, using nearest neighbor with Borda counts. Both R^2-based algorithms uses 5-NN as the final prediction method.



(a) Task 1 (b) Task 2

Figure 11: Experiments on AirBnB price estimation for nonparametric methods, using nearest neighbors of Borda count. Figure (a) uses 50-NN for averaging Borda counts, while Figure (b) uses 5-NN.

Appendix B. Detailed Proofs

B.1. Proof of Theorem 1

Without loss of generality we assume throughout the proof that we re-arrange the samples so that the true ranking of the samples π is the identity permutation, i.e. that $f(X_1) \leq f(X_2) \leq \ldots \leq f(X_n)$. We let $C, c, C_1, c_1, \ldots$ denote universal positive constants. As is

standard in nonparametric regression these constants may depend on the dimension d but we suppress this dependence.

For a random point $X \in \mathcal{X}$, let X_α be the nearest neighbor of X in the labeled set $\mathcal{L}$. We decompose the MSE as

$$\mathbb{E}\left[(\widehat{f}(X) - f(X))^2\right] \leq 2\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] + 2\mathbb{E}\left[(f(X_\alpha) - f(X))^2\right].$$

Under the assumptions of the theorem we have the following two results which provide bounds on the two terms in the above decomposition.

Lemma 13 *For a constant $C > 0$ and $0 < s \leq 1$ we have that,*

$$\mathbb{E}\left[(f(X_\alpha) - f(X))^2\right] \leq Cn^{-2s/d}.$$

Lemma 14 *Let $t_0 = 0$ and $t_{m+1} = n + 1$, and let X_0 and X_{n+1} be two imaginary variables such that $f(X_0) = -M$ and $f(X_{n+1}) = M$ (for easier presentation of the results). For any $0 < \delta \leq 1/2$ we have that there is a constant $C > 0$ such that:*

$$\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] \leq 4\delta M^2 + \frac{C\log(m/\delta)\log n\log(1/\delta)}{m}\left(m^{1/3} + M^2\right).$$

Taking these results as given and choosing $\delta = \max\{n^{-2s/d}, 1/m\}$ we obtain:

$$\mathbb{E}\left[(\widehat{f}(X) - f(X))^2\right] \leq C_1 m^{-2/3}\log^2 n\log m + C_2 n^{-2s/d},$$

as desired. We now prove the two technical lemmas to complete the proof.

B.1.1. Proof of Lemma 13

The proof of this result is an almost immediate consequence of the following result from Györfi et al. (2006).

Lemma 15 (Györfi et al. (2006), Lemma 6.4 and Exercise 6.7) *Suppose that there exist positive constants $p_{\min}$ and p_{max} such that $p_{\min} \leq p(x) \leq p_{\max}$. Then, there is a constant $c > 0$, such that*

$$\mathbb{E}[\|X_\alpha - X\|_2^2] \leq cn^{-2/d}.$$

Using this result and the Hölder condition we have

$$\begin{aligned}\mathbb{E}\left[(f(X_\alpha) - f(X))^2\right] &\leq L\mathbb{E}\left[\|X_\alpha - X\|_2^{2s}\right] \\ &\overset{\text{(i)}}{\leq} L\left(\mathbb{E}\left[\|X_\alpha - X\|_2^2\right]\right)^s \\ &\leq cn^{-2s/d},\end{aligned}$$

where (i) uses Jensen's inequality. We now turn our attention to the remaining technical lemma.

B.1.2. Proof of Lemma 14

We condition on a certain favorable configuration of the samples that holds with high-probability. For a sample $\{X_1, \ldots, X_n\}$ let us denote by

$$q_i := \mathbb{P}(X_\alpha = X_i),$$

where X_α is the nearest neighbor of X. Furthermore, for each k we recall that since we have re-arranged the samples so that π is the identity permutation we can measure the distance between adjacent labeled samples in the ranking by $t_k - t_{k-1}$. The following result shows that the labeled samples are roughly uniformly spaced (up to a logarithmic factor) in the ranked sequence, and that each point X_i is roughly equally likely (up to a logarithmic factor) to be the nearest neighbor of a randomly chosen point.

Lemma 16 *There is a constant $C > 0$ such that with probability at least $1 - \delta$ we have that the following two results hold simultaneously:*

1. *We have that,*

$$\max_{1 \le j \le n} q_j \le \frac{Cd \log(1/\delta) \log n}{n}. \tag{10}$$

2. *Recall that we defined $t_{m+1} = n + 1$ in Lemma 14. We have*

$$\max_{k \in [m+1]} t_k - t_{k-1} \le \frac{Cn \log(m/\delta)}{m}. \tag{11}$$

Denote the event, which holds with probability at least $1 - \delta$ in the above Lemma by $\mathcal{E}_0$. By conditioning on $\mathcal{E}_0$ we obtain the following decomposition:

$$\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] \le \mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2 | \mathcal{E}_0\right] + \delta \cdot 4M^2$$

because both f and $\widehat{f}$ are bounded in $[-M, M]$. We condition all calculations below on $\mathcal{E}_0$. Now we have

$$\begin{aligned}
\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2 | \mathcal{E}_0\right] &= \sum_{i=1}^n \mathbb{P}[X_\alpha = X_i | \mathcal{E}_0] \mathbb{E}\left[(\widehat{f}(X_i) - f(X_i))^2 | \mathcal{E}_0, \alpha = i\right] \\
&\le \sum_{i=1}^n \max_{1 \le j \le n} \mathbb{P}[X_\alpha = X_j | \mathcal{E}_0] \mathbb{E}\left[(\widehat{y}_i - f(X_i))^2 | \mathcal{E}_0, \alpha = i\right] \\
&\le \frac{Cd \log(1/\delta) \log n}{n} \sum_{i=1}^n \mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2 | \mathcal{E}_0, \alpha = i\right], \qquad (12)
\end{aligned}$$

In the first equality, we use the fact that $\widehat{f}(X) = \widehat{f}(X_\alpha)$, and therefore if $\alpha = i$ we have $\widehat{f}(X) = \widehat{f}(X_i)$. In the expectation we drop the condition $X_\alpha = X_i$ since pick of X is independent of the choice of X_i. In the third inequality, recall that $\widehat{y}_{\beta(i)}$ (defined in Algorithm 1) denotes the de-noised (by isotonic regression) y value at the nearest labeled left-neighbor of the point X_i.

We now compute $\mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2 | \mathcal{E}_0, \alpha = i\right]$. We condition the following computation on a set of unlabeled points $\mathcal{U} = \{X_1, ..., X_n\}$ for which the event $\mathcal{E}_0$ happens; now the randomness is only due to the randomness in $\widehat{y}$. Recall that we define $t_0 = 0$ and $f(X_{t_0}) = -M$ in Lemma 14. Equivalently as in Algorithm 1 we are assigning a 0 value to any point with no labeled point with a smaller function value according to the permutation $\widehat{\pi}$. With this definition in place we have that,

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2 | \mathcal{E}_0, \mathcal{U}, \alpha = i\right] \overset{\text{(i)}}{=} \sum_{i=1}^{n} \mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2 | \mathcal{U}\right] \\
\leq &\sum_{i=1}^{n} \left(2\mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_{\beta(i)}))^2 | \mathcal{U}\right] + 2\mathbb{E}\left[(f(X_i) - f(X_{\beta(i)}))^2 | \mathcal{U}\right]\right) \\
\overset{\text{(ii)}}{=} &\sum_{i=1}^{n} \sum_{k=0}^{m} \mathbb{I}[\beta(i) = t_k] \left(2\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | \mathcal{U}\right] + 2\mathbb{E}\left[(f(X_i) - f(X_{t_k}))^2 | \mathcal{U}\right]\right) \\
\overset{\text{(iii)}}{\leq} &\sum_{i=1}^{n} \sum_{k=0}^{m} \mathbb{I}[\beta(i) = t_k] \left(2\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | \mathcal{U}\right] + 2\mathbb{E}\left[(f(X_{t_{k+1}}) - f(X_{t_k}))^2 | \mathcal{U}\right]\right) \\
\overset{\text{(iv)}}{=} &\sum_{k=0}^{m} \left(2\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | \mathcal{U}\right] + 2\mathbb{E}\left[(f(X_{t_{k+1}}) - f(X_{t_k}))^2 | \mathcal{U}\right]\right) \sum_{i=1}^{n} \mathbb{I}[\beta(i) = t_k] \\
\overset{\text{(v)}}{\leq} &\frac{Cn\log(m/\delta)}{m} \sum_{k=1}^{m} \left(2\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | \mathcal{U}\right]\right) + \frac{Cn\log(m/\delta)}{m} \sum_{k=0}^{m} \left(2\mathbb{E}\left[(f(X_{t_{k+1}}) - f(X_{t_k}))^2 | \mathcal{U}\right]\right) \\
\leq &\frac{2Cn\log(m/\delta)}{m} \left[\sum_{k=1}^{m} \left(\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | \mathcal{U}\right]\right) + \sum_{k=0}^{m} \left(\mathbb{E}\left[(f(X_{t_{k+1}}) - f(X_{t_k}))^2 | \mathcal{U}\right]\right)\right].
\end{aligned}
$$

