Interactive Linear Regression with Pairwise Comparisons

Yichong Xu*, Sivaraman Balakrishnan†, Aarti Singh* and Artur Dubrawski‡
*Machine Learning Department,†Department of Statistics, ‡Autonlab
Carnegie Mellon University
Pittsburgh, USA
yichongx@cs.cmu.edu, siva@stat.cmu.edu, {aarti,awd}@cs.cmu.edu

***Abstract*—A general goal of interactive learning is to investigate broad ways of leveraging human feedback, and understand the benefits of learning from potentially complex feedback. We study a special case of linear regression with access to comparisons between pairs of samples. Learning from such queries is motivated by several important applications, where obtaining comparisons can be much easier than direct labels, and/or when comparisons can be more reliable. We develop an interactive algorithm that utilizes both labels and comparisons to obtain a linear estimator, and show that it only requires a very small amount of direct labels to achieve low error. We also give minimax lower bounds for the problem, showing that our algorithm is optimal up to log factors. Finally, experiments show that our algorithm outperforms label-only algorithms when labels are scarce, and it can be practical for real-world applications.**

***Index Terms*—active learning, pairwise comparisons, regression**

I. Introduction

Interactive learning has been drawing attention from the machine learning community, both theoretically and practically, as it helps enrich the feedback that ML systems can handle, and reduce the learning effort [2], [6], [24]. Different than traditional learning and crowdsourcing schemes, interactive learning relies on complex or structured *feedback* from the labelers, to provide more information or more actionable information about each sample. Another important factor of interactive learning is being *active*: Instead of dealing with a fixed dataset, interactive learning selectively chooses the most informative questions throughout the learning process.

We investigate a special case of interactive regression that uses pairwise comparisons. In many applications [20], [23], people can provide more accurate results when they compare the objective for two different samples, than giving direct labels for individual samples. Comparisons also often cost less effort for humans. For example in clinical settings, assessing the health condition of an individual may require laborious review of complex medical data, but comparing the relative health status of two patients can be easier for trained clinicians and carry less subjective bias. Another example involves evaluating face images (e.g., estimating people's age): Crowdsource workers typically have difficulty giving direct evaluations about "how old" is a person, but they are often better at comparing two face images and deciding who looks older. We conduct experiments on this task in Section VI.

We consider interactive pairwise comparisons in the special case of linear regression, one of the most important methods of statistical learning. It can be very effective when we expect a simple relationship between the sample and response. However, often the limited number of labeled training data leads to overfitting models, especially when $n < d$, where n is the number of labels and d is the dimensionality. In such a case, certain sparsity or additive assumptions have been suggested, such as LASSO [22] and SpAM [18]. We take a different approach, showing that when we have enough comparisons, it is possible to learn a linear function even when $n < d$, without making *any* special assumptions. The basic idea behind our algorithm is very simple: we first learn to predict the comparison, which is intrinsically linear classification; with comparison queries we can infer the underlying weights up to scaling transformations[1]. After that, we use labels to infer the scale of the weights, and the label complexity of such inference is *independent* of d. Our contributions come in threefold:

- We develop and analyze an interactive learning algorithm that efficiently learns a linear estimator, using both label and comparison queries. Given n label queries and m comparison queries, we show that the mean squared error (MSE) of our algorithm decays at a rate of $\widetilde{O}\left(\frac{1}{n} + \exp(-\frac{m}{d})\right)$. When sufficient comparisons are available, the label complexity of our algorithm is *independent* of dimension

[1]Note that we cannot learn the underlying weights perfectly with comparisons *only*; for comparisons $y = w^T x$ and $y = 5w^T x$ gives exactly the same results.

d. This establishes that we can achieve good MSE rate without making additional assumptions about the underlying function.
- We give complementary lower bounds to show that our results are almost optimal, up to log factors. In particular, we show that the rate of $O(\frac{1}{n})$, and the total number of queries, are not improvable up to log factors.
- Finally, we conduct experiments on synthetic and real datasets, and show that our algorithm can outperform label-only algorithms when labels are scarce. Experiments on real datasets demonstrate the practicality of our algorithm.

