Linear Label Ranking with Bounded Noise

Dimitris Fotakis
NTUA
fotakis@cs.ntua.gr

Alkis Kalavasis
NTUA
kalavasisalkis@mail.ntua.gr

Vasilis Kontonis
UW Madison
kontonis@wisc.edu

Christos Tzamos
UW Madison
tzamos@wisc.edu

Abstract

Label Ranking (LR) is the supervised task of learning a sorting function that maps feature vectors $x \in \mathbb{R}^d$ to rankings $\sigma(x) \in S_k$ over a finite set of $k$ labels. We focus on the fundamental case of learning linear sorting functions (LSFs) under Gaussian marginals: $x$ is sampled from the $d$-dimensional standard normal and the ground truth ranking $\sigma^*(x)$ is the ordering induced by sorting the coordinates of the vector $W^*x$, where $W^* \in \mathbb{R}^{k \times d}$ is unknown. We consider learning LSFs in the presence of bounded noise: assuming that a noiseless example is of the form $(x, \sigma^*(x))$, we observe $(x, \pi)$, where for any pair of elements $i \neq j$, the probability that the order of $i, j$ is different in $\pi$ than in $\sigma^*(x)$ is at most $\eta < 1/2$. We design efficient non-proper and proper learning algorithms that learn hypotheses within normalized Kendall’s Tau distance $\epsilon$ from the ground truth with $N = \tilde{O}(d \log(k)/\epsilon)$ labeled examples and runtime $\text{poly}(N, k)$. For the more challenging top-$r$ disagreement loss, we give an efficient proper learning algorithm that achieves $\epsilon$ top-$r$ disagreement with the ground truth with $N = \tilde{O}(dkr/\epsilon)$ samples and $\text{poly}(N)$ runtime.

1 Introduction

1.1 Background and Motivation

Label Ranking (LR) is the problem of learning a hypothesis that maps features to rankings over a finite set of labels. Given a feature vector $x \in \mathbb{R}^d$, a sorting function $\sigma(\cdot)$ maps it to a ranking of $k$ alternatives, i.e., $\sigma(x)$ is an element of the symmetric group with $k$ elements, $S_k$. Assuming access to a training dataset of features labeled with their corresponding rankings, i.e., pairs of the form $(x, \pi) \in \mathbb{R}^d \times S_k$, the goal of the learner is to find a sorting function $h(x)$ that generalizes well over a fresh sample. LR has received significant attention over the years [DSM03, SS07, HFCB08, CH08, FHMB08] due to the large number of applications. For example, ad targeting [DGR+14] is an LR instance where for each user we want to use their feature vector to predict a ranking over ad categories and present them with the most relevant. The practical significance of LR has lead to the development of many techniques based on probabilistic models and instance-based methods [CH08, CDH10, GDV12, ZLGQ14], decision trees [CHH09], entropy-based ranking trees [RdSRSK15], bagging [AGM17], and random forests [dSSKC17, ZQ18]. However, almost all of these works come without provable guarantees and/or fail to learn in the presence of noise in the observed rankings.

Linear Sorting Functions (LSFs). In this work, we focus on the fundamental concept class of Linear Sorting functions [HPRZ03]. A linear sorting function parameterized by a matrix $W \in \mathbb{R}^{k \times d}$ with $k$ rows $W_1, \ldots, W_k$ takes a feature $x \in \mathbb{R}^d$, maps it to $Wx = (W_1 \cdot x, \ldots, W_k \cdot x) \in \mathbb{R}^k$ and
We consider the fundamental setting where the feature vector $\Delta$ which they disagree. That is, $x_1, \ldots, x_k \in \mathbb{R}^d$ is generated by a standard normal distribution and the ground-truth ranking for each sample $x_i$ is given by the LSF $\sigma_{W^*}(x)$ for some unknown parameter matrix $W^* \in \mathbb{R}^{k \times d}$. For a fixed $x$, the ranking that we observe comes from an $\eta$-noisy ranking distribution with ground-truth ranking $\sigma_{W^*}(x)$.

Definition 2 (Noisy Linear Label Ranking Distribution). Fix $\eta \in [0, 1/2]$ and some ground-truth parameter matrix $W^* \in \mathbb{R}^{k \times d}$. We assume that the $\eta$-noisy linear label ranking distribution $D$ over $\mathbb{R}^d \times S_k$ satisfies the following:

1. The $x$-marginal of $D$ is the d-dimensional standard normal distribution.
2. For any $(x, \pi) \sim D$, the distribution of $\pi$ conditional on $x$ is an $\eta$-noisy ranking distribution with ground-truth ranking $\sigma_{W^*}(x)$.

At first sight, the assumption that the underlying $x$-marginal is the standard normal may look too strong. However, for $k = 2$, Definition 2 captures the problem of learning linear threshold functions with Massart noise. Without assumptions for the $x$-marginal, it is known [DGT19, CKMY20, DK20, NT22] that optimal learning of halfspaces under Massart noise requires super-polynomial time (in the Statistical Query model of [Kea98]). On the other hand, a lot of recent works [BZ17, MV19, DKTZ20, ZSA20, ZL21] have obtained efficient algorithms for learning Massart halfspaces under Gaussian marginals. The goal of this work is to provide efficient algorithms for the more general problem of learning LSFs with bounded noise under Gaussian marginals.

