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

Yichong Xu [1] **Hariank Muthakana** [1] **Sivaraman Balakrishnan** [2] **Artur Dubrawski** [3] **Aarti Singh** [1]

Abstract

In supervised learning, we leverage a labeled dataset to design methods for function estimation. 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. We focus on a semi-supervised setting where we obtain additional *ordinal (or comparison) information* for potentially 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 nonparametric 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 develop an algorithm called Ranking-Regression (R^2) and analyze its accuracy as a function of size of the labeled and unlabeled datasets and various noise parameters. We also present lower bounds, that establish fundamental limits for the task and show that R^2 is optimal in a variety of settings. Finally, we present experiments that show the efficacy of R^2 and investigate its robustness to various sources of noise and model-misspecification.

1. Introduction

Classical nonparametric regression is centered around the development and analysis of methods that use labeled observations, $\{(X_1, y_1), \ldots, (X_n, y_n)\}$, where $(X_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, in various tasks of estimation and inference. Nonparametric 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 nonparametric methods typically require a large amount of labeled data, scaling exponentially with the dimension, to estimate complex target functions – 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.

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 & Gupta, 2011; Shah et al., 2015)). We investigate a task of this flavor in Section 6. 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 quality landscape 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 provide noisy comparative assessments of synthesis quality. Similarly, in clinical settings, precise assessment of an individual patient's health readings can be difficult, expensive and/or 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]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA [2]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, USA [3]Auton Lab, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Yichong Xu <yichongx@cs.cmu.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Related Work: There is considerable work in supervised and unsupervised learning on incorporating additional types of feedback beyond labels. For instance, the papers (Zou et al., 2015) and (Poulis & Dasgupta, 2017) study the benefits of different types "feature feedback" in clustering and supervised learning respectively. There is also a vast literature on models and methods for analyzing pairwise comparison data, like the classical Bradley-Terry (Bradley & Terry, 1952) and Thurstone (Thurstone, 1927) models. In this literature, 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 work on "noisy sorting" (Braverman & Mossel, 2009) to extract a consensus ranking from noisy pairwise comparisons. Most close in spirit to our own work are the two recent papers (Kane et al., 2017; Xu et al., 2017), which consider binary classification with ordinal information. These works differ from ours in their focus on classification, emphasis on active querying strategies and use of quite different ordinal feedback models. Finally, 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 algorithms from this literature in our work and we comment further on the connections in Section 3. Some salient differences between this literature and our work are that we design methods that work in a semi-supervised setting, and further that our target is an unknown d-dimensional (smooth) regression function as opposed to a univariate shape-constrained function.

Our Contributions: We develop the Ranking-Regression (R^2) algorithm for nonparametric regression that can leverage ordinal information, in addition to direct labels. Theoretical analysis and practical experiments show the strength of our algorithm.

- To establish the usefulness of ordinal information in nonparametric regression, in Section 3 we consider the idealized setting where we obtain a perfect ordering of the unlabeled set. We show that the Mean Squared Error (MSE) of R^2 can be bounded by $\widetilde{O}(m^{-2/3} + n^{-2/d})$[1], 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 dimension. This result establishes that sufficient ordinal information of high quality can allow us to effectively circumvent the curse of dimensionality.
- In Section 4 we analyze R^2 when using a noisy ranking. We show that the MSE is bounded as $\widetilde{O}(m^{-2/3} + \sqrt{\nu} + n^{-2/d})$, where ν is the Kendall-Tau distance between the true and noisy ranking.
- As a corollary, we develop results for R^2 using pairwise comparisons. If the comparison noise is bounded, the R^2 algorithm can be combined with algorithms for ranking from pairwise comparisons (Braverman & Mossel, 2009) to obtain an MSE of $\widetilde{O}(m^{-2/3} + n^{-2/d})$ when $d \geq 4$.
- We give information-theoretic lower bounds to characterize the fundamental limits of combining ordinal and standard supervision. These lower bounds show that our algorithms are almost optimal. In particular, the R^2 algorithm under perfect ranking, as well as under bounded noise comparisons, is optimal up to log factors.
- In our experiments, we test R^2 on simulated data, on UCI datasets and on various age-estimation tasks. Our experimental results show the advantage of R^2 over algorithms that only use labeled data when this labeled data is scarce. Our experiments with the age-estimation data also show the practicality of R^2.

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

2. Background and Problem Setup

We consider a non-parametric 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 $\mathbb{P}_{\mathcal{X}}$. We assume that $\mathbb{P}_{\mathcal{X}}$ has a density $p(x)$ which is upper and lower bounded as $0 < p_{\min} \leq p(x) \leq p_{\max}$ for $x \in [0,1]^d$. Our goal is to estimate a function $f : \mathcal{X} \mapsto \mathbb{R}$, where following classical work (Györfi et al., 2006; Tsybakov, 2009) we assume that f is bounded in $[-M, M]$ and belongs to a Hölder ball $\mathcal{F}_{s,L}$, with $0 < s \leq 1$ where:

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

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 7. 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:

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

where ϵ_i are i.i.d. $\mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2$. 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 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:

$$\sum_{i,j\in[n]} \mathbb{I}[(\pi(i)-\pi(j))(\widehat{\pi}(i)-\widehat{\pi}(j))<0] \le \nu n^2. \tag{1}$$

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

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

As we discuss in Section 5 it is straightforward to extend our results to a setting where only a randomly chosen subset of all pairwise comparisons are observed.

Although classical supervised learning estimates a regression function with labels only and without ordinal supervision, we note that we cannot consistently estimate the underlying function with only ordinal supervision and without direct observations. In the case when no direct measurements are available the underlying function is only identifiable up to certain monotonic transformations.

Our goal is to estimate f, and the quality of an estimate $\widehat{f}$ is assessed using the mean squared error $\mathbb{E}(\widehat{f}(X) - f(X))^2$, where the expectation is taken over the labeled and unlabeled training samples, as well as the 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 (the Hölder parameters s, L and various noise parameters), the minimax risk:

$$\mathfrak{M}(m,n;\eta) = \inf_{\widehat{f}} \sup_{f\in\mathcal{F}_{s,L}} \mathbb{E}(\widehat{f}(X) - f(X))^2, \tag{3}$$

provides an information-theoretic benchmark to assess the performance of an estimator. 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.

3. Nonparametric Regression with Perfect Ranking

To establish 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 3.1, and a lower bound in Section 3.2 which shows that R^2 is optimal up to log factors.