II. Related work

Typically, the focus of current research on pairwise comparison is to elicit a ranking from (noisy) pairwise comparisons on a finite set of items. For instance, [8] gives an efficient algorithm of error $O(\frac{1}{n})$ under the assumption that each comparison is flipped with a bounded probability. More recently, [17], [20] give algorithms that work under the Bradley-Terry [7] and Thurstone [21] models. In contrast, our goal is to elicit a regression function from comparisons, which leads to quite different models and methods.

Active learning [5], [9] focuses on actively selecting questions to present to the user, but primarily focuses on label queries only. The first part of our algorithm relies on previous results in active learning [4], but our setting differs from that of active learning since we introduce comparisons.

Interactive learning has been considered in various circumstances for different kinds of oracles. For example, [11] considers learning from partial corrections, where the user points out the error of prediction with respect to certain components of the input. [25] considers feature discovery through comparison queries. Two recent papers [15], [24] are most closely related to our work. They consider the benefits of learning from comparisons for active classification. However, we focus on regression and our method is very different from theirs.

III. Problem Statement

We assume our samples X come from a distribution $P_{\mathcal{X}}$ on $\mathbb{R}^d$. Following literature in active learning [3], [4], we assume that $P_{\mathcal{X}}$ is isotropic and log-concave; that is, features of X are independent, centered around 0, have covariance I_d; and log of the density function of X is concave. The first assumption can be achieved through standard preprocessing like ICA, and the latter is true for many prevalent choices of distributions, such as uniform and Gaussian [16]. Note that this assumption is weaker than in several regression papers [14], which assumes $P_{\mathcal{X}}$ is independent Gaussian. In addition, let $B(v, r)$ denote the ball of radius r around vector v.

Following previous literature (e.g., [9]), we assume access to a label oracle $\mathcal{O}_l$, which takes a sample $x \in \mathbb{R}^d$ and outputs a label $Y \in \mathbb{R}$. We assume a linear relation between the label and features: $Y = (w^*)^T x + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) = \sigma^2$. Let $r^* = \|w^*\|_2$ and $v^* = \frac{w^*}{\|w^*\|_2}$.

In addition to traditional label queries, we assume access to a (potentially cheaper) comparison oracle $\mathcal{O}_c$. On each query, $\mathcal{O}_c$ receive a pair of samples $(X, X') \sim P_{\mathcal{X}} \times P_{\mathcal{X}}$, and returns a random variable $Z \in \{-1, +1\}$, where $Z = 1$ indicates that the user thinks $f(X) > f(X')$, and $Z = -1$ otherwise. We assume an agnostic noise[2] ν for Z:

$$\Pr(Z \neq \text{sign}(w^* \cdot X - w^* \cdot X')) \leq \nu.$$

That is, a randomly sampled triplet (X, X', Z) is wrong with probability at most ν. Note that the error for a particular example $(X, X') = (x, x')$ can be arbitrary.

Given a (arbitrarily large) set of unlabeled instances $\mathcal{U} = \{X_1, X_2, ...\}$ coming from $P_{\mathcal{X}}$, 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 n and m respectively. We characterize the quality of any such estimator $\widehat{w}$ in terms of the mean squared error (MSE) $E\left[\left((w^* - \widehat{w})^T X\right)^2\right]$, where the expectation is taken over randomness in estimator $\widehat{w}$, oracles $\mathcal{O}_l, \mathcal{O}_c$ and random sample X respectively. We also study information-theoretic limits of any estimator by examining the minimax risk:

$$\mathcal{M}(m, n) = \inf_{\widehat{w}} \sup_{w^*} E\left[\left((w^* - \widehat{w})^T X\right)^2\right]. \quad (1)$$

As a final remark of this section, the classical minimax rate for ordinary least squares(OLS) is of order $O(\frac{d}{n})$, where n is the number of label queries. This rate cannot be improved by active label queries (c.f. [10]).