1.2 Our Results

The main contributions of this paper are the first efficient algorithms for learning LSFs with bounded noise with respect to Kendall’s Tau distance and top-$r$ disagreement loss.

Learning in Kendall’s Tau Distance. The most standard metric in rankings [SSBD14] is Kendall’s Tau (KT) distance which, for two rankings $\pi, \tau \in S_k$, measures the fraction of pairs $(i, j)$ on which they disagree. That is, $\Delta_{KT}(\pi, \tau) = \sum_{i < j} 1(i \succ \tau j) / \binom{k}{2}$. Our first result is an efficient learning algorithm that, given samples from an $\eta$-noisy linear label ranking distribution $D$, computes $\Delta_{KT}(\pi, \tau)$.

We use $i \succ \pi j$ (resp. $i \prec \pi j$) to denote that the element $i$ is ranked higher (resp. lower) than $j$ according to the ranking $\pi$.
a parameter matrix \( W \) that ranks the alternatives almost optimally with respect to the KT distance from the ground-truth ranking \( \sigma_{W^*} (\cdot) \).

**Theorem 1 (Learning LSFs in KT Distance).** Fix \( \eta \in [0, 1/2) \) and \( \epsilon, \delta \in (0, 1) \). Let \( \mathcal{D} \) be an \( \eta \)-noisy linear label ranking distribution satisfying the assumptions of Definition 2 with ground-truth LSF \( \sigma_{W^*} (\cdot) \). There exists an algorithm that draws \( N = \tilde{O} \left( \frac{d}{(1-2\eta)^2} \log(k/\delta) \right) \) samples from \( \mathcal{D} \), runs in sample-polynomial time, and computes a matrix \( W \in \mathbb{R}^{k \times d} \) such that, with probability at least \( 1 - \delta \),

\[
E_{x \sim \mathcal{N}_d} \left[ \Delta_{\text{KT}}(\sigma_W(x), \sigma_{W^*}(x)) \right] \leq \epsilon .
\]

Theorem 1 gives the first efficient algorithm with provable guarantees for the supervised problem of learning noisy linear rankings. We remark that the sample complexity of our learning algorithm is qualitatively optimal (up to logarithmic factors) since, for \( k = 2 \), our problem subsumes learning a linear classifier with Massart noise \(^2\) for which \( \Omega(d/\epsilon) \) are known to be information theoretically necessary [MN06]. Moreover, our learning algorithm is proper in the sense that it computes a linear sorting function \( \sigma_W (\cdot) \). As opposed to improper learners (see also Section 1.3), a proper learning algorithm gives us a compact representation (storing \( W \) requires \( O(kd) \) memory) of the sorting function that allows us to efficiently compute (with runtime \( O(kd + k \log k) \)) the ranking corresponding to a fresh datapoint \( x \in \mathbb{R}^d \).

**Learning in top-\( r \) Disagreement.** We next present our learning algorithm for the top-\( r \) metric formally defined as \( \Delta_{\text{top-}r} (\pi, \tau) = 1 \{ \pi_{1:r} \neq \tau_{1:r} \} \), where by \( \pi_{1:r} \) we denote the ordering on the first \( r \) elements of the permutation \( \pi \). The top-\( r \) metric is a disagreement metric in the sense that it takes binary values and for \( r = 1 \) captures the standard (multiclass) top-1 classification loss. We remark that, in contrast with the top-\( r \) classification loss, which only requires the predicted label to be in the top-\( r \) predictions of the model, the top-\( r \) ranking metric that we consider here requires that the model puts the same elements in the same order as the ground truth in the top-\( r \) positions. The top-\( r \) ranking is well-motivated as, for example, in ad targeting (discussed in Section 1.1) we want to be accurate on the top-\( r \) ad categories for a user so that we can diversify the content that they receive.

**Theorem 2 (Learning LSFs in top-\( r \) Disagreement).** Fix \( \eta \in [0, 1/2) \), \( r \in [k] \) and \( \epsilon, \delta \in (0, 1) \). Let \( \mathcal{D} \) be an \( \eta \)-noisy linear label ranking distribution satisfying the assumptions of Definition 2 with ground-truth LSF \( \sigma_{W^*} (\cdot) \). There exists an algorithm that draws \( N = \tilde{O} \left( \frac{d r k}{(1-2\eta)^2} \log(1/\delta) \right) \) samples from \( \mathcal{D} \), runs in sample-polynomial time and computes a matrix \( W \in \mathbb{R}^{k \times d} \) such that, with probability at least \( 1 - \delta \),

\[
E_{x \sim \mathcal{N}_d} \left[ \Delta_{\text{top-}r}(\sigma_W(x), \sigma_{W^*}(x)) \right] \leq \epsilon .
\]