Algorithm 1 R^2: Ranking-Regression

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

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

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 $i = 1, 2, ..., n$, let $\widetilde{i} = t_k$, where $\widehat{\pi}(t_k)$ is the largest value such that $\widehat{\pi}(t_k) \le \widehat{\pi}(i), k = 0, 1, ..., m$, and $\widetilde{i} = \star$ if no such t_k exists. Set

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

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

3.1. Upper bounds for the R^2 Algorithm

Our non-parametric 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:

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

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:

Theorem 1. *For constants $C_1, C_2 > 0$ the MSE of $\widehat{f}$ is bounded by*

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

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).

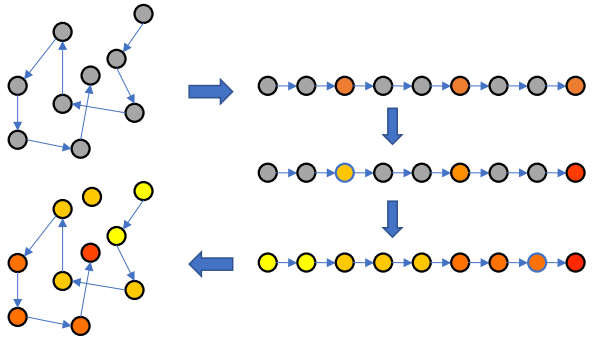


Figure 1. Top Left: A group of unlabeled points are ranked according to function values using ordinal information only. Top Right: We obtain function values of m randomly chosen samples. Middle Right: The values are adjusted using isotonic regression. Bottom Right: Function values of other unlabeled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

The MSE of nonparametric regression using only the labeled samples is $\Theta(m^{-\frac{2s}{2s+d}})$ which is exponential in d and makes non-parametric 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 in d. 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 3.2 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.
(4) In the case where all points are labeled (i.e., $m = n$), the MSE is of order $n^{-2/3} + n^{-2s/d}$, again improving slightly on the rate when no ordinal information is available. The improvement is largest when $m \ll n$.
(5) Finally, we also note in passing that the above theorem provides an upper bound on the minimax risk in (3).

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

$$\mathbb{E}\left[(\widehat{f}(X) - f(X))^2\right] \leq 2\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] + 2\mathbb{E}\left[(f(X_\alpha) - f(X))^2\right]. \quad (5)$$

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 9). Without loss of generality suppose $f(X_{t_1}) \leq \cdots f(X_{t_m})$. By conditioning on a probable configuration of the points and enumerating over choices of the nearest neighbor we find that roughly (see Lemma 9 for a precise statement):

$$\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] \leq \left(\frac{\log^2 n \log m}{m}\right) \times \mathbb{E}\Big(\sum_{k=1}^{m}\Big(\big(\widehat{f}(X_{t_k}) - f(X_{t_k})\big)^2 + \big(f(X_{t_{k+1}}) - f(X_{t_k})\big)^2\Big)\Big). \quad (6)$$

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 $m^{-2/3}$. Putting these pieces together we obtain the theorem. □

3.2. 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 $\mathbb{P}_\mathcal{X}$ 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:

Theorem 2. *For any estimator $\widehat{f}$ we have that 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

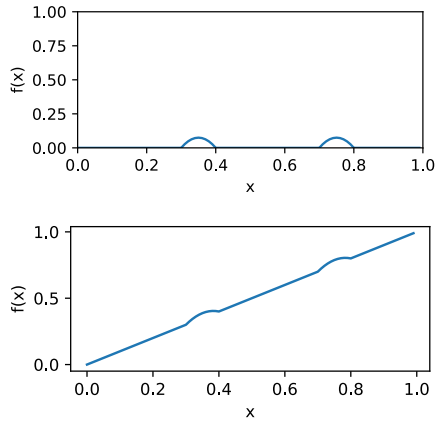


Figure 2. Original construction for nonparametric regression in 1-d (above), and our construction (below).

bounds in non-parametric 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 the Appendix.

□

4. 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 (1) 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 4.2.

4.1. 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 (1), in the following theorem:

Theorem 3. *For constants $C_1, C_2 > 0$, 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. 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.

(3) Finally, in settings where ν is large R^2 can be led astray by the ordinal information, and a standard non-parametric regressor can achieve the (possibly faster) $O\big(m^{-\frac{2s}{2s+d}}\big)$ rate by ignoring the ordinal information. As we show in Appendix E a simple cross-validation procedure can combine the benefits of the two estimators to achieve a rate of $\widetilde{O}\big(m^{-2/3} + \min\{\sqrt{\nu}, m^{-\frac{2s}{2s+d}}\} + n^{-2s/d}\big)$. This rate can converge to 0 if we have sufficiently many labels, even if the comparisons are very noisy. The cross validation process is standard and computationally efficient: 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.

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* (1)

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

Using this result we bound the minimal error of approximating an increasing sequence according to π by an increasing sequence according to the estimated ranking $\widehat{\pi}$. We denote this error 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 (6). 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 Equation (6) we obtain the theorem.

□

4.2. 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 $\mathbb{P}_{\mathcal{X}}$:

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

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 3.2).

Theorem 5. *There is a constant $C > 0$ such that for any estimator $\widehat{f}$ taking input $X_1, ..., X_n, y_1, ..., y_m$ and $\widehat{\pi}$,*

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

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 non-parametric 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. □

5. 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 (2). 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 & Mossel (2009) study noisy sorting algorithms under the noise model described in (2) and establish the following result:

Theorem 6 ((Braverman & 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 7. *For constants $C_1, C_2 > 0$, R^2 with $\widehat{\pi}$ estimated as described above produces an estimator $\widehat{f}$ with MSE*

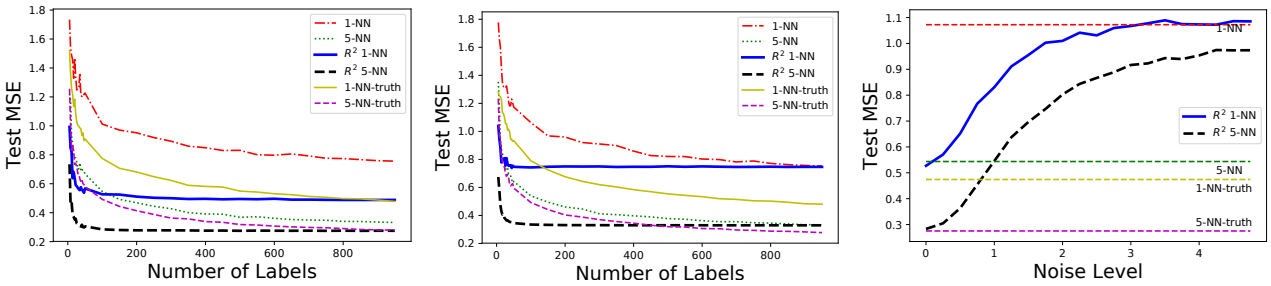
$$\mathbb{E}\big(\widehat{f}(X) - f(X)\big)^2 \leq C_1 m^{-2/3} \log^2 n \log m + C_2 \max\{n^{-2s/d}, n^{-1/2} \log^2 n \log m\}.$$

Remarks: (1) From a technical standpoint this result is an immediate corollary of Theorems 3 and 6, 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 Theorem 7 is also optimal up to log factors (following the lower bound in Section 3.2).