IV. Algorithm & Analysis

Our algorithm is described in Algorithms 1 and 2. We first consider comparisons as a classification problem with samples $((X - X'), Z)$, and use active linear classification to learn an estimated $\widehat{v} \approx v^*$. The active classification algorithm we use comes from [4], and is presented in Algorithm 2. In each iteration, it find the best weights that minimize the hinge loss

$$l_\tau(u, W) = \sum_{i=1}^{m_k} \max(0, 1 - \frac{z_i(u \cdot (x_i - x_i'))}{\tau}), \quad (2)$$

[2]Our model can also be adapted to the bounded noise model case using a different algorithm from active learning; See Section VII for details.

where $W = \{(x_i, x_i', z_i)\}_{i=1}^{m_k}$ is the labeled dataset. This vector is normalized and then used as the criterion for selecting pairs. The final result of the classification is a unit vector $\widehat{v} \approx v^*$.

After classification, we use the estimated $\widehat{v}$ along with actual label queries to learn an estimated weight norm $\widehat{r}$, through OLS. Combining the results we get $\widehat{w} = \widehat{r} \cdot \widehat{v}$.

Algorithm 1 Linear Regression with Comparison Queries

Input: comparison budget m, label budget n, comparison oracle $\mathcal{O}_c$, set of parameters for Algorithm 2

1: Run Algorithm 2 with $m, \mathcal{O}_c$ and obtain $\widehat{v}$
2: Query random samples $\{(X_i, Y_i)\}_{i=1}^n$
3: Let $\widehat{r} = \dfrac{\sum_{i=1}^n \widehat{v}^T X_i Y_i}{\sum_{i=1}^n (\widehat{v}^T X_i)^2}$.

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

Algorithm 2 Active-Comparison

Input: Comparison oracle $\mathcal{O}_c$, comparison budget m, sample sizes m_k, sequences r_k, b_k, τ_k, precision value κ.

1: Draw m_1 pairs i.i.d. into set U and ask $\mathcal{O}_c$ to label them. Obtain labeled set $W = \{(x_i, x_i', z_i)\}_{i=1}^{m_1}$.
2: Iteration counter $k \leftarrow 0$
3: **while** Comparison budget not exhausted **do**
4: Find $u_k \in B(v_{k-1}, r_k)$ that approximately minimize training hinge loss (2) over W, with length at most 1:

$$l_{\tau_k}(u_k, W) \le \min_{u \in B(v_{k-1}, r_k) \cap B(0,1)} l_{\tau_k}(u, W) + \kappa/8.$$

5: $v_k \leftarrow \frac{u_k}{\|u_k\|_2}$.
6: Sample another dataset U of m_k unlabeled sample pairs.
7: $U' = \{(x, x') \in U : |v_k \cdot (x - x')| \le b_k\}$.
8: Ask $\mathcal{O}_c$ to label all samples in U' and obtain labeled set W.
9: Increment counter $k \leftarrow k + 1$

Output: Return v_k.

Theorem 1. *There exists some constants C, N such that if $n > N$, the MSE of Algorithm 1 satisfies*[3]

$$\begin{aligned}&\mathbb{E}[((w^*)^T X - \widehat{w}^T X)^2] \\ \le &\widetilde{O}\left(\frac{1}{n} + \log^2 n \exp\left(-\frac{Cm}{d \log^3(nd)}\right) + \nu^2\right).\end{aligned}$$

Several remarks are in order before we turn to details of the proof.

[3]We use $\widetilde{O}$ to represent expressions without the $\log\log(\cdot)$ terms.