As a direct corollary of our result, we obtain a proper algorithm for learning the top-1 element with respect to the standard 0-1 loss that uses \( \tilde{O}(kd) \) samples. In fact, for small values of \( r \), i.e., \( r = \tilde{O}(1) \), our sample complexity is essentially tight. It is known that \( \Theta(kd) \) samples are information theoretically necessary [Nat89] for top-1 classification. \(^3\) For the case \( r = k \), i.e., when we want to learn the whole ranking with respect to the 0-1 loss, our sample complexity is \( O(k^2d) \). However, using arguments similar to [DSBDSS11], one can show that in fact \( O(dk) \) ranking samples are sufficient in order to learn the whole ranking with respect to the 0-1 loss. In this case, it is unclear whether a better sample complexity can be achieved with an efficient algorithm and we leave this as an interesting open question for future work.

### 1.3 Our Techniques

**Learning in Kendall’s Tau distance.** Our proper learning algorithm consists of two steps: an improper learning algorithm that decomposes the ranking problem to \( O(k^2) \) binary linear classification problems and a convex (second order conic) program that “compresses” the \( k^2 \) linear classifiers

---

\(^2\)Notice that in this case Kendall’s Tau distance is simply the standard 0-1 binary loss.

\(^3\)Strictly speaking, those lower bounds do not directly apply in our setting because our labels are whole rankings instead of just the top classes but, in the Appendix D, we show that we can adapt the lower bound technique of [DSBDSS11] to obtain the same sample complexity lower bound for our ranking setting.
to obtain a $k \times d$ matrix $W$. Our improper learning algorithm splits the ranking learning problem into $O(k^2)$ binary, $d$-dimensional linear classification problems with Massart noise. In particular, for every pair of elements $i, j \in [k]$, each binary classification task asks whether element $i$ is ranked higher than element $j$ in the ground-truth permutation $\sigma_W(x)$. As we already discussed, we have that, under the Gaussian distribution, there exist efficient Massart learning algorithms [BZ17, MV19, DKTZ20, ZSA20, ZL21] that can recover linear classifiers $\text{sgn}(v_{ij} \cdot x)$ that correctly order the pair $i, j$ for all $x$ apart from a region of $O(\varepsilon)$-Gaussian mass. However, we still need to aggregate the results of the approximate binary classifiers in order to obtain a ranking of the $k$ alternatives for each $x$. We first show that we can design a “voting scheme” that combines the results of the binary classifiers using an efficient constant factor approximation algorithm for the Minimum Feedback Arc Set (MFAS) problem [ACN08]. This gives us an efficient but improper algorithm for learning LSFs in Kendall’s Tau distance. In order to obtain a proper learning algorithm, we further “compress” the $O(k^2)$ approximate linear classifiers with normal vectors $v_{ij}$ and obtain a matrix $W \in \mathbb{R}^{k \times d}$ with the property that the difference of every two rows $W_i - W_j$ is $O(\varepsilon)$-close to the vector $v_{ij}$. More precisely, we show that, given the linear classifiers $v_{ij} \in \mathbb{R}^d$, we can efficiently compute a matrix $W \in \mathbb{R}^{k \times d}$ such that the following angle distance with $W^*$ is small:

$$d_{\text{angle}}(W, W^*) \triangleq \max_{i,j} \theta(W_i - W_j, W_i^* - W_j^*) \leq O(\varepsilon).$$

It is not hard to show that, as long as the above angle metric is at most $O(\varepsilon)$, then (in expectation over the standard Gaussian) Kendall’s Tau distance between the LSFs is also $O(\varepsilon)$. A key technical difficulty that we face in this reduction is bounding the “condition number” of the convex (second order conic) program that finds the matrix $W$ given the vectors $v_{ij}$, see Claim 2. Finally, we remark that the proper learning algorithm of Theorem 1 results in a compact and efficient sorting function that requires: (i) storing $O(k)$ weight vectors as opposed to the initial $O(k^2)$ vectors of the improper learner; and (ii) evaluating $k$ inner products with $x$ to find its ranking (instead of $O(k^2)$).

**Learning in top-$r$ Disagreement.** We next turn our attention to the more challenging top-$r$ ranking disagreement metric. In particular, suppose that we are interested in recovering only the top element of the ranking. One approach would be to directly use the improper learning algorithm for this task and ask for KT distance of order roughly $\varepsilon/k^2$. The resulting hypothesis would produce good predictions for the top element but the required sample complexity would be $O(dk^2)$. While it seems that training $O(k^2)$ $d$-dimensional binary classifiers inherently requires $O(dk^2)$ samples, we show that, using the proper KT distance learning algorithm of Theorem 1, we can also obtain improved sample complexity results for the top-$r$ metric. Our main technical contribution here is a novel estimate of the top-$r$ disagreement in terms of the angle metric. In general, one can show that the top-$r$ disagreement is at most $O(k^2)$ $d_{\text{angle}}(W, W^*)$. We significantly sharpen this estimate by showing the following lemma.