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

6. Experiments & Simulations

To verify our theoretical results and test R^2 in practice, we perform three sets of experiments. First, we conduct experiments on simulated data, where the noise in the labels and ranking can be controlled separately. Second, we test R^2 on UCI datasets, where the rankings are simulated using labels. We present these results in Appendix A. Finally, we



(a) R^2 with perfect ranking (b) R^2 with noisy ranking (c) Effect of ranking noise

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

consider a practical application of predicting people's age from portraits and we test R^2 on two realistic estimation tasks.

We compare R^2 with k-NN algorithms in all experiments. We choose k-NN methods because they are near-optimal theoretically, and are widely used in practice. Theoretical guidelines suggest using the tuning parameter $k_m = m^{\frac{2}{d+2}}$ when we have access to m labeled samples; however for all m, d values we considered, $m^{\frac{2}{d+2}}$ is very small (< 5). Instead we choose a range of different constant values of k (that do not change with m) in our experiments. We repeat each experiment 20 times and report the average MSE[2].

6.1. Simulated Data

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

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

with p sampled uniformly random in $[0, 10]$. We rescale $f(x)$ so that it has 0 mean and unit variance. The labels are generated as $y = f(x) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. We generate a training and a test set of $n = 1000$ samples respectively. 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 all test data $X_1^{\text{test}}, ..., X_n^{\text{test}}$.

Variants of R^2. We consider two variants of R^2. The first variant is exactly Algorithm 1 using 1-NN as the final estimator; the second variant uses 5-NN as the final estimator. Using 5-NN does not change the asymptotic behavior of our bounds. However, we find that using 5-NN improves our estimator empirically.

Baselines. We compare R^2 with the following baselines: i) 1-NN and 5-NN using noisy labels (x, y). 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 $m = n$ they serve as a benchmark for our algorithms.

R^2 with perfect rankings. In our first experiment, R^2 had access to the ranking over all 1000 training samples, while the k-NN baseline algorithms only had access to labeled samples. We varied the number of labeled samples for all algorithms from $m = 5$ to $m = 1000$. The results are depicted in Figure 3(a). 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, R^2 1-NN and R^2 5-NN performed nearly the same as 1-NN or 5-NN using all 1000 perfect labels, while only requiring around 50 labeled samples.

R^2 with noisy rankings. We then consider noisy rankings; particularly in Figure 3(b), the input ranking of R^2 is obtained from noisy labels. This eliminates the need for isotonic regression in Algorithm 1, but we find that the ranking still provides useful information for the unlabeled samples. In this setting R^2 outperformed the 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 $n = m$.

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

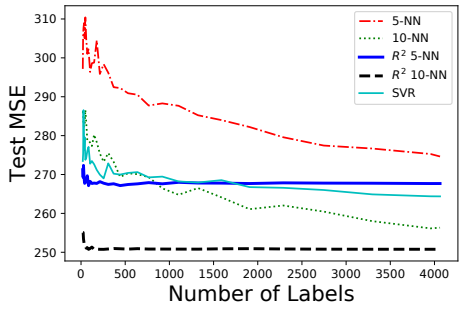
$$y' = f(x) + \varepsilon'$$

where $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$. We varied σ from 0 to 5 and plotted the MSE. As σ goes up, the error of both variants of the R^2 algorithm increases as expected.

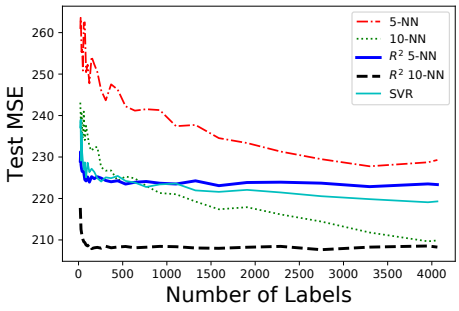
6.2. Predicting Ages from Portraits

To further validate R^2 in practice, we consider the task of estimating people's age from portraits. We use the APPA-

[2]Our plots are best viewed in color.



(a) Task 1



(b) Task 2

Figure 4. Experiments on age prediction.

REAL dataset (E Agustsson, 2017), which contains 7,591 images, where each image is associated a biological age and an apparent age. The biological age is the person's actual age, whereas the apparent ages are collected by crowdsourcing. 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 4113 train, 1500 validation and 1978 test samples, and we only use the train and validation samples for our experiments.

Features and Models. We extract the 128-dim feature for each image using the last layer of FaceNet (Schroff et al., 2015). We rescale the features so that every $X \in [0,1]^d$. We use 5-NN and 10-NN in this experiment. To further show the effectiveness of R^2, we also compare to kernelized support vector regression (SVR). We used the standard parameter configuration in scikit-learn (Pedregosa et al., 2011), using penalty parameter $C = 1$, RBF kernel, and tolerance of 0.1.

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

1. In the first task, the goal is to predict biological age. The labels were biological age, whereas the ranking came from apparent ages. This is motivated by the collection process of most modern datasets where typically, an aggregated (de-noised) label is obtained through majority vote. For example, we may have the truthful biological age 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. So we assume additional access to a ranking that comes from apparent ages.

2. In the second task, the goal is to predict the apparent age. Both labels and ranking were generated using the standard deviation provided in APPA-REAL. Labels were generated according to a Gaussian distribution with mean equal to the apparent age, and standard deviation provided in the dataset. The ranking was generated by first generating a sample of all labels using the same distribution, and ranking according to the sample. This resembles the case where we ask one single labeler for each label and comparison, to collect data for more samples. Such policy is also used in e.g., (Bi et al., 2014; Khetan et al., 2017). Note that in real applications, the ranking will have less noise than in our experiment; (Shah et al., 2015) considered exactly the same task, and showed that comparisons are more reliable than labels.

Results are depicted in Figure 4. The 10-NN version of R^2 gave the best overall performance in both tasks. R^2 5-NN and R^2 10-NN both outperformed other algorithms when the number of labeled samples was less than 500. The performance of SVR was between 5-NN and 10-NN in our experiments. 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, indicating the high reliability of the ordinal information for this task.