In equality (i) we drop the dependence on $\alpha = i$ because α is determined by $\mathcal{U}$ and the draw of X, which is independent of $\widehat{y}$. We drop dependence on $\mathcal{E}_0$ because $\mathcal{U}$ contains $\mathcal{E}_0$. In equality (ii) we enumerate the possible values of $\beta(i)$, and symbolically define $\widehat{y}_0 = f(X_0) = -M$. The inequality (iii) follows by noticing that if $\beta(i) = t_k$, $f(X_i) - f(X_{\beta(i)})$ is upper bounded by $f(X_{t_{k+1}}) - f(X_{t_k})$. We interchange the order of summations to obtain the inequality (iv). The inequality in (v) uses Lemma 16.

The first term in the upper bound above is simply the MSE in an isotonic regression problem, and using standard risk bounds for isotonic regression (see for instance, Theorem 2.2 in Zhang (2002)) we obtain that for a constant $C > 0$:

$$\sum_{k=1}^{m} \left(\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | \mathcal{U}\right]\right) \leq Cm^{1/3}.$$

Furthermore, since $f(X_{t_{m+1}}) - f(X_{t_0}) \leq 2M$, and the function values are increasing we obtain that:

$$\sum_{k=0}^{m} \left(\mathbb{E}\left[(f(X_{t_{k+1}}) - f(X_{t_k}))^2 | \mathcal{U}\right]\right) \leq 4M^2.$$

Taking expectation over all $\mathcal{U}$ and plugging this expression back in to (12), we obtain the Lemma. Thus, to complete the proof it only remains to establish the result in Lemma 16.

B.1.3. Proof of Lemma 16

We prove that both of the two results hold with a probability at least $1 - \delta/2$, and in turn Lemma 16 follows as a corollary.

Proof of Inequality (10)**:** As a preliminary, we need the following Vapnik-Cervonenkis result from Chaudhuri and Dasgupta (2010):

Lemma 17 (Chaudhuri and Dasgupta (2010), Lemma 7) *Suppose we draw a sample $\{X_1, \ldots, X_n\}$, from a distribution $\mathbb{P}$, then there exists a universal constant C' such that with probability $1 - \delta$, every (open) ball B with probability:*

$$\mathbb{P}(B) \geq \frac{C' \log(1/\delta) d \log n}{n},$$

contains at least one of the sample points.

We now show that under this event we have

$$\max_i q_i \leq \frac{C' p_{\max} \log(1/\delta) d \log n}{p_{\min} n}.$$

Recall that $\mathcal{U} = \{X_1, X_2, ..., X_n\}$. Fix any point $X_i \in \mathcal{U}$, and for a new point X, let $r = \|X_i - X\|_2$. If X_i is X's nearest neighbor in T, there is no point in the ball $B(X, r)$. Comparing this with the event in Lemma 17 (and also using $\delta/2$ instead of δ, adapting the constant C') we have

$$p_{\min} v_d r^d \leq \frac{C' \log(1/\delta) d \log n}{n},$$

where v_d is the volume of the unit ball in d dimension.

Hence we obtain an upper bound on r. Now since $p(x)$ is upper and lower bounded we can bound the largest q_i as

$$\max_i q_i \leq p_{\max} v_d r^d \leq \frac{C' p_{\max} \log(1/\delta) d \log n}{p_{\min} n}.$$

Thus we obtain the inequality (10).

Proof of Inequality (11)**:** Recall that we define $t_{m+1} := n + 1$. Notice that $t_1, \ldots, t_m$ are randomly chosen from $[n]$. So for each $k \in [m]$ we have

$$\mathbb{P}[t_k - t_{k-1} \geq t] \leq \frac{n - t + 1}{n} \left(\frac{n - t}{n}\right)^{m-1} \leq \left(\frac{n - t}{n}\right)^{m-1},$$

since we must randomly choose t_k in $X_t, X_{t+1}, \ldots, X_n$, and choose the other $m - 1$ samples in $X_1, \ldots, X_{t_k - t}, X_{t_k + 1}, \ldots, X_n$. Similarly we also have

$$\mathbb{P}[t_{m+1} - t_m \geq t] \leq \left(\frac{n - t}{n}\right)^{m-1}.$$

So

$$\mathbb{P}[\max_{k\in[m+1]} t_k - t_{k-1} \geq t] \leq \sum_{k=1}^{m+1} \mathbb{P}[t_k - t_{k-1} \geq t] \leq (m+1)\left(\frac{n-t}{n}\right)^{m-1}.$$

Setting this to be less than or equal to $\delta/2$, we have

$$\frac{t}{n} \geq 1 - \left(\frac{\delta}{2(m+1)}\right)^{\frac{1}{m-1}}.$$

Routine calculation shows that it suffices for $t \geq C\frac{n\log(m/\delta)}{m}$ such that

$$\mathbb{P}[\max_{k\in[m+1]} t_k - t_{k-1} \geq t] \leq \delta/2.$$

B.2. Proof of Theorem 2

To prove the result we separately establish lower bounds on the size of the labeled set of samples m and the size of the ordered set of samples n. Concretely, we show the following pair of claims, for a positive constant $C > 0$,

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}_{s,L}} \mathbb{E}\left[(f(X) - \widehat{f}(X))^2\right] \geq Cm^{-2/3} \tag{13}$$

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}_{s,L}} \mathbb{E}\left[(f(X) - \widehat{f}(X))^2\right] \geq Cn^{-2s/d}. \tag{14}$$

We prove each of these claims in turn and note that together they establish Theorem 2.

Proof of Claim (13)**:** We establish the lower bound in this case by constructing a suitable packing set of functions, and using Fano's inequality. The main technical novelty, that allows us to deal with the setting where both direct and ordinal information is available, is that we construct functions that are all increasing functions of the first covariate x_1, and for these functions the ordinal measurements provide no additional information.

Without loss of generality we consider the case when $d = 1$, and note that the claim follows in general by simply extending our construction using functions for which $f(x) = f(x_1)$. We take the covariate distribution $\mathbb{P}_X$ to be uniform on $[0, 1]$. For a kernel function K that is 1-Lipschitz on $\mathbb{R}$, bounded and supported on $[-1/2, 1/2]$, with

$$\int_{-1/2}^{1/2} K^2(x)dx > 0$$

we define:

$$u = \lceil m^{\frac{1}{3}} \rceil, \quad h = 1/u,$$
$$x_k = \frac{k-1/2}{u}, \quad \phi_k(x) = \frac{Lh}{2}K\left(\frac{x-x_k}{h}\right), \quad \text{for} \quad k = \{1, 2, \ldots, u\},$$
$$\Omega = \{\omega : \omega \in \{0, 1\}^u\}.$$

With these definitions in place we consider the following class of functions:

$$\mathcal{G} = \left\{ f_\omega : [0,1] \mapsto \mathbb{R} : f_\omega(x) = \frac{Lx}{2} + \sum_{k=1}^{u} \omega_k \phi_k(x), \text{ for } x \in [0,1] \right\}.$$

We note that the functions in $\mathcal{G}$ are L-Lipschitz, and thus satisfy the Hölder constraint (for any $0 < s \leq 1$).