**Lemma 1 (Top-$r$ Disagreement via Parameter Distance).** Consider two matrices $W, W^* \in \mathbb{R}^{k \times d}$ and let $\mathcal{N}_d$ be the standard Gaussian in $d$ dimensions. We have that

$$\Pr_{x \sim \mathcal{N}_d} [\sigma_{1..r}(Wx) \neq \sigma_{1..r}(W^*x)] \leq \tilde{O}(kr) \ d_{\text{angle}}(W, W^*).$$

We remark that Lemma 1 is a general geometric tool that we believe will be useful in other distribution-specific multiclass learning settings. The proof of Lemma 1 mainly relies on geometric Gaussian surface area computations that we believe are of independent interest. For the details, we refer the reader to Section 4. An interesting question with a convex-geometric flavor is whether the sharp bound of Lemma 1 also holds under the more general class of isotropic log-concave distributions.

**1.4 Related Work**

**Robust Supervised Learning.** We start with a summary of prior work on PAC learning with Massart noise. The Massart noise model was formally defined in [MN06] but similar variants had been defined by Vapnik, Sloan and Rivest [Vap06, Slo88, Slo92, RS94, Slo96]. This model is a strict extension of the Random Classification Noise (RCN) model [AL88], where the label noise is uniform, i.e., context-independent and is a special case of the agnostic model [Hau18, KSS94], where the label noise is fully adversarial and computational barriers are known to exist [GR09, FGKP06, Dan16, DKKZ20, GGK20, DKPZ21, HSSVG22]. Our work partially builds upon on the algorithmic task of
PAC learning halfspaces with Massart noise [BH20]. In the distribution-independent setting, known efficient algorithms [DGT19, CKMY20, DKT21] achieve error $\eta + \epsilon$ and the works of [DK20, NT22] indicate that this error bound is the best possible in the Statistical Query model [Kea98]. This lower bound motivates the study of the distribution-specific setting (which is also the case of our work).

There is an extensive line of work in this direction: [ABHU15, ABHZ16, YZ17, ZLC17, BZ17, MV19, DKTZ20, ZSA20, ZL21] with the currently best algorithms succeeding for all $\eta < 1/2$ with a sample and computational complexity poly($d, 1/\epsilon, 1/(1-2\eta)$) under a class of distributions including isotropic log-concave distributions. For details, see [DKK+21]. In this work we focus on Gaussian marginals but some of our results extend to larger distribution classes.

**Label Ranking.** Our work lies in the area of Label Ranking, which has received significant attention over the years [SS07, HFCB08, CH08, HPRZ03, FHMB08, DSM03]. There are multiple approaches for tackling this problem (see [VG10], [ZLY+14]). Some of them are based on probabilistic models [CH08, CDH10, GDV12, ZLCC14] or may be tree based, such as decision trees [CHH09], entropy based ranking trees and forests [RdRSK15, dSSK17], bagging techniques [AGM17] and random forests [ZQ18]. There are also works focusing on supervised clustering [GDV13]. Finally, [CH08, CDH10, CHH09] adopt an instance-based approaches using nearest neighbors approaches. The above results are industrial. From a theoretical perspective, LR has been mainly studied from a statistical learning theory framework [CV20, CKS18, KGB18, KCS17]. [FKP21] provide some computational guarantees for the performance of decision trees in the noiseless case and some experimental results on the robustness of random forests to noise. The setting of [DGR+14] is close to ours but is investigated from an experimental standpoint. We remark that while reducing LR to multiple binary classification tasks has been used in prior literature [HFCB08, CH12, FKP21], standard reductions can not tolerate noise in rankings (nevertheless, from an experimental perspective, e.g., random forests seem robust to noise but lack formal theoretical guarantees). Our reduction crucially relies on the existence of efficient learning algorithms for binary linear classification with Massart noise.

## 2 Notation and Preliminaries

**General Notation.** We use $\tilde{O}(\cdot)$ to omit poly-logarithmic factors. A learning algorithm has sample-polynomial runtime if it runs in time polynomial in the size of the description of the input training set. We denote vectors by boldface $\mathbf{x}$ (with elements $x_i$) and matrices with $W$, where we let $W_i \in \mathbb{R}^d$ denote the $i$-th row of $W \in \mathbb{R}^{k \times d}$ and $W_{ij}$ its elements. We denote $a \cdot b$ the inner product of two vectors and $\theta(a, b)$ their angle. Let $\mathcal{N}_d$ denote the $d$-dimensional standard normal and $\Gamma(\cdot)$ the Gaussian surface area.