7. Discussion and Conclusion

We design minimax-optimal algorithms for nonparametric 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 algorithm 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.

In future work motivated by practical applications in crowdsourcing, we hope to address the setting where both direct and ordinal supervision are actively acquired. Another possible direction is to consider *partial orders*, where we have several subsets of unlabeled data ranked, but the relation between these subsets is unknown. It would also be interesting to consider other models for ordinal information and to more broadly understand settings where indirect feedback is beneficial. Also, several recent papers (Bellec & 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 would induce a low-complexity isotonic regression problem is an interesting avenue for future work.

Acknowledgements

This work is supported by AFRL grant FA8750-17-2-0212, NSF CCF-1763734, NSF DMS-1713003 and DARPA award FA8750-17-2-0130.

References

Barlow, R. E. Statistical inference under order restrictions: The theory and application of isotonic regression. Technical report, 1972.

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

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

Bi, W., Wang, L., Kwok, J. T., and Tu, Z. Learning to predict from crowdsourced data. In *Uncertainty in Artificial Intelligence*, pp. 82–91, 2014.

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

Braverman, M. and Mossel, E. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.

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

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

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

E Agustsson, R Timofte, S. E. X. B. I. G. R. R. 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.

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

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. Isotonic regression in general dimensions. *arXiv preprint arXiv:1708.09468*, 2017.

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. *Nonparametric and semiparametric models*. Springer Science & Business Media, 2012.

Kane, D. M., Lovett, S., Moran, S., and Zhang, J. Active classification with comparison queries. *arXiv preprint arXiv:1704.03564*, 2017.

Khetan, A., Lipton, Z. C., and Anandkumar, A. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Poulis, S. and Dasgupta, S. Learning with feature feedback: From theory to practice. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial Intelligence and Statistics*, pp. 856–865, 2015.

Shah, N., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, pp. 11–20, 2016.

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

Tsukida, K. and Gupta, M. R. How to analyze paired comparison data. Technical report, DTIC Document, 2011.

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

Tsybakov, A. B. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

Xu, Y., Zhang, H., Miller, K., Singh, A., and Dubrawski, A. Noise-tolerant interactive learning from pairwise comparisons with near-minimal label complexity. *arXiv preprint arXiv:1704.05820*, 2017.

Xue, D., Balachandran, P. V., Hogden, J., Theiler, J., Xue, D., and Lookman, T. Accelerated search for materials with targeted properties by adaptive design. *Nature Communications*, 7, 2016.

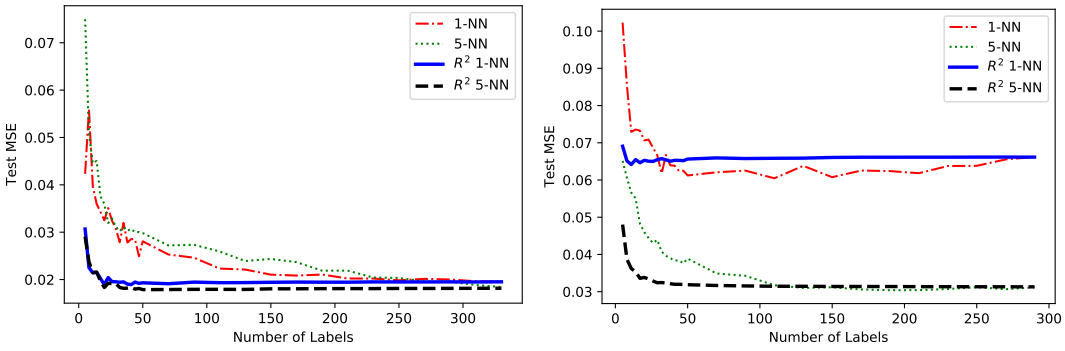
Zhang, C.-H. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.

Zou, J. Y., Chaudhuri, K., and Kalai, A. T. 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.*, pp. 198, 2015.

A. Experiment Results on UCI Datasets

In this section we provide experiment results on two UCI regression datasets, *Boston-Housing* and *diabetes*. Since we do not have "truthful" labels in this case, all rankings were generated directly from the labels in the datasets. For the same reason we only compared R^2 with 1-NN and 5-NN for UCI datasets.

Results on *Boston-housing* and *diabetes* are depicted in Figure 5(a) and 5(b) respectively. Our algorithm both outperformed their label-only counterparts, but the performance gain on *Boston-housing* was larger than that on *diabetes*. This might be because *diabetes* has a higher dimension (13) than *Boston-housing*(10).



(a) Boston Housing (b) Diabetes

Figure 5. Experiments on UCI datasets. Legends follow Figure 3.

B. Proof of Theorem 1

Proof. 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.

For a random point $X \in \mathcal{X}$, let X_α be the nearest neighbor of X in T. 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 8. *For a constant $C > 0$ we have that,*

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

Lemma 9. *For any $\delta > 0$ we have that there is a constant $C > 0$ such that:*

$$\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] \leq \frac{C\log(m/\delta)\log n \log(1/\delta)}{m}\left[\sum_{k=1}^{m}\left(\mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2\right]\right) + \sum_{k=0}^{m}\left(\mathbb{E}\left[(f(X_{t_{k+1}}) - f(X_{t_k}))^2\right]\right)\right] + 4\delta M^2.$$

Taking these results as given we can now complete the proof of the theorem. We first note that the first term in the upper bound in Lemma 9 is simply the MSE in an isotonic regression problem, and using standard risk bound for isotonic regression (see (Zhang, 2002)) we obtain that for a constant $C > 0$:

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

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

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

Now, 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. Proof of Lemma 8

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

Lemma 10 ((Györfi et al., 2006), Lemma 6.4 and Exercise 6.7)**.** *Suppose that there exists 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.2. Proof of Lemma 9

We condition on a certain favorable configuration of the samples that holds with high-probability. Conditioning on the samples $\{X_1, \ldots, X_n\}$ 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 11. *There is a constant $C > 0$ such that with probability at least $1 - \delta$ we have that the following two results hold:*

1. $$\max_{1 \leq j \leq n} q_i \leq \frac{Cd \log(1/\delta) \log n}{n}. \tag{7}$$