We note that these functions are all increasing so the permutation π contains no additional information and can be obtained simply by sorting the samples (according to their first coordinate). Furthermore, in this case the unlabeled samples contribute no additional information as their distribution $\mathbb{P}_X$ is known (since we take it to be uniform). Concretely, for any estimator in our setup which uses $\{(X_1, y_1), \ldots, (X_m, y_m), X_{m+1}, \ldots, X_n, \pi\}$ with

$$\sup_{f \in \mathcal{G}} \mathbb{E}\left[(f(X) - \widehat{f}(X))^2 \right] < C m^{-2/3},$$

we can construct an equivalent estimator that uses only $\{(X_1, y_1), \ldots, (X_m, y_m)\}$. In particular, we can simply augment the sample by sampling $X_{m+1}, \ldots, X_n$ uniformly on $[0,1]$ and generating π by ranking X in increasing order.

In light of this observation, in order to complete the proof of Claim (13) it suffices to show that $Cm^{-2/3}$ is a lower bound for estimating functions in $\mathcal{G}$ with access to only noisy labels. For any pair $\omega, \omega' \in \Omega$ we have that,

$$\begin{aligned}
\mathbb{E}[(f_\omega(X) - f_{\omega'}(X))^2] &= \sum_{k=1}^{u} (\omega_k - \omega'_k)^2 \int \phi_k^2(x) dx \\
&= L^2 h^3 \|K\|_2^2 \rho(\omega, \omega'),
\end{aligned}$$

where $\rho(\omega, \omega')$ denotes the Hamming distance between x and x', and $\|K\|^2 = \int_{-1/2}^{1/2} K^2(x) dx$ is a finite constant.

Denote by P_0 the distribution induced by the function f_0 with $\omega = (0, \ldots, 0)$, with the covariate distribution being uniform on $[0,1]$. We can upper bound the KL divergence between the distribution induced by any function in $\mathcal{G}$ and P_0 as:

$$\begin{aligned}
\mathrm{KL}(P_j^m, P_0^m) &= m \int_{\mathcal{X}} p(x) \int_{\mathbb{R}} p_j(y|x) \log \frac{p_0(y|x)}{p_j(y|x)} dy dx \\
&= m \int_{\mathcal{X}} p(x) \sum_{i=1}^{u} \omega_i^j \phi_i^2(x) dx \\
&\leq m L^2 h^3 \|K\|_2^2 u.
\end{aligned}$$

Now, the Gilbert-Varshamov bound (Varshamov, 1957; Gilbert, 1952), ensures that if $u > 8$, then there is a subset $\Omega' \subseteq \Omega$ of cardinality $2^{u/8}$, such that the Hamming distance between each pair of elements $\omega, \omega' \in \Omega'$ is at least $u/8$. We use the following version of Fano's inequality:

Theorem 18 (Theorem 2.5, Tsybakov (2008), adapted) *Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_1, ..., \theta_M$ such that: (i) $d(\theta_j, \theta_k) \geq 2s > 0, \forall 0 \leq j < k \leq$*

M; (ii) $\frac{1}{M}\sum_{j=1}^{M} KL(P_j, P_0) \le \alpha \log M$, *with* $0 < \alpha < 1/8$ *and* $P_j = P_{\theta_j}, j = 0, 1, ..., M$. *Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \ge s) \ge \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0.$$

Using Theorem 18 with $\Theta = \Omega$, $\alpha = mL^2h^3\|K\|_2^2$, $s = L^2h^3\|K\|_2^2 u$, we pick h such that

$$mL^2h^3\|K\|_2^2 u \le cu,$$

for a small constant c. So for another constant c' we have the error of any estimator is lower bounded as:

$$\sup_{f \in \mathcal{G}} \mathbb{E}[(\widehat{f}(X) - f(X))^2] \ge c' L^2 h^3 \|K\|_2^2 u \ge c' m^{-2/3},$$

establishing the desired claim.

Proof of Claim (14)**:** We show this claim by reducing to the case where we have n points with noiseless evaluations, i.e., we observe $\{(X_1, f(X_1)), (X_2, f(X_2)), \ldots, (X_n, f(X_n))\}$. We notice that the ranking π provides no additional information when all the points are labeled without noise. Formally, if we have an estimator $\widehat{f}$ which when provided $\{(X_1, y_1), \ldots, (X_m, y_m), X_{m+1}, \ldots, X_n, \pi\}$ obtains an error less than $cn^{-2s/d}$ (for some sufficiently small constant $c > 0$) then we can also use this estimator in setting where all n samples are labeled without noise, by generating π according to the noiseless labels, adding Gaussian noise to the labels of m points, and eliminating the remaining labels before using the estimator $\widehat{f}$ on this modified sample.

It remains to show that, $cn^{-2s/d}$ is a lower bound on the MSE of any estimator that receives n noiseless labels. To simplify our notation we will assume that $(2n)^{1/d}$ is an integer (if not we can establish the lower bound for a larger sample-size for which this condition is true, and conclude the desired lower bound with an adjustment of various constants). For a given sample size n, we choose

$$h = (2n)^{-1/d},$$

and consider the grid with $2n$ cubes with side-length h. Denote the centers of the cubes as $\{x_1, \ldots, x_{2n}\}$. For a kernel function K supported on $[-1/2, 1/2]^d$, which is 1-Lipschitz on $\mathbb{R}^d$, bounded and satisfies:

$$\int_{[-1/2,1/2]^d} K^2(x)dx > 0$$

we consider a class of functions:

$$\mathcal{G} = \left\{ f_\omega : f_\omega(x) = Lh^s \sum_{i=1}^{2n} \omega_i K\left(\frac{x - x_i}{h}\right), \text{for } \omega \in \{0,1\}^{2n} \right\}.$$

We note that these functions are all in the desired Hölder class with exponent s. Given n samples (these may be arbitrarily distributed) $\{(X_1, f(X_1)), \ldots, (X_n, f(X_n))\}$, we notice that we are only able to identify at most n of the ω_i (while leaving at least n of the ω_i

completely unconstrained) and thus any estimator $\widehat{f}$ must incur a large error on at least one of the functions f_ω consistent with the obtained samples. Formally, we have that

$$\sup_{f\in\mathcal{G}} \mathbb{E}(\widehat{f}(X) - f(X))^2 \geq \frac{nL^2h^{2s+d}\|K\|_2^2}{4} \geq cn^{-2s/d},$$

as desired. This completes the proof of Claim (14).

B.3. Proof of Theorem 3

Throughout this proof without loss of generality we re-arrange the samples so that the estimated permutation $\widehat{\pi}$ is the identity permutation. To simplify the notation further, we let $X_{(i)} = X_{\pi^{-1}(i)}$ be the i-th element according to true permutation π. This leads to $f(X_{(1)}) \leq f(X_{(2)}) \leq \cdots \leq f(X_{(n)})$.

We begin with a technical result that bounds the error of using $\widehat{\pi}$ instead of the true permutation π in the R^2 algorithm.

The first part of the proof is the same as that of Theorem 1. We have

$$\begin{aligned}\mathbb{E}\left[(\widehat{f}(X) - f(X))^2\right] &\leq 2\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] + 2\mathbb{E}\left[(f(X_\alpha) - f(X))^2\right] \\ &\leq 2\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] + Cn^{-2s/d}.\end{aligned}$$

And for event $\mathcal{E}_0$ we have (note that $\beta(i)$ is the nearest neighbor index defined in Algorithm 1)

$$\begin{aligned}\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] &\leq C\frac{d\log(1/\delta)\log n}{n}\sum_{i=1}^n \mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2|\mathcal{E}_0\right] + \delta \\ &\leq C\left(\frac{\log^2 n}{n}\sum_{i=1}^n \mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2|\mathcal{E}_0\right] + n^{-2s/d}\right). \qquad (15)\end{aligned}$$

The second inequality is obtained by letting $\delta = n^{-2s/d}$. To bound the sum of expectations above, we first prove a lemma bounding the difference between $X_1, \ldots, X_n$ and $X_{(1)}, \ldots, X_{(n)}$:

Lemma 4(Restated). Suppose the ranking $(X_1, \ldots, X_n)$ is of at most ν error with respect to the true permutation. Then

$$\sum_{i=1}^n (f(X_i) - f(X_{(i)})^2 \leq 8M^2\sqrt{2\nu}n.$$

Proof Let $\theta_i = f(X_i)$ and $\theta_{(i)} = f(X_{(i)})$, and let $g(\theta_1, \ldots, \theta_n) = \sum_{i=1}^n (\theta_i - \theta_{(i)})^2$. Consider g as a function of θ_i; θ_i appears twice in g, one as $(\theta_i - \theta_{(i)})^2$, and the other as $(\theta_{\pi(i)} - \theta_{(\pi(i))})^2 = (\theta_{\pi(i)} - \theta_i)^2$. If $\pi(i) = i$, then θ_i does not influence value of g; otherwise, g is a quadratic function of θ_i, and it achieves maximum either when $\theta_i = M$ or $\theta_i = -M$. So when g achieves maximum it must be $\theta_i \in \{-M, M\}$. Now notice that $\theta_{(1)} \leq \cdots \leq \theta_{(n)}$, so the maximum is achieved when for some $0 \leq k \leq n$ that $\theta_{(i)} = -M$ for $i \leq k$, and $\theta_{(i)} = M$ for $i > k$.