**Rankings.** We let $\text{argsort}_{i \in [k]} v$ denote the ranking of $[k]$ in decreasing order according to the values of $v$. For a ranking $\pi$, we let $\pi(i)$ denote the position of the $i$-th element. If $\pi = \pi(x)$, we may also write $\pi(x)(i)$ to denote the position of $i$. We often refer to the elements of a ranking as *alternatives*. For a ranking $\sigma$, we let $\sigma_{1:r}$ denote the top-$r$ part of $\sigma$. When $\sigma = \sigma(x)$, we may also write $\sigma_{1:r}(x)$ and $\sigma_l(x)$ will be the alternative at the $l$-th position. We let $\Delta_{\text{KT}}$ denote the (normalized) KT distance, i.e., $\Delta_{\text{KT}}(\pi, \tau) = \sum_{i < j} 1\{i \succtau j\}/(\binom{k}{2})$ for $\pi, \tau \in \mathcal{S}_k$.

## 3 Learning in KT distance: Theorem 1

In this section, we present the main tools required to obtain our proper learning algorithm of Theorem 1. Our proper algorithm adopts a two-step approach: it first invokes an efficient *improper* algorithm which, instead of a linear sorting function (i.e., a matrix $W \in \mathbb{R}^{k \times d}$), outputs a list of $O(k^2)$ linear classifiers. We then design a novel convex program in order to find the matrix $W$ satisfying the guarantees of Theorem 1. Let us begin with the improper learner for LSFs with bounded noise with respect to the KT distance, whose description can be found in Algorithm 1.

### 3.1 Improper Learning Algorithm

Let us assume that the target function is $\sigma^*(x) = \sigma_W^*(x) = \text{argsort}(W^*x)$ for some $W^* \in \mathbb{R}^{k \times d}$.

**Step 1: Binary decomposition and Noise Structure.** For each drawn example $(x, \pi)$ from the $\eta$-noisy linear label ranking distribution $\mathcal{D}$ (see Definition 2), we create $\binom{k}{2}$ binary examples $(x, y_{ij})$.
Proper Learning Algorithm

Algorithm 2 will efficiently give the matrix \( W_y \) with \( \epsilon < \delta \). Initially, the algorithm starts similarly with the improper learner and obtains a collection of binary classifiers. Having obtained the improper learning algorithm, we can now describe our properAlgorithm 2.

Step 1: Calling Non-proper Learners. As a first step, the algorithm calls Algorithm 1 with parameters \( \epsilon, \delta \) and \( \eta \in [0, 1/2] \) and obtains a list of binary classifiers with normal vectors \( \nu_{ij} \) for \( i < j \). Without loss of generality, assume that \( \| \nu_{ij} \|_2 = 1 \).

**Algorithm 1** Non-proper Learning Algorithm ImproperLSF

| Input: Training set \( T = \{ (x^i, \pi^i) \}_{i \in [N]} \), \( \epsilon, \delta \in (0, 1) \), \( \eta \in [0, 1/2] \) |
| Output: Sorting function \( h : \mathbb{R}^d \rightarrow S_k \) |

For any \( 1 \leq i < j \leq k \), create \( T_{ij} = \{ (x^i, \pi^i(i) - \pi^j(j)) \} \)

For any \( 1 \leq i < j \leq k \), compute \( \nu_{ij} = \text{MassartLTF}(T_{ij}, \frac{\epsilon}{\log k}, \eta) \) \( \triangleright \) See Appendix A.1

**Ranking Phase**: Given \( x \in \mathbb{R}^d \):

(a) Construct directed graph \( G \) with \( V(G) = [k] \) and edges \( e_{i \rightarrow j} \) only if \( \nu_{ij} : x > 0 \forall i \neq j \)

(b) Output \( h(x) = \text{MFAS}(G) \) \( \triangleright \) See Appendix A.1

with \( y_{ij} = \text{sgn}(\pi(i) - \pi(j)) \) for any \( 1 \leq i < j \leq k \). We have that

\[
\Pr_{x \sim \mathcal{D}} [ y_{ij} : \text{sgn}(W_i^* - W_j^* \cdot x) < 0 | x ] = \Pr_{\pi \sim \mathcal{M}(\sigma^*(x))} [ \pi(i) < \pi(j) | W_i^* \cdot x < W_j^* \cdot x ].
\]

Since \( \mathcal{M}(\sigma^*(x)) \) is an \( \eta \)-noisy ranking distribution (see Definition 1), we get that the above quantity is at most \( \eta < 1/2 \). Therefore, each sample \( (x, y_{ij}) \) can be viewed as a sample from a distribution \( D_{ij} \) with Gaussian \( x \)-marginal, optimal linear classifier \( \text{sgn}(W_i^* - W_j^* \cdot x) \), and Massart noise \( \eta \). Hence, we have reduced the task of learning noisy LSFs to a number of \( \binom{n}{2} \) sub-problems concerning the learnability of halfspaces in the presence of bounded (Massart) noise.