2. *Let us take $t_{m+1} := n + 1$, then*

$$\max_{k \in [m+1]} t_k - t_{k-1} \leq \frac{Cn \log(m/\delta)}{m}. \tag{8}$$

Denote the event, which holds with probability at least $1-\delta$ in the above Lemma by E_0. By conditioning on E_0 we obtain the following decomposition:

$$\mathbb{E}\left[(\widehat{f}(X)-f(X_\alpha))^2\right] \leq \mathbb{E}\left[(\widehat{f}(X)-f(X_\alpha))^2|E_0\right]+\delta\cdot 4M^2$$

because both f and $\widehat{f}$ are bounded in $[-M,M]$. We condition all calculations below on E_0 but omit this from the notation. Now we have

$$\begin{aligned}
\mathbb{E}\left[(\widehat{f}(X)-f(X_\alpha))^2\right] &= \sum_{i=1}^n \mathbb{P}[X_\alpha=X_i]\mathbb{E}\left[(\widehat{f}(X_i)-f(X_i))^2\right] \\
&\leq \sum_{i=1}^n \max_{1\leq j\leq n}\mathbb{P}[X_\alpha=X_j]\mathbb{E}\left[(\widehat{y}_i-f(X_i))^2\right] \\
&\leq \frac{Cd\log(1/\delta)\log n}{n}\sum_{i=1}^n \mathbb{E}\left[(\widehat{y}_{\widetilde{i}}-f(X_i))^2\right], \qquad (9)
\end{aligned}$$

where we recall that $\widehat{y}_{\widetilde{i}}$ (defined in Algorithm 1) denotes the de-noised (by isotonic regression) y value at the nearest labeled left-neighbor of the point X_i. For convenience we defined $f(X_{t_0})=0$, then we have that,

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\left[(\widehat{y}_{\widetilde{i}}-f(X_i))^2\right] &\leq \sum_{i=1}^n \left(2\mathbb{E}\left[(\widehat{y}_{\widetilde{i}}-f(X_{\widetilde{i}}))^2\right]+2\mathbb{E}\left[(f(X_i)-f(X_{\widetilde{i}}))^2\right]\right) \\
&= \sum_{i=1}^n\sum_{k=0}^m I[\widetilde{i}=t_k]\left(2\mathbb{E}\left[(\widehat{y}_{t_k}-f(X_{t_k}))^2\right]I[k\neq 0]+\ 2\mathbb{E}\left[(f(X_i)-f(X_{t_k}))^2\right]\right) \\
&\overset{\text{(i)}}{\leq} \sum_{i=1}^n\sum_{k=0}^m I[\widetilde{i}=t_k]\left(2\mathbb{E}\left[(\widehat{y}_{t_k}-f(X_{t_k}))^2\right]I[k\neq 0]+\ 2\mathbb{E}\left[(f(X_{t_{k+1}})-f(X_{t_k}))^2\right]\right) \\
&\overset{\text{(ii)}}{=} \sum_{k=0}^m \left(2\mathbb{E}\left[(\widehat{y}_{t_k}-f(X_{t_k}))^2\right]I[k\neq 0]+2\mathbb{E}\left[(f(X_{t_{k+1}})-f(X_{t_k}))^2\right]\right)\sum_{i=1}^n I[\widetilde{i}=t_k] \\
&\leq \frac{Cn\log(m/\delta)}{m}\sum_{k=1}^m\left(2\mathbb{E}\left[(\widehat{y}_{t_k}-f(X_{t_k}))^2\right]\right)+\frac{Cn\log(m/\delta)}{m}\sum_{k=0}^m\left(2\mathbb{E}\left[(f(X_{t_{k+1}})-f(X_{t_k}))^2\right]\right) \\
&\leq \frac{2Cn\log(m/\delta)}{m}\left[\sum_{k=1}^m\left(\mathbb{E}\left[(\widehat{y}_{t_k}-f(X_{t_k}))^2\right]\right)+\sum_{k=0}^m\left(\mathbb{E}\left[(f(X_{t_{k+1}})-f(X_{t_k}))^2\right]\right)\right].
\end{aligned}$$

The inequality (i) follows by noticing that if $\widetilde{i}=t_k$, $f(X_i)-f(X_{\widetilde{i}})$ is upper bounded by $f(X_{t_{k+1}})-f(X_{t_k})$. We interchange the order of summations to obtain the equation (ii). Plugging this expression back in to (9) we obtain the Lemma. Thus, to complete the proof it only remains to establish the result in Lemma 11.

B.2.1. Proof of Lemma 11

We prove each of the two results in turn.

Proof of inequality (7)**:** As a preliminary, we need the following Vapnik-Cervonenkis result from (Chaudhuri & Dasgupta, 2010):

Lemma 12. *There exists a universal constant C' such that with probability $1-\delta$, every ball B with*

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

contains at least one sample from $T=\{X_1,...,X_n\}$.

We now show that under this event we have

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

Fix any point $X_i \in T$, 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 12 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 (7).

Proof of inequality (8)**:** Notice that $t_1, ..., t_m$ are randomly chosen from $[n]$. So for each $k \in [m]$ we have

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

I.e., we randomly pick t_k in $X_t, X_{t+1}, ..., X_n$, and choose the other $m-1$ samples in $X_1, ..., X_{t_k - t}, X_{t_k+1}, ..., X_n$. Similarly we also have

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

So

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

Let the RHS less than or equal to δ, we have

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

Let $u = \log\left(1 - \left(\frac{\delta}{m+1}\right)^{\frac{1}{m-1}}\right) = -C\frac{\log(m/\delta)}{m}$, we have $1 - e^u = O(-u)$ since u is small and bounded. So it suffices for $t \geq C\frac{n \log(m/\delta)}{m}$ such that

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

C. Proof of Theorem 2

Proof. To prove the result we prove separately lower bounds on the error in terms of m and n in Lemmas 13 and 15 respectively.

Lemma 13. $\inf_{\widehat{f}} \sup_f \mathbb{E}\left[(f(X) - \widehat{f}(X))^2\right] \geq C m^{-\frac{2}{3}}$.

Proof. We consider the $d = 1$ case in the proof, and results in $d = 1$ can be extended for $d > 1$ by using $f(x) = f(x_1)$, where x_1 is the first dimension of x.

Define $u = \lceil m^{\frac{1}{3}} \rceil$, $h = 1/u$, $x_k = \frac{k - 1/2}{u}$, $\phi_k(x) = \frac{L}{2} h K\left(\frac{x - x_k}{h}\right)$, $k = 1, 2, ..., u$, $\Omega = \{\omega = (\omega_1, ..., \omega_u), \omega_i \in \{0, 1\}\}$, where K is a kernel function that is 1-Lipschitz, bounded and supported on $[-1/2, 1/2]$. Consider the following class of functions

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

Functions in $\mathcal{E}$ are L-Lipschitz, and thus satisfy the Hölder constraint.