Note that $\sum_{i=1}^n(\theta_i-\theta_{(i)})^2=\sum_{i=1}^n(\theta_{\pi^{-1}(i)}-\theta_{(\pi^{-1}(i))})^2=\sum_{i=1}^n(\theta_{(i)}-\theta_{(\pi^{-1}(i))})^2$. From the discussion above, when g achieves its maximum we must have $(\theta_{(i)}-\theta_{(\pi^{-1}(i))})^2=4M^2$ if i and $\pi^{-1}(i)$ lies on different sides of k, or otherwise it is 0. To further bound the sum, we use the Spearman Footrule distance between π and $(1,2,\ldots,n)$, which Diaconis and Graham (1977) shows that it can be bounded as

$$\sum_{i=1}^n|\pi(i)-i|\le 2\sum_{1\le i,j\le n}I\left[(\pi(i)-\pi(j))(i-j)<0\right].$$

The RHS can be bounded by $2\nu n^2$ since the agnostic error of $\widehat{\pi}$ is at most ν. We also have that

$$\sum_{i=1}^n|\pi(i)-i|=\sum_{i=1}^n|\pi(\pi^{-1}(i))-\pi^{-1}(i)|=\sum_{i=1}^n|i-\pi^{-1}(i)|.$$

Let $U_1=\{i:\pi^{-1}(i)\le k,i>k\}$ and $U_2=\{\pi^{-1}(i)>k,i\le k\}$. So when g achieves the maximum value we have

$$\sum_{i=1}^n(\theta_{\pi(i)}-\theta_i)^2=4M^2(|U_1|+|U_2|).$$

Now notice that for $i\in U_1$, we have $|\pi^{-1}(i)-i|\ge i-k$; and for $i\in U_2$ we have $|\pi^{-1}(i)-i|\ge k-i+1$. Considering the range of i we have

$$\sum_{j=1}^{|U_1|}j+\sum_{j=1}^{|U_2|}j\le\sum_{i=1}^n|\pi^{-1}(i)-i|\le 2\nu n^2.$$

So $|U_1|+|U_2|\le 2\sqrt{2\nu}n$. And

$$\sum_{i=1}^n(f(X_i)-f(X_{(i)})^2=\sum_{i=1}^n(\theta_{\pi(i)}-\theta_i)^2\le 4M^2(|U_1|+|U_2|)\le 8M^2\sqrt{2\nu}n.$$

Thus we prove the lemma. ■

Now back to the original proof. Under event $\mathcal{E}_0$ we have

$$\begin{aligned}
&\sum_{i=1}^n E[(\widehat{y}_i-f(X_i))^2|\mathcal{E}_0]\\
=&\sum_{i=1}^n E[(\widehat{y}_{\beta(i)}-f(X_i))^2|\mathcal{E}_0]\\
\le&\sum_{i=1}^n\mathbb{E}\left[2(\widehat{y}_{\beta(i)}-f(X_{\beta(i)}))^2+2(f(X_{\beta(i)})-f(X_i))^2|\mathcal{E}_0\right]\\
\le&\frac{Cn\log(m/\delta)}{m}\sum_{k=1}^m\mathbb{E}\left[(\widehat{y}_{t_k}-f(X_{t_k}))^2|\mathcal{E}_0\right]+2\sum_{i=1}^n\mathbb{E}\left[(f(X_{\beta(i)})-f(X_i))^2|\mathcal{E}_0\right]. \qquad (16)
\end{aligned}$$

We omit the condition $\mathcal{E}_0$ in discussion below, and we will bound the two terms separately. For the first term, we use the following theorem adapted from Zhang (2002):

Theorem 19 (Zhang (2002), adapted from Theorem 2.3 and Remark 4.2) *Suppose X_{t_k}, y_{t_k} are fixed for $k \in [m]$, and $f(X_{t_k})$ is arbitrary in order. Let*

$$S = \min_u \sum_{k=1}^{m} (u_k - f(X_{t_k}))^2,$$

where the minimum is taken over all sequence of $u \in \mathbb{R}^m$ that is non-decreasing. The risk of isotonic regression satisfies

$$\frac{1}{m^{1/3}M^{1/3}} \left(\mathbb{E}\left[\sum_{k=1}^{m} (\widehat{y}_{t_k} - f(X_{t_k}))^2 \right] - S \right) \leq C$$

for some universal constant C.

From Theorem 19 we know that

$$\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 \right] \leq \mathbb{E}[S] + Cm^{1/3},$$

where the expectation in $E[S]$ is taken w.r.t. the randomness in t_k. From Lemma 4 we know that

$$\sum_{i=1}^{n} (f(X_i) - f(X_{(i)}))^2 \leq C\sqrt{\nu} n.$$

Note that since t_k is chosen at random from $[n]$, each element X_i has equal probability $\frac{m}{n}$ to be picked; so

$$\mathbb{E}[S] \leq \mathbb{E}\left[\sum_{i=1}^{m} (f(X_{(t_k)}) - f(X_{t_k}))^2 \right] \leq C\sqrt{\nu} m.$$

Now we bound the second term in (16). We have

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[(f(X_{\beta(i)}) - f(X_i))^2 \right] \\
&\leq 3\sum_{i=1}^{n} \mathbb{E}\left[(f(X_{\beta(i)}) - f(X_{(\beta(i))}))^2 \right] + 3\sum_{i=1}^{n} \mathbb{E}\left[(f(X_{(\beta(i))}) - f(X_{(i)}))^2 \right] + 3\sum_{i=1}^{n} \mathbb{E}\left[(f(X_{(i)}) - f(X_i))^2 \right] \\
&\leq \frac{Cn\log m}{m} \sum_{k=1}^{m} \mathbb{E}\left[(f(X_{t_k}) - f(X_{(t_k)}))^2 \right] + \frac{Cn\log m}{m} \sum_{k=1}^{m} \mathbb{E}\left[(f(X_{(t_{k+1})}) - f(X_{(t_k)}))^2 \right] \\
&\qquad + 3\sum_{i=1}^{n} \mathbb{E}\left[(f(X_{(i)}) - f(X_i))^2 \right] \\
&\leq \frac{Cn\log m}{m} \sqrt{\nu} m + \frac{Cn\log m}{m} \cdot 1 + C\sqrt{\nu} n \\
&\leq C\sqrt{\nu} n \log m.
\end{aligned}$$

The first inequality is by noticing $(x+y+z)^2 \leq 3x^2 + 3y^2 + 3z^2$ for any number $x, y, z \in \mathbb{R}$; the second inequality is by grouping values of $\beta(i)$, and the choice of t_k; the third

inequality comes from analysis of the first term on $\sum_{k=1}^{m} \mathbb{E}\left[(f(X_{t_k}) - f(X_{(t_k)}))^2\right]$, the fact that $f(X_{(t_m)}) - f(X_{(t_1)}) \leq 1$, and Lemma 4.

Summarizing the two terms we have

$$\mathbb{E}\left[(\widehat{y}_{\beta(i)} - f(X_i))^2 | \mathcal{E}_0\right] \leq C(\sqrt{\nu}n + m^{-2/3}n)\log m.$$

Take this back to (15) we prove the theorem.

B.4. Proof of Theorem 5

To simplify notation, we suppose that we have m labeled samples $T = \{(X_i, y_i)\}_{i=1}^{m}$ for training and another m labeled samples $V = \{(X_i, y_i)\}_{i=m+1}^{2m}$ for validation. We consider the following models:

1. R^2 using both the ordinal data and the labeled samples in T, and we denote by $\widehat{f}_0$,
2. k-NN regression using only the labeled samples in T for $k \in [m]$ which we denote by $\{\widehat{f}_1, \ldots, \widehat{f}_m\}$.