**Step 2**: Solving Binary Sub-problems. We can now apply the algorithm \( \text{MassartLTF} \) for LTFs with Massart noise under standard Gaussian marginals [ZSA20] (for details, see Appendix A.1): for all the pairs of alternatives \( 1 \leq i < j \leq k \) with accuracy parameter \( \epsilon' \), confidence \( \delta' = O(\delta/k^2) \), and a total number of \( N = \tilde{\Omega}(\frac{d^d}{\epsilon'^4 \delta^4}) \) i.i.d. samples from \( D \), we can obtain a collection of linear classifiers with normal vectors \( \nu_{ij} \) for \( i < j \). We remark that each one of these halfspaces \( \nu_{ij} \) achieves \( \epsilon \) disagreement with the ground-truth halfspaces \( W_i^* - W_j^* \) with high probability, i.e.,

\[
\Pr_{x \sim \mathcal{N}_d} [ \text{sgn}(\nu_{ij} : x) \neq \text{sgn}(W_i^* - W_j^* \cdot x) ] \leq \epsilon'.
\]

**Step 3**: Ranking Phase. We now have to aggregate the linear classifiers and compute a single sorting function \( h : \mathbb{R}^d \rightarrow S_k \). Given an example \( x \), we create the tournament graph \( G \) with \( k \) nodes that contains a directed edge \( e_{i \rightarrow j} \) if \( \nu_{ij} : x > 0 \). If \( G \) is acyclic, we output the induced permutation; otherwise, the graph contains cycles which should be eliminated. In order to output a ranking, we remove cycles from \( G \) with an efficient, 3-approximation algorithm for MFAS [ACN08, VZW09]. Hence, the output \( h(x) \) and the true target \( \sigma^*(x) \) will have \( E_{x \sim \mathcal{N}_d}[\Delta_{KT}(h(x), \sigma^*(x))] \leq \epsilon' + 3\epsilon' = 4\epsilon' \). This last equation indicates why a constant factor approximation algorithm suffices for our purposes – we can always pick \( \epsilon' = \epsilon/4 \) and complete the proof. For details, see Appendix A.1.

**3.2 Proper Learning Algorithm: Theorem 1**

Having obtained the improper learning algorithm, we can now describe our proper Algorithm 2. Initially, the algorithm starts similarly with the improper learner and obtains a collection of binary linear classifiers. The crucial idea is the next step: the design of an appropriate convex program which will efficiently give the matrix \( W \). We proceed with the details. For the proof, see Appendix A.2.

**Algorithm 2** Proper Learning Algorithm ProperLSF

| Input: Training set \( T = \{ (x^i, \pi^i) \}_{i \in [N]} \), \( \epsilon, \delta \in (0, 1) \), \( \eta \in [0, 1/2] \) |
| Output: Linear Sorting function \( h : \mathbb{R}^d \rightarrow S_k \), i.e., \( h(\cdot) = \sigma_W(\cdot) \) for some matrix \( W \in \mathbb{R}^{k \times d} \) |

\[
\text{Compute } (\nu_{ij})_{1 \leq i < j \leq k} = \text{ImproperLSF}(T, \epsilon, \delta, \eta) \quad \triangleright \text{See Algorithm 1}
\]

Setup the CP 1 and compute \( W = \text{Ellipsoid}(\text{CP}) \) \( \triangleright \text{See Appendix A.2}

**Ranking Phase**: Given \( x \in \mathbb{R}^d \), output \( h(x) = \text{argsort}(Wx) \)

| Step 1: Calling Non-proper Learners. As a first step, the algorithm calls Algorithm 1 with parameters \( \epsilon, \delta \) and \( \eta \in [0, 1/2] \) and obtains a list of linear classifiers with normal vectors \( \nu_{ij} \) for \( i < j \). Without loss of generality, assume that \( \| \nu_{ij} \|_2 = 1 \).
Step 2: Designing and Solving the CP 1. Our main goal is to find a matrix $W$ whose LSF is close to the true target in KT distance. We show the following lemma that connects the KT distance between two LSFs with the angle metric $d_{\text{angle}}(\cdot, \cdot)$ defined in Eq. (1). The proof can be found in the Appendix A.2.

**Lemma 2.** For $W, W^* \in \mathbb{R}^{k \times d}$, it holds $\mathbb{E}_{x \sim N_0}[\Delta_{\text{KT}}(\sigma_W(x), \sigma_{W^*}(x))] \leq d_{\text{angle}}(W, W^*)$.

The above lemma states that, for our purposes, it suffices to control the $d_{\text{angle}}$ metric between the guess $W$ and the true matrix $W^*$. It turns out that, given the binary classifiers $v_{ij}$, we can design a convex program whose solution will satisfy this property. Thinking of the binary classifier as a proxy for $W^* - W^*$, we want each difference $W_i - W_j$ to have small angle with $v_{ij}$ or equivalently to have large correlation with it, i.e., $(W_i - W_j) \cdot v_{ij} \approx ||W_i - W_j||_2$. To enforce this condition, we can therefore use the second order conic constraint $(W_i - W_j) \cdot v_{ij} \geq (1 - \phi)||W_i - W_j||_2$.