Since the functions in $\mathcal{E}$ are all increasing, and the distribution $\mathbb{P}_{\mathcal{X}}$ is taken to be uniform two important simplifications arise: since the functions are increasing 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 since their distribution $\mathbb{P}_{\mathcal{X}}$ is known and their ranking is also known. Concretely, for any estimator using $X_1, ..., X_n, y_1, ..., y_m, \pi$ with

$$\sup_f \mathbb{E}\left[(f(X) - \widehat{f}(X))^2\right] < Cm^{-2/3},$$

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 $X_{m+1}, ..., X_n$ uniformly in $[0,1]^d$ and generating π by ranking X in increasing order. Now we show that $Cm^{-2/3}$ is a lower bound to approximate functions in $\mathcal{E}$ with access to only noisy labels. For all $\omega, \omega' \in \Omega$ we have

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

where $\rho(\omega, \omega')$ denotes the hamming distance between x and x'. We use the following theorem from (Tsybakov, 2009):

Theorem 14 ((Tsybakov, 2009), Theorem 2.5)**.** *Suppose $M \geq 2$ and $\Theta = \{\theta_1, ..., \theta_M\}$ such that $d(\theta_j, \theta_k) > 2t > 0, \forall 0 \leq j < k \leq M$. Let P_j be a distribution induced by parameter θ_j for all $j \in [M]$. If P_j is absolute continuous with respect to P_0, and $\frac{1}{M}\sum_{j=1}^{M} KL(P_j, P_0) \leq \alpha \log M$ with $0 < \alpha < 1/8$, then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\theta, \widehat{\theta}) \geq t) \geq C$$

where C is some positive constant.

By the Varshamov-Gilbert bound, if $u > 8$ there exists a $2^{u/8}$ subset $\Omega' = \{\omega^{(0)}, \omega^{(1)}, ..., \omega^{(M)}\}$ of Ω such that the distance between each element $\omega^{(i)}, \omega^{(j)}$ is at least $u/8$. So $d(\theta_i, \theta_j) \geq 1/u = m^{-1/3}$. Now for P_j, P_0 we have

$$\begin{aligned}KL(P_j, P_0) &= m\int_{\mathcal{X}} p(x) \int_{\dagger} p_j(y|x) \log \frac{p_0(y|x)}{p_j(y|x)} dy dx \\ &= m\int_{\mathcal{X}} p(x) \sum_{i=1}^{u} \omega_j^{(i)} \phi^2_{\omega^{(i)}}(x) dx \\ &\leq m \cdot Ch^3 \cdot u = Cm^{1/3} = Cu.\end{aligned}$$

This is by the same process as in the proof of Theorem 2.7 in (Tsybakov, 2009). Since $Cu \leq \alpha \log M$, we apply Theorem 14 to obtain a lower bound of $t^2 = m^{-2/3}$.

□

Lemma 15. *Suppose $n > 2$. Then for any estimator $\widehat{f}$, $\inf_{\widehat{f}} \sup_f \mathbb{E}\left[(f(X) - \widehat{f}(X))^2\right] \geq Cn^{-\frac{2s}{d}}$.*

Proof. We show this lemma by reduction to the case where we have n points with noiseless evaluations, i.e., $X_1, X_2, ..., X_n, f(X_1), ..., f(X_n)$. Suppose $R(\widehat{f})$ is the risk of $\widehat{f}$. Same as in Lemma 13, if there exists an estimator $\widehat{f}$ with access to $X_1, ..., X_n, y_1, ..., y_m, \pi$ such that $R(\widehat{f}) \leq Cn^{-2s/d}$, then we can obtain an estimator with access to $X_1, ..., X_n, f(X_1), ..., f(X_n)$ by the following process: We generate $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $y_i = f(X_i) + \varepsilon_i$, and derive π by ranking X_i according to $f(X_i)$. Then we use $\widehat{f}$ as the estimator. So we can also obtain an estimator with access to $X_1, ..., X_n, f(X_1), ..., f(X_n)$ with risk $Cn^{-2s/d}$. So an lower bound in the new (noiseless label) case is also a lower bound in the label+comparison case.

Now we prove Lemma 15 by showing the lower bound in the following lemma:

Lemma 16. *Suppose $n > 2$. Let $S' = \{X_1, ..., X_n\}$ be n random points from $[0,1]^d$. Then for any estimator $\widehat{f}'$ with access to S' and $f(X_1), ..., f(X_n)$, $\inf_{\widehat{f}'} \sup_f \mathbb{E}_{S'}\left[(f(X) - \widehat{f}(X))^2\right] \geq Cn^{-\frac{2s}{d}}$.*

Proof. Let $h = n^{-1/d}$ and $p = 1/h$. Divide $[0,1]^d$ to a grid of n points, with step size h. Let the grid points be $\gamma^{(1)}, \gamma^{(2)}, ..., \gamma^{(n)} \in [0,1]^d$. Let $x^{(k)} = \frac{\gamma^{(k)} - 1/2}{p}$, and $\phi_k(x) = Lh^s K(\frac{x - x^{(k)}}{h}), k = 1, 2, ..., n$, 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 $[hx^{(k)} - 1/2h, hx^{(k)} + 1/2h]$, and is s-Hölder. Let $\Omega = \{\omega = (\omega_1, ..., \omega_{p^d}), \omega_i \in \{0,1\}\}$, and

$$\mathcal{E} = \left\{ f_\omega(x) = Lx_1 + \sum_{i=1}^{k} \omega_i \phi_k(x), x \in [0,1]^d \right\}.$$

Let $\widehat{f}'$ be any estimator. We now compute $E_{\omega,S'}[(f_w - \widehat{f}')^2]$, where w is randomly chosen from Ω. For any grid point area $[hx^{(k)} - 1/2h, hx^{(k)} + 1/2h]$, the probability of one particular point falling in it is $1/n$; so the probability that no points falls in it is $(1 - 1/n)^n$. In all, the expected number of empty areas is $n(1-1/n)^n \geq 1/4n$ since $n > 2$. So by Markov's inequality, with probability 1/2 the number of empty areas is at least $1/8n$. Since choice of $X_1, X_2, ..., X_n$ is independent of $\widehat{f}'$ each empty area incurs error of

$$\int \phi_k^2(x)dx = L^2 h^{d+2s}.$$

For $1/8n$ empty areas, we have

$$E_{\omega,S'}[(f_w - \widehat{f}')^2] \geq \frac{1}{8} n \cdot L^2 h^{d+2s} = Cn^{-2s/d}.$$

So this means there exists at least one $\omega \in \Omega$ such that f_ω incurs at least $Cn^{-2s/d}$ error on $\widehat{f}'$ with probability at least $1/2$. This means in general the risk of $\widehat{f}'$ is at least $\frac{C}{2} n^{-2s/d}$, which proves the lemma. □

Combining Lemma 13 and Lemma 15 we prove Theorem 2. □

D. 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. Notice that in this setting $f(X_{\widehat{\pi}(i)}) = f(X_i)$, since $\widehat{\pi}(i)$ is the position of X_i in the true ranking. Also to simplify notations, let $X_{(i)} = X_{\pi^{-1}(i)}$ be the i-th element according to π. This leads to $f(X_{(1)}) \leq f(X_{(2)}) \leq \cdots \leq f(X_{(n)})$. We use these notations throughout the proof.