We select the best model according to performance on validation set. We further restrict all estimators to be bounded in $[-M, M]$; i.e., when $\widehat{f}_j(x) < -M$ for some x and j, we clip its value by setting $\widehat{f}_j(x) = -M$ and we analogously clip the function when it exceeds M. We note in passing that this only reduces the MSE since the true function f is bounded between $[-M, M]$. Throughout the remainder of the proof we condition on the training set T but suppress this in our notation. We define the empirical validation risk of a function $\widehat{f}$ to be:

$$\widehat{R}^V\big(\widehat{f}\big) = \frac{1}{m}\sum_{i=m+1}^{2m}\big(y_i - \widehat{f}(X_i)\big)^2,$$

and the population MSE of a function $\widehat{f}$ to be

$$\mathrm{err}(\widehat{f}) = \mathbb{E}\big[\big(\widehat{f}(X) - f(X)\big)^2\big],$$

where $\epsilon \sim N(0,1)$ denotes the noise in the direct measurements. Now let

$$\widehat{f}^* = \operatorname*{arg\,min}_{j=0,\ldots,m} \widehat{R}^V(\widehat{f}_j),$$

be the best model selected using cross validation and

$$f^* = \operatorname*{arg\,min}_{j=0,\ldots,m} \mathrm{err}(\widehat{f}_j),$$

be the estimate with lowest MSE in $\widehat{f}_0, \ldots, \widehat{f}_m$. Let us denote:

$$\mathcal{G} = \{\widehat{f}_0, \ldots, \widehat{f}_m\}.$$

Recall that f denotes the true unknown regression function. Then in the sequel we show the following result:

Lemma 20 *With probability at least $1-\delta$, for any $\widehat{f} \in \mathcal{G}$ we have that the following hold for some constant $C > 0$,*

$$err(\widehat{f}) \leq 2\left[\widehat{R}^V(\widehat{f}) - \widehat{R}^V(f)\right] + \frac{C\log(m/\delta)}{m}, \tag{17}$$

$$\left[\widehat{R}^V(\widehat{f}) - \widehat{R}^V(f)\right] \leq 2err(\widehat{f}) + \frac{C\log(m/\delta)}{m}. \tag{18}$$

Since $\widehat{R}^V(\widehat{f^*}) \leq \widehat{R}^V(f^*)$ we obtain using (17) that with probability at least $1-\delta$,

$$\text{err}(\widehat{f^*}) \leq 2\left[\widehat{R}^V(f^*) - \widehat{R}^V(f)\right] + \frac{C\log(m/\delta)}{m}.$$

Since $f^* \in \mathcal{G}$, we can use (18) to obtain that with probability $1-\delta$,

$$\text{err}(\widehat{f^*}) \leq 4\text{err}(f^*) + \frac{2C\log(m/\delta)}{m}.$$

Since $\text{err}(\widehat{f^*})$ is a positive random variable, integrating this bound we obtain that,

$$\begin{aligned}\mathbb{E}[\text{err}(\widehat{f^*})] &= \int_0^\infty \mathbb{P}(\text{err}(\widehat{f^*}) \geq t)dt, \\ &\leq 4\text{err}(f^*) + \frac{6C\log(m)}{m}.\end{aligned}$$

So far we have implicitly conditioned throughout on the training set T. Taking an expectation over the training set yields:

$$\mathbb{E}[\text{err}(\widehat{f^*})] \leq 4\mathbb{E}[\text{err}(f^*)] + \frac{6C\log(m)}{m}.$$

We now note that,

$$\mathbb{E}[\text{err}(f^*)]] \leq \min_{j\in\{0,\ldots,m\}} \mathbb{E}[\text{err}(\widehat{f_j})].$$

Standard results on k-NN regression (for instance, a straightforward modification of Theorem 6.2 in Györfi et al. (2006) to deal with $0 < s \leq 1$ in the Hölder class) yield that for a constant $C > 0$,

$$\min_{j\in\{1,\ldots,m\}} \mathbb{E}[\text{err}(\widehat{f_j})] \leq Cm^{-2s/(2s+d)}.$$

Theorem 3 yields that,

$$\mathbb{E}[\text{err}(\widehat{f_0})] \leq C_1\left(\log^2 n \log m\left(m^{-2/3} + \sqrt{\nu}\right)\right) + C_2 n^{-2s/d},$$

and putting these together we obtain that,

$$\mathbb{E}[\text{err}(f^*)] \leq \widetilde{O}\left(m^{-2/3} + \min\{\sqrt{\nu}, m^{-\frac{2s}{2s+d}}\} + n^{-2s/d}\right),$$

and thus it only remains to prove Lemma 20 to complete the proof of the theorem.

B.4.1. Proof of Lemma 20

For a fixed classifier $\widehat{f}$ and for samples in the validation set $i \in \{m+1, \ldots, 2m\}$ we define the random variables:

$$Z_i = (y_i - \widehat{f}(X_i))^2 - (y_i - f(X_i))^2 = (\widehat{f}(X_i) - f(X_i))^2 + 2\epsilon_i(f(X_i) - \widehat{f}(X_i)),$$

and note that $\mathbb{E}[Z_i] = \mathrm{err}(\widehat{f})$. In order to obtain tail bounds on the average of the Z_i let us bound the absolute central moments of Z_i. Using the inequality that $(x+y)^k \leq 2^{k-1}(x^k + y^k)$, for $k > 2$ we obtain that,

$$\begin{aligned}\mathbb{E}|Z_i - \mathbb{E}[Z_i]|^k &= \mathbb{E}|(\widehat{f}(X_i) - f(X_i))^2 + 2\epsilon_i(f(X_i) - \widehat{f}(X_i)) - \mathrm{err}(\widehat{f})|^k \\ &\leq 2^{k-1}\mathbb{E}|(\widehat{f}(X_i) - f(X_i))^2 - \mathrm{err}(\widehat{f})|^k + 2^{k-1}\mathbb{E}|\epsilon_i(f(X_i) - \widehat{f}(X_i))|^k. \quad (19)\end{aligned}$$

We bound each of these terms in turn. Since $(\widehat{f}(X_i) - f(X_i))^2 \in [0, 4M^2]$, we obtain that,

$$\mathbb{E}|(\widehat{f}(X_i) - f(X_i))^2 - \mathrm{err}(\widehat{f})|^k \leq \mathrm{var}((\widehat{f}(X_i) - f(X_i))^2)(4M^2)^{k-2},$$

and using the fact that ϵ_i are Gaussian we obtain that,

$$\begin{aligned}\mathbb{E}|\epsilon_i(f(X_i) - \widehat{f}(X_i))|^k &\leq \mathbb{E}|\epsilon_i|^{k-2}\mathbb{E}(f(X_i) - \widehat{f}(X_i))^2(2M)^{k-2} \\ &\leq \mathrm{var}(\epsilon_i(f(X_i) - \widehat{f}(X_i)))k! \times (2M)^{k-2}.\end{aligned}$$

Since ϵ_i is independent of the other terms in Z_i we have that,

$$\mathrm{var}(Z_i) = \mathrm{var}(\epsilon_i(f(X_i) - \widehat{f}(X_i))) + \mathrm{var}((\widehat{f}(X_i) - f(X_i))^2).$$

Putting these pieces together with (19) we obtain,

$$\begin{aligned}\mathbb{E}|Z_i - \mathbb{E}[Z_i]|^k &\leq 2^{k-1}\mathrm{var}(Z_i)\left[k! \times (2M)^{k-2} + (4M^2)^{k-2}\right] \\ &\leq \frac{\mathrm{var}(Z_i)}{2}k!(16M + 32M^2)^{k-2}.\end{aligned}$$

Let us denote $r := 16M + 32M^2$. It remains to bound the variance. We have that,

$$\mathrm{var}(Z_i) \leq \mathbb{E}[Z_i^2] \leq 2\mathbb{E}((\widehat{f}(X_i) - f(X_i))^4) + 8\mathbb{E}(f(X_i) - \widehat{f}(X_i))^2,$$

and using the fact that the functions are bounded in $[-M, M]$ we obtain that,

$$\mathrm{var}(Z_i) \leq (8M^2 + 8)\mathrm{err}(\widehat{f}). \quad (20)$$

Now, applying the inequality in Lemma 24, we obtain that for any $c < 1$ and for any $t \leq c/r$ that,

$$\mathrm{err}(\widehat{f}) \leq \frac{1}{m}\sum_{i=m+1}^{2m} Z_i + \frac{\log(1/\delta)}{mt} + \frac{8t(M^2+1)\mathrm{err}(\widehat{f})}{2(1-c)},$$

we choose $c = 1/2$ and $t = \min\{1/(2r), 1/(16(M^2+1))\}$, and rearrange to obtain that,

$$\mathrm{err}(\widehat{f}) \leq \frac{2}{m}\sum_{i=m+1}^{2m} Z_i + \frac{2\log(1/\delta)}{m}\max\{2r, 16(M^2+1)\}, \leq \frac{2}{m}\sum_{i=m+1}^{2m} Z_i + \frac{C\log(1/\delta)}{m},$$

and using a union bound over the $m+1$ functions $\widehat{f} \in \mathcal{G}$ we obtain (17). Repeating this argument with the random variables $-Z_i$ we obtain (18) completing the proof of the Lemma.