Remarks. (1) The MSE rate for classical ordinary least square (OLS) is $\frac{d}{n}$ (see e.g., [12]). Theorem 1 reduces this to $\frac{1}{n}$, which is *independent* of n. This is critical for high-dimensional linear regression, where we typically have $d > n$.
(2) The dependence on m, however, is dependent on d, and we generally require $m > d$ to obtain a low MSE. This suggests that comparisons are most helpful when they are ample.
(3) Somewhat surprisingly, the dependence on $\frac{m}{d}$ is exponential, making our algorithm query efficient once it reaches $m > d$. Suppose we aim at a MSE of γ for some small $\gamma > 2\nu^2$, classical OLS requires $O(d/\gamma)$ labels, whereas Algorithm 2 only needs a much-less $n + m = \widetilde{O}(1/\gamma + d\log(d/\gamma))$ queries in total. We show in Section V that this quantity is optimal up to log factors.
(4) Finally, we remark that Theorem 1 indicates an upper bound on the minimax risk of (1).

The full proof is deferred to the Appendix due to space limits; we give a sketch here.

Proof Sketch. First, using results in [4], we obtain an estimator $\|\widehat{v} - v^*\|_2 \le \varepsilon = O\left(\exp\left(-\frac{m}{C_2 d \log^3(nd)}\right) + \nu\right)$ when we finish the active-comparison in Algorithm 2. Now let $T_i = \widehat{v}^T X_i$, and we have

$$\begin{aligned}\widehat{r} &= \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i^2} \\ &= r^* + \frac{\sum_{i=1}^n T_i r^* (v^* - \widehat{v})^T X_i + \varepsilon_i}{\sum_{i=1}^n T_i^2}.\end{aligned}$$

And thus

$$\begin{aligned}&(w^*)^T X - (\widehat{w})^T X \\ =&r^*(v^* - \widehat{v})^T X - \frac{\sum_{i=1}^n T_i r^* (v^* - \widehat{v})^T X_i}{\sum_{i=1}^n T_i^2} \widehat{v}^T X + \\ &\frac{\sum_{i=1}^n T_i \varepsilon_i}{\sum_{i=1}^n T_i^2} \widehat{v}^T X\end{aligned}$$

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

V. Lower Bounds

Now we turn to information-theoretic lower bounds of the minimax risk (1). We consider *any* active estimator $\widehat{w}$ with access to the two oracles $\mathcal{O}_c, \mathcal{O}_l$, using m comparisons and n labels, and show that the upper bound of MSE in Theorem 1 is optimal up to log factors. Our results come in two parts: Theorem 2 shows a lower bound that captures dependency on n, for any $\widehat{w}$ using

comparisons. Then, Theorem 3 shows lower bound on the total number of queries $(m+n)$.

Theorem 2. *Suppose X is uniform in $[-1,1]^d$, and $\varepsilon \sim \mathcal{N}(0,\sigma^2)$. Then for any (active) estimator $\widehat{w}$ with access to both label and comparison oracles, there is a constant C such that*

$$\inf_{\widehat{w}} \sup_{w^*} \mathbb{E}\left[\left((w^*-\widehat{w})^T X\right)^2\right] \geq \frac{C}{n}.$$

Theorem 2 shows that the $O(\frac{1}{n})$ term in Theorem 1 is necessary. The proof is quite standard using Le Cam's method for two increasing functions when $d=1$, and is included in the Appendix.

Theorem 3. *For any (active) estimator $\widehat{w}$ with access to n labels and m comparisons, there exists a ground truth weight $\widetilde{w}$ and a global constant C, such that when $w^*=\widetilde{w}$ and $2m+n<d$,*

$$\mathbb{E}\left[\left((\widehat{w}-w^*)^T X\right)^2\right] \geq C.$$

Theorem 3 shows a lower bound on the total number of queries in order to get low error. Combining with Theorem 2, 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 1, we need $n+m=\widetilde{O}(1/\gamma+d\log(d/\gamma))$ for Algorithm 1 to reach γ MSE. So Algorithm 1 is optimal in terms of total queries, up to log factors.