We formulate the following convex program 1 with variable the matrix $W$:

\[
\begin{align*}
\text{Find} & \quad W \in \mathbb{R}^{k \times d}, \quad ||W||_F \leq 1, \\
\text{such that} & \quad (W_i - W_j) \cdot v_{ij} \geq (1 - \phi) \cdot ||W_i - W_j||_2 \quad \text{for any} \ 1 \leq i < j \leq k, \quad (1)
\end{align*}
\]

for some $\phi \in (0, 1)$ to be decided. Intuitively, since any $v_{ij}$ has good correlation with $W_i^* - W_j^*$ (by the guarantees of the improper learning algorithm) and the CP 1 requires that its solution $W$ similarly correlates well with $v_{ij}$, we expect that $d_{\text{angle}}(W, W^*)$ will be small. We show that:

**Claim 1.** The convex program 1 is feasible and any solution $W$ of 1 satisfies $d_{\text{angle}}(W, W^*) \leq \epsilon$.

To see this, note that any solution of CP 1 is a matrix $W$ whose angle metric (see Eq. (1)) with the true matrix is small by an application of the triangle inequality between the angles of $(v_{ij}, W_i - W_j)$ and $(v_{ij}, W_i^* - W_j^*)$ for any $i \neq j$. We next have to deal with the feasibility of CP 1. Our goal is to determine the value of $\phi$ that makes the CP 1 feasible. For the pair $1 \leq i < j \leq k$, the guess $v_{ij}$ and the true normal vector $W_i^* - W_j^*$ satisfy, with high probability,

\[
\Pr_{x \sim D_i}[\sgn(v_{ij} \cdot x) \neq \sgn((W_i^* - W_j^*) \cdot x)] \leq \epsilon. \quad (2)
\]

Under the Gaussian distribution (which is rotationally symmetric), it is well known that the angle $\theta(u, v)$ between two vectors $u, v \in \mathbb{R}^{d}$ is equal to $\pi \cdot \Pr_{x \sim N_0}[\sgn(u \cdot x) \neq \sgn(v \cdot x)]$. Hence, using Eq. (2), we get that the angle between the guess $v_{ij}$ and the true normal vector $W_i^* - W_j^*$ is $\theta(W_i^* - W_j^*, v_{ij}) \leq \epsilon$. For sufficiently small $\epsilon$, this bound implies that the cosine of the above angle is of order $1 - (\epsilon c)^2$ and so the following inequality will hold (since $v_{ij}$ is unit):

\[
(W_i^* - W_j^*) \cdot v_{ij} \geq (1 - 2(\epsilon c)^2) \cdot ||W_i^* - W_j^*||_2.
\]

Hence, by setting $\phi = 2(\epsilon c)^2$, the convex program 1 with variables $W \in \mathbb{R}^{k \times d}$ will be feasible; since $||W^*||_F \leq 1$ comes without loss of generality, $W^*$ will be a solution with probability $1 - \delta$.

Next, we have to control the volume of the feasible region. This is crucial in order to apply the ellipsoid algorithm (for details, see in Appendix A.2.1) and, hence, solve the convex program. We show the following claim (see Appendix A.2.1 for the proof):

**Claim 2.** There exists $\rho \geq 2^{-\text{poly}(d,k,1/\epsilon, \log(1/\delta))}$ so that the feasible set of CP 1 with $\phi = O(\epsilon^2)$ contains a ball (with respect to the Frobenius norm) of radius $\rho$.

Critically, the runtime of the ellipsoid algorithm is logarithmic in $1/\rho$. So, the ellipsoid runs in time polynomial in the parameters of the problem and outputs the desired matrix $W$.

4 Learning in top-$r$ Disagreement: Theorem 2

In this section we show that the proper learning algorithm of Section 3.2 learns noisy LSFs in the top-$r$ disagreement metric. We have seen that, with $O(d \log(k)/\epsilon)$ samples, Algorithm 2 of Section 3.2 computes a matrix $W$ such that $d_{\text{angle}}(W, W^*) \leq \epsilon$, see Claim 1. Let us be more specific. Lemma 2 relates the expected KT distance with the angle metric of the two matrices (see also Equation (1)). Our Algorithm 2 essentially gives an upper bound on this angle metric. When we shift our objective and our goal is to control the top-$r$ disagreement, we can still apply Algorithm 2 which essentially controls the angle metric. The crucial ingredient that is missing is the relation between the loss we
have to control, i.e., the expected top-$r$ disagreement and the angle metric of Equation 1. This relation is presented right after and essentially says that the expected top-$r$ disagreement is at most $O(kr)$ times this angle metric. Hence, in order to get top-$r$ disagreement of order $\epsilon$, it suffices to apply our Algorithm 2 with $\epsilon' = O(\epsilon/(kr))$.

We continue with our main contribution which is the following lemma that connects the top-$r$ disagreement metric with the geometric distance $d_{\text{angle}}(\cdot, \cdot)$, recall Lemma 1. To keep this sketch simple we shall present a sketch of the proof of Lemma 1 for the special case of top-1 classification, which we restate below. The proof of the top-1 case can be found at the Appendix B. The detailed proof of the general case ($r>1$) can be found in the Appendix C.