B.5. Proof of Theorem 6

We prove a slightly stronger result, and show Theorem 6 as a corollary.

Theorem 21 *Assume the same modeling assumptions for $X_1, \ldots, X_n, y_1, \ldots, y_m$ as in Theorem 2. Also permutation $\widehat{\pi}$ satisfies $\mathbb{P}[(f(X_i) - f(X_j))(\pi(i) - \pi(j)) < 0] \leq \nu$. Then for any estimator $\widehat{f}$ taking input $X_1, \ldots, X_n, y_1, \ldots, y_m$ and $\widehat{\pi}$, we have*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{s,L}} \mathbb{E}\big(f(X) - \widehat{f}(X)\big)^2 \geq C(m^{-\frac{2}{3}} + \min\{\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}}, 1\} m^{-\frac{2}{d+2}} + n^{-2s/d}).$$

Proof of Theorem 21 In this proof, we use x_i to represent i-th dimension of x, and upper script for different vectors $x^{(1)}, x^{(2)}, \ldots$. Let $u = \lceil m^{\frac{1}{2+d}} \rceil, h = 1/u$, and $t = \min\left\{\left(\nu m^{\frac{1}{2+d}}\right)^{\frac{1}{2d}}, 1\right\}$. Let $\Gamma = \{(\gamma_1, \ldots, \gamma_d), \gamma_i \in \{1, 2, \ldots, u\}\}$. Choose an arbitrary order on Γ to be $\Gamma = \{\gamma^{(1)}, \ldots, \gamma^{(u^d)}\}$. Let $x^{(k)} = \frac{\gamma^{(k)} - 1/2}{u}$, and $\phi_k(x) = \frac{L}{2} thK(\frac{x - tx^{(k)}}{th}), k = 1, 2, \ldots, u^d$, where K is a kernel function in d dimension supported on $[-1/2, 1/2]^d$, i.e., $\int K(x)dx$ and $\max_x K(x)$ are both bounded, K is 1-Lipschitz. So $\phi_k(x)$ is supported on $[thx^{(k)} - 1/2th, thx^{(k)} + 1/2th]$. Let $\Omega = \{\omega = (\omega_1, \ldots, \omega_{u^d}), \omega_i \in \{0, 1\}\}$, and

$$\mathcal{E} = \left\{ f_\omega(x) = \frac{L}{2} x_1 + \sum_{i=1}^{k} \omega_i \phi_k(x), x \in [0, 1]^d \right\}.$$

Functions in $\mathcal{E}$ are L-Lipschitz. The function value is linear in x_1 for $x \notin [0, t]^d$ in all functions in $\mathcal{E}$. Consider the comparison function $Z(x, x') = I(x_1 < x'_1)$ that ranks x according to the first dimension. Since K is 1-Lipschitz, it only makes an error when both x, x' lies in $[0, t]^d$, and both x_1, x'_1 lie in the same grid segment $[tk/u, t(k+1)/u]$ for some $k \in [u]$. So the error is at most $t^{2d}(1/u)^2 \cdot u \leq \nu$ for any function $f \in \mathcal{E}$. Thus, if there exists one estimator with $\sup_f \mathbb{E}[(f - \widehat{f})^2] < C \min\{\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}}, 1\} m^{-\frac{2}{d+2}}$, then we can obtain one estimator for functions in $\mathcal{E}$ by using $\widehat{f}$ on $\mathcal{E}$, and responding to all comparisons and rankings as $Z(x, x') = I(x_1 < x'_1)$. So a lower bound on learning $\mathcal{E}$ is also a lower bound on learning any $f \in \mathcal{F}_{s,L}$ with ν-agnostic comparisons.

Now we show that $Ct^{d+2}h^2 = C \min\{\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}}, 1\} m^{-\frac{2}{d+2}}$ is a lower bound to approximate functions in $\mathcal{E}$. For all $\omega, \omega' \in \Omega$ we have

$$\begin{aligned}
\mathbb{E}[(f_\omega - f_{\omega'})^2]^{1/2} &= \left(\sum_{k=1}^{p^d} (\omega_k - \omega'_k)^2 \int \phi_k^2(x) dx \right)^{1/2} \\
&= \left(\rho(\omega, \omega') L^2 t^{d+2} h^{d+2} \right)^{1/2} \\
&= L(th)^{\frac{d+2}{2}} \|K\|_2 \sqrt{\rho(\omega, \omega')},
\end{aligned}$$

where $\rho(\omega, \omega')$ denotes the Hamming distance between x and x'.

By the Varshamov-Gilbert lemma, we can have a $M = O(2^{u^d/8})$ subset $\Omega' = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M)}\}$ of Ω such that the distance between each element $\omega^{(i)}, \omega^{(j)}$ is at least $u^d/8$. So

$d(\theta_i, \theta_j) \geq h^s t^{(d+2)/2}$. Now for P_j, P_0 (P_0 corresponds to f_ω when $\omega = (0, 0, \ldots, 0)$) we have

$$\begin{aligned} KL(P_j, P_0) &= m \int_{\mathcal{X}} p(x) \int_{\dagger} p_j(y|x) \log \frac{p_0(y|x)}{p_j(y|x)} dy dx \\ &= m \int_{\mathcal{X}} p(x) \sum_{i=1}^{u^d} \omega_i^{(j)} \phi^2_{\omega^{(j)}}(x) \\ &\leq m \cdot C h^{d+2} t^{d+2} u^d = C u^d t^{d+2}. \end{aligned}$$

We have $Cu^d t^{d+2} \leq cu^d \leq \alpha \log M$ (since $t \leq 1$), so again using Theorem 2.5 in Tsybakov (2008) we obtain a lower bound of $d(\theta_i, \theta_j)^2 = Ch^2 t^{d+2} = \min\{\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}}, 1\} m^{-\frac{2}{d+2}}$. ■

Now we can prove Theorem 6.

Proof [Proof of Theorem 6] We only need to show

$$\min\{\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}}, 1\} m^{-\frac{2}{d+2}} \geq \min\{\nu^2, m^{-\frac{2}{d+2}}\}. \tag{21}$$

If $\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}} \geq 1$, we have $\nu^2 \geq m^{-\frac{2}{d+2}}$. In this case both sides of (21) equals $m^{-\frac{2}{d+2}}$. If $\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}} \leq 1$, we have $m \leq \nu^{-(d+2)}$, and thus LHS of (21) have term $\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}} m^{-\frac{2}{d+2}} \geq \nu^2$, which equals RHS. ■

B.6. Proof of Theorem 9

Proof We first list some properties of log-concave distributions:

Theorem 22 (Lovász and Vempala (2007); Awasthi et al. (2014)) *The following statements hold for an isotropic log-concave distribution $\mathbb{P}_X$:*

1. *Projections of $\mathbb{P}_X$ onto subspaces of $\mathbb{R}^d$ are isotropic log-concave.*
2. *$\mathbb{P}[\|X\|_2 \geq \alpha\sqrt{d}] \leq e^{1-\alpha}$.*
3. *There is an absolute constant C such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $C\|v - u\|_2 \leq \mathbb{P}(sign(u \cdot X) \neq sign(v \cdot X))$.*