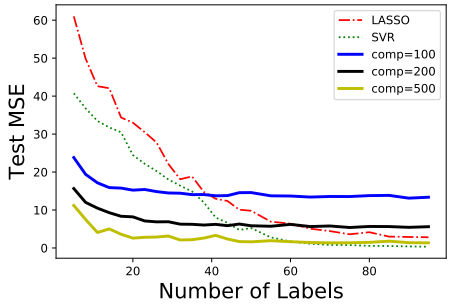
The proof of Theorem 3 is done by considering an estimator with access to $n+2m$ *noiseless* labels $\{(x_i, w^*\cdot x_i)\}_{i=1}^{n+2m}$, which can be used to generate m comparisons and n labels. We sample w^* from a prior distribution in $B(0,1)$, and show that the expectation of MSE in this case is at least a constant. Thus there exists a weight vector $\widetilde{w}$ that leads to constant error. The full proof is deferred to the Appendix.

VI. Experiments

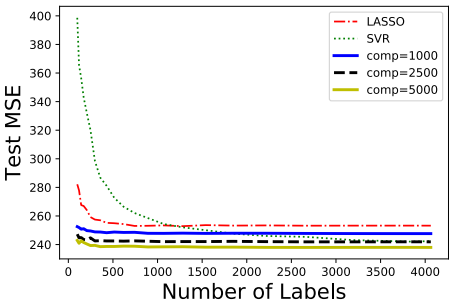
We test our algorithm through experiments to verify its practical use. We compare Algorithm 1 with two strong baselines in linear regression: LASSO [22] and support vector regression. We first conduct experiments in a controlled setting using synthetic data. After that, we consider a practical task of estimating people's ages from portraits. We repeat each experiment 20 times and report the average MSE.

A. Synthetic Dataset

For synthetic data, we set $d=50$ and generate both X and w^* from $\mathcal{N}(0,I_d)$, and $\varepsilon \sim \mathcal{N}(0,0.5^2)$. The comparison oracle generates response using the same noise model: $Z=\text{sign}((w^*\cdot x+\varepsilon)-(w^*\cdot x'-\varepsilon'))$ for input (x,x'), with $\varepsilon,\varepsilon' \sim \mathcal{N}(0,0.5^2)$. We generate a training set of size $n_{\text{train}}=4000$ and test set of size



(a)



(b)

Fig. 1: (a) Experiment result on synthetic dataset. (b) Experiment result for age estimation. Best viewed in color.

$n_{\text{test}}=1000$. At test time we compute the empirical MSE $\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}[(\widehat{w}-w^*)^T X_i^{\text{test}}]^2$.

Model performances are compared in Figure 1a. We vary the number of labels n from 5 to 100, and the number of comparisons m in $\{100,200,500\}$. When the number of labels is small, our algorithm consistently outperforms the baselines. For larger numbers of labels, our algorithm achieves similar performance as SVR, when $m=500$.

B. Age Estimation

In this section, we consider a practical task of estimating people's ages from their portraits. The APPA-REAL dataset [1] contains 7,591 images each associated with a biological age and an apparent age. The biological age is the person's actual age, whereas the apparent ages are crowdsourced by human. 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. We extract the 128-dim feature for each image using the last layer of FaceNet [19]. The features are centralized to have zero mean and unit variance.

We consider the task of predicting biological ages; typically it is hard to obtain direct biological labels of people in portraits, but it is relatively easy to crowdsource comparisons. Note that comparisons in this case are based on apparent ages instead of biological ages. So in our experiments, the comparison oracle $\mathcal{O}_c$ returns

labels according to the apparent age, and the label oracle $\mathcal{O}_l$ returns biological ages directly.

We vary the number of labels from $n = 100$ to 4,063 (all labels) and number of comparisons in $m \in \{1000, 2500, 5000\}$. Results (Figure 1b) show that our algorithm requires fewer labels than baseline methods to achieve low MSE. Also, the total number of queries that our algorithm makes is smaller than that of baseline methods. This verifies our theoretical results in Section IV and V, and demonstrates practicality of our method.