Proof of Theorem 3. The first part of the proof is the same as that of Theorem 1. We have

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

And for event E_0 we have

$$\begin{aligned}
\mathbb{E}\left[(\widehat{f}(X) - f(X_\alpha))^2\right] \leq\ & C\frac{d\log(1/\delta)\log n}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\widehat{y}_{\widetilde{i}} - f(X_i))^2 | E_0\right] + \delta \\
\leq\ & C\left(\frac{\log^2 n}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\widehat{y}_{\widetilde{i}} - f(X_i))^2 | E_0\right] + n^{-2s/d}\right). \qquad (10)
\end{aligned}$$

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 $X_1, ..., X_n$ and $X_{(1)}, ..., X_{(n)}$:

Lemma 4(Restated). Suppose the ranking $(X_1, ..., X_n)$ 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 $g(\theta_1, ..., \theta_n) = \sum_{i=1}^n (\theta_i - \theta_{(i)})^2$. 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 $\theta_{(1)} \le \cdots \le \theta_{(n)}$, so the maximum is achieved when for some $0 \le k \le n$ that $\theta_{(i)} = -M$ for $i \le k$, 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, in the maximum case $(\theta_{(i)} - \theta_{(\pi^{-1}(i))})^2 = 1$ iff i and $\pi^{-1}(i)$ lies on different sides of k, and otherwise it is 0. To further bound the sum, we use the Spearman Footrule distance between π and $(1, 2, ..., n)$, which (Diaconis & Graham, 1977) shows that it can be bounded as

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

And the RHS can be bounded by $2\nu n^2$ since the agnostic error of ranking is at most ν. We also have $\sum_{i=1} |\pi(i) - i| = \sum_{i=1} |\pi(\pi^{-1}(i)) - \pi^{-1}(i)| = \sum_{i=1} |i - \pi^{-1}(i)|$. Let $U_1 = \{i : \pi^{-1}(i) \le k, i > k\}$ and $U_2 = \{\pi^{-1}(i) > k, i \le k\}$. So in the maximum case 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 E_0 we have

$$\begin{aligned}
\sum_{i=1}^n E[(\widehat{y}_i - f(X_i))^2 | E_0] = & \sum_{i=1}^n E[(\widehat{y}_{\widetilde{i}} - f(X_i))^2 | E_0] \\
\le & \sum_{i=1}^n \mathbb{E}\left[2(\widehat{y}_{\widetilde{i}} - f(X_{\widetilde{i}}))^2 + 2(f(X_{\widetilde{i}}) - f(X_i))^2 | E_0\right] \\
\le & \frac{Cn\log(m/\delta)}{m} \sum_{k=1}^m \mathbb{E}\left[(\widehat{y}_{t_k} - f(X_{t_k}))^2 | E_0\right] + 2\sum_{i=1}^n \mathbb{E}\left[(f(X_{\widetilde{i}}) - f(X_i))^2 | E_0\right]. \qquad (11)
\end{aligned}$$

We omit the condition E_0 in discussion below. We bound the two terms separately. For the first term, (Zhang, 2002) shows that for isotonic regression

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

for some universal constant C, where

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

The minimum is taken over all sequence of $u \in \mathbb{R}^m$ that is non-decreasing. The expectation in $E[S]$ is taken w.r.t. the randomness in t_k. From Lemma 4 we know that

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

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

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

Now we bound the second term in (11). We have

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

The first inequality is by noticing $(x+y+z)^2 \le 3x^2+3y^2+3z^2$ for any number $x, y, z \in \mathbb{R}$; the second inequality is by grouping values of $\widetilde{i}$, 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 $f(X_{(t_m)}) - f(X_{(t_1)}) \le 1$, and Lemma 4.

Summarizing the two terms we have

$$\mathbb{E}\left[(\widehat{y}_{\widetilde{i}} - f(X_i))^2 | E_0\right] \le C(\sqrt{\nu}n + m^{-2/3}n)\log m.$$

Take this back to (10) we prove the theorem.

□

E. Model Selection

In this section we consider the case of selecting between the R^2 algorithm and a non-parametric regressor that ignores the ordinal information. We prove the following result:

Theorem 17. *Under the same assumptions as Theorem 3, there exists an estimator $\widehat{f}$ such that*

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

Proof. To simplify notation, suppose 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 train R^2 using ordinal data and k-NN with k varying in $[m]$ using only labeled data, and select the best model according to performance on validation set. Formally, let $\widehat{f}_k$ be the k-NN estimator trained on T, and $\widehat{f}_0$ be the R^2 estimator. 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 change it to $\widehat{f}'_j(x) = -M$, and similar for M; this only reduces the MSE. Define empirical risk of function $\widehat{f}$ to be $\widehat{R}^V(\widehat{f}) = \frac{1}{m}\sum_{i=m+1}^{2m}\left(y_i - \widehat{f}(X_i)\right)$, and error of $\widehat{f}$ to be $\mathrm{err}(\widehat{f}) = \mathbb{E}\left[\left(\widehat{f}(X) - f(X)\right)\right]$. Now let $\widehat{f}^* = \arg\min_{j=0,\ldots,m}\widehat{R}^V(\widehat{f}_j)$ be the best model using cross validation and $f^* = \arg\min_{j=0,\ldots,m}\mathrm{err}(\widehat{f}_j)$ be the actual best model in $\widehat{f}_0, \ldots, \widehat{f}_m$.

We use the Craig-Bernstein's inequality (Craig, 1933):

Lemma 18. *Let $X_1, \ldots, X_n$ be random variables and suppose that for $k \ge 3$,*

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

for some $r > 0$. Then with probability $> 1 - \delta$:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \le \frac{\log(1/\delta)}{nt} + \frac{tvar(X_i)}{2(1-c)},$$

for $0 \le tr \le c < 1$.