From property of $\mathcal{A}^c$ and point 3 in Theorem 22 we can get $\|\widehat{v} - v^*\|_2 \leq C\varepsilon_{\mathcal{A}^c}(n, \delta/4)$ using n comparisons, with probability $1 - \delta/4$. We use a shorthand notion $\varepsilon := C\varepsilon_{\mathcal{A}^c}(n, \delta/4)$ for this error. Now consider estimating r^*. The following discussion is conditioned on a fixed $\widehat{v}$ that satisfies $\|\widehat{v} - v^*\|_2 \leq \varepsilon$. For simplicity let $T_i = \langle \widehat{v}, X \rangle_i$. We have

$$\begin{aligned} \widehat{r} &= \frac{\sum_{i=1}^m T_i y_i}{\sum_{i=1}^m T_i^2} \\ &= \frac{\sum_{i=1}^m T_i r^* \langle v^*, X_i \rangle + T_i \varepsilon_i}{\sum_{i=1}^m T_i^2} \\ &= r^* + \frac{\sum_{i=1}^m T_i r^* \langle v^* - \widehat{v}, X_i \rangle + T_i \varepsilon_i}{\sum_{i=1}^m T_i^2}. \end{aligned}$$

Now we have

$$\begin{aligned}\langle w^* - \widehat{w},\, X\rangle &= r^*\langle v^*,\, X\rangle - \widehat{r}\langle \widehat{v},\, X\rangle \\ &= r^*\langle v^* - \widehat{v},\, X\rangle - \frac{\sum_{i=1}^m T_i r^*\langle v^* - \widehat{v},\, X_i\rangle + T_i\varepsilon_i}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle.\end{aligned}$$

So

$$\begin{aligned}&\mathbb{E}\left[\langle w^* - \widehat{w},\, X\rangle^2\right] \\ &\leq 3\mathbb{E}\left[(r^*\langle v^* - \widehat{v},\, X\rangle)^2\right] + 3\mathbb{E}\left[\left(\frac{\sum_{i=1}^m T_i r^*\langle v^* - \widehat{v},\, X_i\rangle}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle\right)^2\right] + 3\left[\left(\frac{\sum_{i=1}^m T_i\varepsilon_i}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle\right)^2\right]\end{aligned} \tag{22}$$

The first term can be bounded by

$$(r^*)^2\mathbb{E}[\langle \widehat{v} - v^*,\, X\rangle^2] = (r^*)^2\|\widehat{v} - v^*\|_2^2 \leq (r^*)^2\varepsilon^2.$$

For the latter two terms, we first bound the denominator $\sum_{i=1}^m T_i^2$ using Hoeffding's inequality. Firstly since $\|\widehat{v}\|_2 = 1$, from point 1 in Theorem 22, each T_i is also isotropic log-concave. Now using point 2 in Theorem 22 with $\alpha = 1 - \log(\delta/(4em))$ we get that with probability $1 - \delta/4$, $T_i \leq \log(4em/\delta)$ for all $i \in \{1, 2, \ldots, m\}$. Let E_δ^T denote this event, and $\mathbb{P}'_X$ is the distribution of X such that $T_i \leq \log(4em/\delta)$. Now using Hoeffding's inequality, under E_δ^T for any $t > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^m T_i^2 - \mathbb{E}_{\mathbb{P}'_X}[\langle \widehat{v},\, X\rangle^2]\right| \geq t\right] \leq \exp\left(-\frac{2mt^2}{\log^2(4em/\delta)}\right).$$

Note that $\mathbb{E}_{\mathbb{P}'_X}[\langle \widehat{v},\, X\rangle^2] \leq \mathbb{E}_{\mathbb{P}_X}[\langle \widehat{v},\, X\rangle^2] = 1$. Also we have

$$\begin{aligned}1 = \mathbb{E}[T_i^2] &\leq \mathbb{E}_{\mathbb{P}'_X}[T_i^2] + \int_{\log^2(4em/\delta)}^{+\infty} t\mathbb{P}[T_i^2 \geq t]dt \\ &\leq \mathbb{E}_{\mathbb{P}'_X}[T_i^2] + \int_{\log^2(4em/\delta)}^{+\infty} te^{-\sqrt{t}+1}dt \\ &\leq \mathbb{E}_{\mathbb{P}'_X}[T_i^2] + \frac{3\delta}{2m}\end{aligned}$$

Let $t = \frac{1}{4}\mathbb{E}_{\mathbb{P}'_X}[\langle \widehat{v},\, X\rangle^2]$, we have

$$\sum_{i=1}^m T_i^2 \leq \left[\frac{3m}{4}\mathbb{E}_{\mathbb{P}'_X}[\langle \widehat{v},\, X\rangle^2], \frac{5m}{4}\mathbb{E}_{\mathbb{P}'_X}[\langle \widehat{v},\, X\rangle^2]\right] \subseteq [m/2, 2m]. \tag{23}$$

with probability $1 - \delta/4$, when $m = \Omega(\log^3(1/\delta))$.

Let $\mathcal{E}_\delta$ denote the event when $m/2 \le \sum_{i=1}^m T_i^2 \le 2m$ and T_i are bounded by $\log(4em/\delta)$ for all i . Condition on $\mathcal{E}_\delta$ for the second term in (22) we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\sum_{i=1}^m T_i r^*\langle v^* - \widehat{v},\, X_i\rangle}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle\right)^2\right] &\le \frac{\mathbb{E}\left[\left(\sum_{i=1}^m T_i r^*\langle v^* - \widehat{v},\, X_i\rangle\right)^2\right]}{\frac{m^2}{4}}\mathbb{E}[(\langle \widehat{v},\, X\rangle)^2] \\ &= \frac{4\mathbb{E}\left[\left(\sum_{i=1}^m T_i r^*\langle v^* - \widehat{v},\, X_i\rangle\right)^2\right]}{m^2}\end{aligned}$$

Now notice that $\frac{\widehat{v}-v^*}{\|\widehat{v}-v^*\|_2}X$ is also isotropic log-concave; using point 2 in Theorem 22 we have with probability $1-\delta/4$, $(\widehat{v}-v^*)^T X_i \le \|\widehat{v}-v^*\|_2 \log(4em/\delta)$ for all $i \in \{1,2,\ldots,m\}$. So

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^m T_i r^*\langle v^* - \widehat{v},\, X_i\rangle\right)^2\right] &\le (r^*)^2\varepsilon^2\log^2(4em/\delta)\mathbb{E}\left[\left(\sum_{i=1}^m |T_i|\right)^2\right] \\ &\le (r^*)^2\varepsilon^2\log^2(4em/\delta)\mathbb{E}\left[m\sum_{i=1}^m T_i^2\right] \\ &= (r^*)^2\varepsilon^2\log^2(4em/\delta)m^2\end{aligned}$$

For the third term in (22), also conditioning on $\mathcal{E}_\delta$ we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\sum_{i=1}^m T_i\varepsilon_i}{\sum_{i=1}^m T_i^2}\langle \widehat{v},\, X\rangle\right)^2\right] &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^m T_i\varepsilon_i}{\sum_{i=1}^m T_i^2}\right)^2\right]\mathbb{E}\left[\langle \widehat{v},\, X\rangle^2\right] \\ &\le \frac{\mathbb{E}\left[\left(\sum_{i=1}^m T_i\varepsilon_i\right)^2\right]}{\frac{m^2}{4}} \\ &\le \frac{4\mathbb{E}\left[\sum_{i=1}^m T_i^2\sigma^2\right]}{m^2} = \frac{4\sigma^2}{m}.\end{aligned}$$

Combining the three terms and considering all the conditioned events, we have

$$\begin{aligned}\mathbb{E}\left[(\langle w^*,\, X\rangle - \langle \widehat{w},\, X\rangle)^2\right] \le& 4(r^*)^2\varepsilon^2 + (r^*)^2\varepsilon^2\log^2(4em/\delta) + \frac{4\sigma^2}{m} + C'\delta \\ \le& O\left(\frac{1}{m} + \log^2(m/\delta)\varepsilon_{\mathcal{A}^c}(n,\delta/4) + \nu^2 + \delta\right)\end{aligned}$$

Taking $\delta = \frac{4}{m}$ we obtain the desired result.