VII. Discussion and Conclusion

We develop interactive learning algorithms for linear regression with access to both label and comparison oracles. Our results show that when comparison labels are copious, only a very small amount of direct labels is required to learn a linear estimator with good accuracy. We also provide complementary lower bounds to show the optimality of our algorithm. Experiments on both synthetic and real-world datasets show the practicality of our algorithm. In applications, comparisons are typically easier to obtain than labels, and our algorithm becomes more effort-efficient than label-only algorithms in these cases.

We analyze our method under the assumption of agnostic comparisons. The same results can be easily obtained for the case where each comparison is flipped (w.r.t. ground truth) with some probability $\eta < 1/2$, using algorithm in [13], however that algorithm is computationally inefficient. Currently the best efficient active learning algorithm under such "bounded noise" setting requires $m \geq O\left(d^{O(\frac{1}{(1-2\eta)^4})}\right)$ comparisons [3]. So it remains open to solve such cases using around $\widetilde{O}(d)$ comparisons. It would also be interesting to consider other comparison models such as BTL [7] and Thurstone [21].

Other interesting directions include extending our method to more complex models, such as generalized linear and graphical models. A broader goal would be to understand the fundamental benefits and limits of utility of a general class of indirect queries, along with their applications.

References

[1] 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 *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 87–94. IEEE, 2017.

[2] J. Attenberg, P. Melville, and F. Provost. A unified approach to active dual supervision for labeling features and examples. In *Machine Learning and Knowledge Discovery in Databases*, pages 40–55. Springer, 2010.

[3] P. Awasthi, M.-F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory*, pages 152–192, 2016.

[4] P. Awasthi, M. F. Balcan, and P. 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.

[5] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[6] M.-F. Balcan and S. Hanneke. Robust interactive learning. In *COLT*, pages 20–1, 2012.

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

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

[9] K. Chaudhuri, P. Jain, and N. Natarajan. Active heteroscedastic regression. In *International Conference on Machine Learning*, pages 694–702, 2017.

[10] K. Chaudhuri, S. M. Kakade, P. Netrapalli, and S. Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015.

[11] S. Dasgupta and M. Luby. Learning from partial correction. *arXiv preprint arXiv:1705.08076*, 2017.

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

[13] S. Hanneke. *Theoretical foundations of active learning*. ProQuest, 2009.

[14] L. Janson, R. F. Barber, and E. Candes. Eigenprism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1037–1065, 2017.

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

[16] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[17] L. Maystre and M. Grossglauser. Just sort it! a simple and effective approach to active preference learning. *arXiv preprint arXiv:1502.05556*, 2015.

[18] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1201–1208. Curran Associates Inc., 2007.

[19] F. Schroff, D. Kalenichenko, and J. 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.

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

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

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

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

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

[25] J. Y. Zou, K. Chaudhuri, and A. T. Kalai. Crowdsourcing feature discovery via adaptively chosen comparisons. *arXiv preprint arXiv:1504.00064*, 2015.

APPENDIX

A. Proof of Theorem 1

Proof. We first need the following theorem for log-concave distributions:

Theorem 4 ([16]). *The following holds for isotropic log-concave distribution $P_{\mathcal{X}}$:*
1. The convolution of two log-concave functions is also log-concave.
2. $\Pr[\|X\|_2 \geq \alpha\sqrt{d}] \leq e^{1-\alpha}$.

Using point 1 we know that $X - X'$ also follows a isotropic log-concave distribution. We use the following theorem, adapted from [4]:

Theorem 5 ([4]). *Suppose $X - X'$ is isotropic log-concave. There exists sequences r_k, b_k, τ_k, κ and constant C such that, for every $\varepsilon > 0$, if $\nu < C\varepsilon$, algorithm 2 returns a vector $\|\widehat{v} - v^*\|_2 \leq \varepsilon$ using label budget*

$$m = O(d\log(1/\varepsilon)\log^3(d/\delta))$$

with probability $1 - \delta$. [4]

Let $\varepsilon = O\left(e^{-\frac{m}{Cd\log^3(d/\delta)}} + \nu\right)$, from Theorem 5 we obtain that we can get $\|\widehat{v} - v^*\|_2 \leq \varepsilon$ using m comparisons, with probability $1 - \delta/4$.