Let $U_i^j = (y_i - f(X_i))^2 - \left(y_i - \widehat{f}_k(X_i)\right)^2$, for $i = m+1, ..., 2m, j = 0, 1, ..., m$. We have $\sum_{i=m+1}^{2m} U_i^j = \widehat{R}^V(f) - \widehat{R}^V(\widehat{f}_j)$. Also

$$E[U_i^j] = -(f(X_i))^2 + 2f(X_i)\widehat{f}_j - \left(\widehat{f}(X_i)\right)^2 = -\text{err}(\widehat{f}_j).$$

We now bound the variance of U_i^j:

$$\begin{aligned}
\text{var}(U_i^j) \le \mathbb{E}\left[\left(U_i^j\right)^2\right] &\le \mathbb{E}\left[\left(\left(y_i - \widehat{f}_k(X_i)\right)^2 - (y_i - f(X_i))^2\right)^2\right] \\
&\le \mathbb{E}\left[\left(\left(\varepsilon_i + f(X_i) - \widehat{f}_k(X_i)\right)^2 - \varepsilon_i^2\right)^2\right] \\
&\le \mathbb{E}\left[\left(f(X_i) - \widehat{f}_k(X_i)\right)^4 + 4\varepsilon_i\left(f(X_i) - \widehat{f}_k(X_i)\right)^3 + 4\varepsilon_i^2\left(f(X_i) - \widehat{f}_k(X_i)\right)^2\right] \\
&\le (4M^2 + 4\sigma^2)E[U_i^j] \triangleq BE[U_i^j].
\end{aligned}$$

The last inequality is because both $\text{var}(\varepsilon_i), f(X_i), \hat{f}(X_i)$ are bounded. So the assumption of Lemma 18 holds for $r = B$. Now let $t < \frac{1}{2B}$ and $c = tr < 1$, and $a = \frac{tB}{1-c} < 1$. Now by Lemma 18 we have

$$\widehat{R}^V(f) - \widehat{R}^V(\widehat{f}_j) + (1-a)\text{err}(\widehat{f}_j) \le \frac{\log(1/\delta)}{nt}.$$

So using union bound for all $\hat{f}_j$ we have

$$\text{err}(\widehat{f}_j) \le \frac{1}{1-a}\left[\widehat{R}^V(\widehat{f}_j) - \widehat{R}^V(f) + \frac{\log(n/\delta)}{nt}\right].$$

Thus

$$\text{err}(\widehat{f}^*) \le \frac{1}{1-a}\left[\widehat{R}^V(\widehat{f}^*) - \widehat{R}^V(f) + \frac{\log(n/\delta)}{nt}\right].$$

Take expectation w.r.t. V we have

$$\text{err}(\widehat{f}^*) \le \frac{1}{1-a}\left[R(\widehat{f}^*) - R(f) + \frac{\log(n/\delta)}{nt}\right].$$

Now take expectation w.r.t. T and use Theorem 3 as well as rate of k-NN estimates we have

$$\begin{aligned}
\text{err}(\widehat{f}^*) &\le \widetilde{O}\left(m^{-2/3} + \min\{\sqrt{\nu}, m^{-\frac{2s}{2s+d}}\} + n^{-\frac{2s}{d}} + \frac{1}{n}\right) \\
&\le \widetilde{O}\left(m^{-2/3} + \min\{\sqrt{\nu}, m^{-\frac{2s}{2s+d}}\} + n^{-\frac{2s}{d}}\right)
\end{aligned}$$

□

F. Proof of Theorem 5

We prove a slightly stronger result, and show Theorem 5 as a corollary.

Theorem 19. *Assume the same modeling assumptions for $X_1, ..., X_n, y_1, ..., y_m$ as in Theorem 2. Also permutation $\widehat{\pi}$ satisfies $\mathbb{P}[(f(X_i) - f(X_j))(\pi(i) - \pi(j)) < 0] \le \nu$. Then for any estimator $\widehat{f}$ taking input $X_1, ..., X_n, y_1, ..., y_m$ and $\widehat{\pi}$, we have*

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

Proof of Theorem 19. In this proof, we use x_i to represent i-th dimension of x, and upper script for different vectors $x^{(1)}, x^{(2)}, ...$ 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 $\Gamma = \{(\gamma_1, ..., \gamma_d), \gamma_i \in \{1, 2, ..., u\}\}$. Choose an arbitrary order on Γ to be $\Gamma = \{\gamma^{(1)}, ..., \gamma^{(u^d)}\}$. Let $x^{(k)} = \frac{\gamma^{(k)} - 1/2}{u}$, and $\phi_k(x) = \frac{L}{2} th K(\frac{x - t x^{(k)}}{th}), k = 1, 2, ..., u^d$, 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 $\Omega = \{\omega = (\omega_1, ..., \omega_{u^d}), \omega_i \in \{0, 1\}\}$, 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 $t^{2d}(1/u)^2 \cdot u \le \nu$ 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 $\Omega' = \{\omega^{(0)}, \omega^{(1)}, ..., \omega^{(M)}\}$ of Ω such that the distance between each element $\omega^{(i)}, \omega^{(j)}$ is at least $u^d/8$. So $d(\theta_i, \theta_j) \ge h^s t^{(d+2)/2}$. Now for P_j, P_0 (P_0 corresponds to f_ω when $\omega = (0, 0, ..., 0)$) we have

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

We have $Cu^d t^{d+2} \le cu^d \le \alpha \log M$ (since $t \le 1$), so using Theorem 14 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 5.

Proof of Theorem 5. We only need to show

$$\min\{\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}}, 1\} m^{-\frac{2}{d+2}} \geq \min\{\nu^2, m^{-\frac{2}{d+2}}\}. \tag{12}$$

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 (12) 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 (12) have term $\nu^{\frac{d+2}{2d}} m^{\frac{1}{2d}} m^{-\frac{2}{d+2}} \geq \nu^2$, which equals RHS. □

G. Additional Discussion

Comparison models. To circumvent problems in parametric comparison models like BTL and Thurstone, we may assume more relaxed assumptions like Tsybakov noise condition (Tsybakov, 2004) typically used for classification (Xu et al., 2017); the main obstacle here is the lack of efficient algorithms that goes from comparisons to ranking under Tsybakov noise. Strong stochastic transitivity (SST)(Shah et al., 2016), which is recently proposed and more relaxed, does not apply to our setting: SST assumes that for any three samples X_i, X_j, X_k that $f(X_i) \leq f(X_j)$, the probability that X_i beats X_k is larger than the probability that X_j beats X_k. Such assumption is not enough to guarantee the quality of ranking in our setting.