■

B.7. Proof of Theorem 11

Our proof ideas come from Castro and Nowak (2008). We use Le Cam's method, explained in the lemma below:

Lemma 23 (Theorem 2.2, Tsybakov (2008)) *Suppose $\mathcal{P}$ is a set of distributions parametrized by $\theta \in \Theta$. $P_0, P_1 \in \mathcal{P}$ are two distributions, parametrized by θ_0, θ_1 respectively, and $KL(P_1, P_0) \leq \alpha \leq \infty$. Let d be a semi-distance on Θ, and $d(\theta_0, \theta_1) = 2a$. Then for any estimator $\widehat{\theta}$ we have*

$$\begin{aligned}\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}[d(\widehat{\theta}, \theta) \geq a] &\geq \inf_{\widehat{\theta}} \sup_{j \in \{0,1\}} \mathbb{P}[d(\widehat{\theta}, \theta_j) \geq a] \\ &\geq \max\left(\frac{1}{4}\exp(-\alpha), -\frac{1-\sqrt{\alpha/2}}{2}\right)\end{aligned}$$

We consider two functions: $w_0^* = (\xi, 0, 0, \ldots, 0)^T$ and $w_1^* = (\frac{1}{\sqrt{m}}, 0, 0, \ldots, 0)^T$, where ξ is a very small constant. Note that for these two functions comparisons provide no information about the weights (comparisons can be carried out directly by comparing $x^{(1)}$, the first dimension of x). So differentiating w_0^* and w_1^* using two oracles is the same as that using only active labels. We have $d(w_0^*, w_1^*) = E\left[\left((w_0^* - w_1^*)^T X\right)^2\right] = \left(\frac{1}{\sqrt{m}} - \xi\right)^2$. For any estimator $\widehat{w}$, let $\{(X_i, y_i)\}_{i=1}^m$ be the set of samples and labels obtained by $\widehat{w}$. Note that X_{j+1} might depend on $\{(X_i, y_i)\}_{i=1}^j$. Now for the KL divergence we have

$$\begin{aligned}\mathrm{KL}(P_1, P_0) &= \mathbb{E}_{P_1}\left[\log \frac{P_1\left(\{(X_i, y_i)\}_{i=1}^m\right)}{P_0\left(\{(X_i, y_i)\}_{i=1}^m\right)}\right] \\ &= \mathbb{E}_{P_1}\left[\log \frac{\prod_{j=1}^m P_1(Y_j|X_j) P(X_j|\{(X_i, y_i)\}_{i=1}^j)}{\prod_{j=1}^m P_1(Y_j|X_j) P(X_j|\{(X_i, y_i)\}_{i=1}^j)}\right] \\ &= \mathbb{E}_{P_1}\left[\log \frac{\prod_{j=1}^m P_1(y_j|X_j)}{\prod_{j=1}^m P_0(y_j|X_j)}\right] \\ &= \sum_{i=1}^m \mathbb{E}_{P_1}\left[\mathbb{E}_{P_1}\left[\log \frac{\prod_{j=1}^m P_1(y_j|X_j)}{\prod_{j=1}^m P_0(y_j|X_j)}\middle| X_1, \ldots, X_m\right]\right] \\ &\leq n \max_x \mathbb{E}_{P_1}\left[\log \frac{\prod_{j=1}^m P_1(y_j|X_j)}{\prod_{j=1}^m P_0(y_j|X_j)}\middle| X_1 = x\right].\end{aligned}$$

The third equality is because decision of X_j is independent of the underlying function giving previous samples. Note that given $X = x$, y is normally distributed; by basic properties of Gaussian we have

$$\mathbb{E}_{P_1}\left[\log \frac{\prod_{j=1}^m P_1(y_j|X_j)}{\prod_{j=1}^m P_0(y_j|X_j)}\middle| X_1 = x\right] = \frac{(\frac{1}{\sqrt{m}-\xi})^2}{2\sigma^2}.$$

Now by taking ξ sufficiently small we have for some constants C_1, C_2,

$$\mathrm{KL}(P_1, P_0) \leq C_1, d(\theta_0, \theta_1) \geq \frac{C_2}{m}.$$

Combining with Lemma 23 we obtain the lower bound.

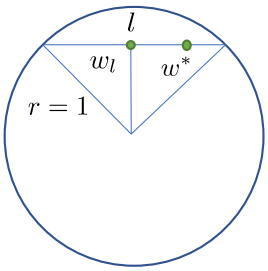


Figure 12: Graphic illustration about the sampling process in the proof of Theorem 12.

B.8. Proof of Theorem 12

Proof We just prove the theorem for $2n + m = d - 1$. Note that this case can be simulated by considering an estimator with access to $2n + m$ *truthful* samples; that is, $Y_i = w^* \cdot X_i$ for $i = 1, 2, \ldots, 2n + m$. In this way truthful comparisons can be simulated by query two labels. We now prove a lower bound for this new case with $2n + m = d - 1$ truthful samples.

We randomly sample w^* as below: first uniformly sample v^* on the surface of $B(0, 1)$, and then uniformly sample $r^* \in [0, 1]$. Let this distribution be $\mathcal{P}_{w^*}$. Since we only have $d-1$ labels, for any set of samples $x_1, \ldots, x_{d-1}$ there exists a line $l \in B(0, 1)$ such that every $w \in l$ produces the same labels on all the samples. Not losing generality, suppose such l is unique (if not, we can augment $\widehat{w}$ such that it always queries until l is unique). Now for any active estimator $\widehat{w}$, let $X_1, \ldots, X_{d-1}$ denote the sequence of queried points when w^* is randomly picked as above. Now note that for every w,

$$\mathbb{P}[w^* = w | \{X_i, y_i\}_{i=1}^{d-1}, l] = \mathbb{P}[w^* = w | \{X_i, y_i\}_{i=1}^{d-1}] \propto \mathbb{P}[w^* = w] I(w \in l).$$

The first equality is because l is a function of $\{X_i, y_i\}_{i=1}^{d-1}$; the second statement is because all $w \in l$ produces the same dataset on $X_1, \ldots, X_{d-1}$, and every $w \notin l$ is impossible given the dataset. Notice that r^* is uniform on $[0, 1]$; so with probability at least a half, the resulting l has distance less than $1/2$ to the origin (since l contains w^*, and $\|w^*\|$ is uniform on $[0, 1]$). Denote by w_l the middle point of l (see Figure 12). For any such line l, the error is minimized by predicting the middle point of l: Actually we have

$$\mathbb{E}\left[\langle w^* - \widehat{w}, X\rangle^2 | l \text{ has distance less than } 1/2\right] \geq \int_{u=0}^{|l|/2} u^2 dP(\|w^* - \widehat{w}\|_2 \geq u | w^* \in l) \quad (24)$$

$$\geq \int_{u=0}^{|l|/2} u^2 dP(\|w^* - w_l\|_2 \geq u | w^* \in l) \quad (25)$$

Note that the distribution of $w^* \in l$ is equivalent to that we sample from the circle containing l and centered at origin, and then condition on $w^* \in l$ (see Figure 12). Notice that this sampling process is the same as when $d = 2$; and with some routine calculation we can show that (25) is a constant C. So overall we have

$$\mathbb{E}\left[\langle w^* - \widehat{w}, X\rangle^2\right] \geq \frac{1}{2} C,$$

where the expectation is taken over randomness of w^* and randomness of $\widehat{w}$. Now since the expectation is a constant, there must exists some w such that

$$\mathbb{E}\left[\langle w^* - \widehat{w}, X\rangle^2 \middle| w^* = w\right] \geq \frac{1}{2}C,$$

which proves the theorem. ■

Appendix C. Auxiliary Technical Results

We use some well-known technical results in our proofs and collect them here to improve readability. We use the Craig-Bernstein inequality (Craig, 1933):

Lemma 24 *Suppose we have $\{X_1, \ldots, X_n\}$ be independent random variables and suppose that for $k \geq 2$, for some $r > 0$,*

$$\mathbb{E}[|X_i - \mathbb{E}[X_i]|^k] \leq \frac{var(X_i)}{2} k! r^{k-2}.$$

Then with probability at least $1 - \delta$, for any $c < 1$ and for any $t \leq c/r$ we have that:

$$\frac{1}{n}\sum_{i=1}^{n} (\mathbb{E}[X_i] - X_i) \leq \frac{\log(1/\delta)}{nt} + \frac{t\ var(X_i)}{2(1-c)}.$$