Now consider estimating r^*. The following discussion is conditioned on a fixed $\widehat{v}$. For simplicity let $T_i = \widehat{v}^T X_i$. We have

$$\begin{aligned}\widehat{r} &= \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i^2} \\ &= \frac{\sum_{i=1}^n T_i r^*(v^*)^T X_i + T_i\varepsilon_i}{\sum_{i=1}^n T_i^2} \\ &= r^* + \frac{\sum_{i=1}^n T_i r^*(v^* - \widehat{v})^T X_i + T_i\varepsilon_i}{\sum_{i=1}^n T_i^2}.\end{aligned}$$

Now we have

$$\begin{aligned}&(w^*)^T X - (\widehat{w})^T X \\ =&r^*(v^*)^T X - \widehat{r}(\widehat{v})^T X \\ =&r^*(v^* - \widehat{v})^T X - \frac{\sum_{i=1}^n T_i r^*(v^* - \widehat{v})^T X_i + T_i\varepsilon_i}{\sum_{i=1}^n T_i^2}\widehat{v}^T X.\end{aligned}$$

[4]The original result in [4] is with dependency $d^2\log^2(1/\varepsilon)$; this can be improved to $d\log(1/\varepsilon)$ by results in [24].

So

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

The first term can be bounded by

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

For the latter two terms, we first bound the denominator $\sum_{i=1}^n T_i^2$ using Hoeffding's inequality. First notice that $E[T_i^2] = E[(\widehat{v}^T X)^2] = \|\widehat{v}\|_2^2 = 1$, since X is isotropic. So from point 1 in Theorem 4, each T_i is also isotropic log-concave. Now using point 2 in Theorem 4 with $\alpha = 1 - \log(\delta/(4en))$ we get that with probability $1 - \delta/4$, $T_i \leq \log(4en/\delta)$ for all $i \in \{1, 2, ..., n\}$. Now using Hoeffding's inequality, for any $t > 0$

$$\begin{aligned}&\Pr\left[\frac{1}{n}\sum_{i=1}^n T_i^2 - \mathbb{E}[\left(\widehat{v}^T X\right)^2] \leq -t\right] \\ \leq&\exp\left(-\frac{2nt^2}{\log^2(4en/\delta)}\right).\end{aligned}$$

Let $t = \frac{1}{2}E[\left(\widehat{v}^T X\right)^2]$, we have

$$\sum_{i=1}^n T_i^2 \geq \frac{n}{2}E[\left(\widehat{v}^T X\right)^2] = \frac{n}{2}. \tag{4}$$

with probability $1 - \delta/4$, when $n = \widetilde{O}(\log^3(1/\delta))$. Similarly we have $\sum_{i=1}^n T_i^2 \leq 2n$ with probability $1 - \delta/4$. Let E_δ denote the event when $n/2 \leq \sum_{i=1}^n T_i^2 \leq 2n$ and T_i are bounded by $\log(4en/\delta)$ for all i . Condition on E_δ for the second term in (3) we have

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

Now notice that $\frac{\widehat{v} - v^*}{\|\widehat{v} - v^*\|_2}X$ is also isotropic log-concave; using point 2 in Theorem 4 we have with probability

$1-\delta/4$, $(\widehat{v}-v^*)^T X_i \le \|\widehat{v}-v^*\|_2 \log(4en/\delta)$ for all $i \in \{1,2,...,n\}$. So

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^n T_i r^*(v^*-\widehat{v})^T X_i\right)^2\right] \\
\le&(r^*)^2\varepsilon^2\log^2(4en/\delta)\mathbb{E}\left[\left(\sum_{i=1}^n T_i\right)^2\right] \\
=&(r^*)^2\varepsilon^2\log^2(4en/\delta)\mathbb{E}\left[n\sum_{i=1}^n T_i^2\right] \\
\le&2(r^*)^2\varepsilon^2\log^2(4en/\delta)n^2
\end{aligned}$$

For the third term in (3), also conditioning on E_δ we have

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

Combining the three terms and removing all conditioning, we have

$$\begin{aligned}
&\mathbb{E}\left[\left((w^*)^T X-(\widehat{w})^T X\right)^2\right] \\
\le&4(r^*)^2\varepsilon^2+(r^*)^2\varepsilon^2\log^2(4en/\delta)+\frac{4\sigma^2}{n}+C'\delta \\
\le&O\left(\frac{1}{n}+\log^2(n/\delta)\exp\left(-\frac{m}{Cd\log^3(d/\delta)}\right)+\nu^2+\delta\right)
\end{aligned}$$

Taking $\delta=\frac{1}{n}$ obtain our desired result.

□

B. Proof of Theorem 2

Proof. We use the Le Cam's method, explained in the lemma below:

Lemma 1. *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 $KL(P_1,P_2) \le \alpha \le \infty$. Let d be a semi-distance on Θ, and $d(\theta_0,\theta_1)=2a$. Then for any estimator $\widehat{\theta}$ we have*

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

We consider two functions: $w_0^*=(\xi,0,0,...,0)^T$ and $w_1^*=(\frac{1}{\sqrt{n}},0,0,...,0)^T$, 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 $d(w_0^*,w_1^*)=E\left[\left((w_0^*-w_1^*)^T X\right)^2\right]=\left(\frac{1}{\sqrt{n}}-\xi\right)^2$. For any estimator $\widehat{w}$, let $\{(X_i,Y_i)\}_{i=1}^n$ 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 KL-divergence we have

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

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

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

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

$$\mathrm{KL}(P_1,P_0)\le C_1, d(\theta_0,\theta_1)\ge\frac{C_2}{n}.$$

Combining with Lemma 1 we obtain the lower bound.

□

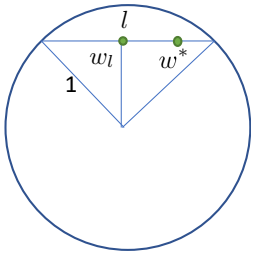


Fig. 2: Graphic illustration about the sampling process in the proof of Theorem 3.

C. Proof of Theorem 3

Proof. We just prove the theorem for $2m + n = d - 1$. Note that this case can be simulated by considering an estimator with access to $2m + n$ *truthful* samples; that is, $Y_i = w^* \cdot X_i$ for $i = 1, 2, ..., 2m + n$. In this way comparison and labels can be simulated. We now prove a lower bound for this new case with $2m + n = d - 1$ 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 $x_1, ..., x_{d-1}$ 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 $X_1, ..., X_{d-1}$ denote the sequence of queried points when w^* is randomly picked as above. Now note that for every w,

$$\begin{aligned} &\Pr[w^* = w | \{X_i, Y_i\}_{i=1}^{d-1}, l] \\ =&\Pr[w^* = w | \{X_i, Y_i\}_{i=1}^{d-1}] \\ \propto&\Pr[w^* = w] I(w \in l) \end{aligned}$$

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 $X_1, ..., X_{d-1}$, 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^*). Denote by w_l the middle point of l (see Figure 2). For any such line l, the best estimator is to predict the middle point of l. So we have

$$\begin{aligned} &\mathbb{E}\left[\left((w^* - \widehat{w})^T X\right)^2 | l \text{ has distance less than } 1/2\right] \\ \geq &\int_{u=0}^{|l|/2} u^2 dP(\|w^* - w_l\|_2 \geq u | w^* \in l). \end{aligned} \tag{5}$$

Note that the distribution of $w^* \in l$ is equivalent to that we sample from the circle containing l and origin, and then condition on $w^* \in l$ (see Figure 2). Notice that this sampling process is the same as when $d = 2$; and with some routine calculation we can show that (5) is a constant C. So overall we have

$$\mathbb{E}\left[\left((w^* - \widehat{w})^T\right)^2\right] \geq \frac{1}{2}C$$

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

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

which proves the theorem. □