**Lemma 3 (Top-1 Disagreement Loss via $d_{\text{angle}}(\cdot, \cdot)$).** Consider two matrices $U, V \in \mathbb{R}^{k \times d}$ and let $N_d$ be the standard Gaussian in $d$ dimensions. We have that

$$\Pr_{x \sim N_d} [\sigma_1(Ux) \neq \sigma_1(Vx)] \leq O \left( k \sqrt{\log k} \right) d_{\text{angle}}(U, V).$$

We observe that

$$\Pr_{x \sim N_d} [\sigma_1(Ux) \neq \sigma_1(Vx)] = \sum_{i \in [k]} \Pr_{x \sim N_d} [\sigma_1(Ux) = i, \sigma_1(Vx) \neq i].$$

(1)

We denote by $C_U^{(i)} \triangleq 1 \{ x : \sigma_1(Ux) = i \} = \prod_{j \neq i} 1 \{(U_i - U_j) \cdot x \geq 0\}$, i.e., this is the set where the ranking corresponding to $U$ picks $i$ as the top element. Note that $C_U^{(i)}$ is the indicator of a homogeneous polyhedral cone since it can be written as the intersection of homogeneous halfspaces. Using these cones we can rewrite the top-1 disagreement of Eq. (1) as

$$\Pr_{x \sim N_d} [\sigma_1(Ux) \neq \sigma_1(Vx)] = \sum_{i \in [k]} \Pr_{x \sim N_d} [C_U^{(i)}(x) = 1, C_V^{(i)}(x) = 0].$$

(2)

Hence, our task is to control the mass of the disagreement region of two cones. The next Lemma 4 achieves this task and, combined with Eq. (2) directly gives the conclusion of Lemma 3.

Next we work with two general homogeneous polyhedral cones with set indicators $C_1, C_2$:

**Lemma 4 (Cone Disagreement).** Let $C_1, C_2 : \mathbb{R}^d \rightarrow \{0, 1\}$ be homogeneous polyhedral cones defined by the $k$ unit vectors $v_1, \ldots, v_k$ and $u_1, \ldots, u_k$ respectively. For some universal constant $c > 0$, it holds that $\Pr_{x \sim N_d} [C_1(x) \neq C_2(x)] \leq c \sqrt{\log k} \max_{i \in [k]} \theta(v_i, u_i)$.

**Roadmap of the Proof of Lemma 4:** Assume that we rotate one face of the polyhedral cone $C_1$ by a very small angle $\theta$ to obtain the perturbed cone $C_2$. At a high-level, we expect the probability of the disagreement region between the new cone $C_2$ and $C_1$ to be roughly (this is an underestimation) equal to the size of the perturbation $\theta$ times the (Gaussian) surface area of the face of the convex cone that we perturbed. The Gaussian Surface Area (GSA) of a convex set $A \subset \mathbb{R}^d$, is defined as $\Gamma(A) \triangleq \int_A \phi_d(x) d\mu(x)$, where $d\mu(x)$ is the standard surface measure in $\mathbb{R}^d$ and $\phi_d(x) = (2\pi)^{-d/2} \exp(-\|x\|^2/2)$. In fact, in Claim 5 below, we show that the total probability of the disagreement region is bounded above by the perturbation size $\theta$ times the sum of the Gaussian surface area of every face (times a logarithmic blow-up factor):

$$\Pr_{x \sim N_d} [C_1(x) \neq C_2(x)] \leq O(\theta) \sum_{i=1}^k \Gamma(F_i) \sqrt{\log(1/\Gamma(F_i) + 1)}.$$ 

Surprisingly, for homogeneous convex cones, the above sum cannot grow very fast with $k$. In fact, we show that it can be at most $O(\sqrt{\log k})$. To prove this, we crucially rely on the following convex geometry result showing that the Gaussian surface area of a homogeneous convex cone is $O(1)$ regardless of the number of its faces $k$.

**Lemma 5 ([Naz03]).** Let $C$ be a homogeneous polyhedral cone with $k$ faces $F_1, \ldots, F_k$. Then $C$ has Gaussian surface area $\Gamma(C) = \sum_{i=1}^k \Gamma(F_i) \leq 1$.  

8
Using an inequality similar to the fact that the maximum entropy of a discrete distribution on $k$ elements is at most $\log k$, and, since, from Lemma 5, it holds that $\sum_{i=1}^{k} \log(1/\Gamma(\Gamma(F_i))) = O(\sqrt{\log k})$. Therefore, with the above lemma we conclude that, if the maximum angle perturbation that we perform on $C_1$ is $\theta$, then the probability of the disagreement region is $O(\theta)$. We next give the formal proof resulting in the upper bound of $O(\sqrt{\log \theta})$ for the disagreement.

**Single Face Perturbation Bound:** **Claim 3:** We will use the following notation for the positive orthant indicator $R(z) = \prod_{i=1}^{k} \mathbf{1}\{z_i \geq 0\}$. Notice that the homogeneous polyhedral cone $C_1$ can be written as $C_1(x) = R(Vx) = R(v_1 \cdot x, \ldots, v_k \cdot x)$. Claim 3 below shows that the disagreement of two cones that differ on a single normal vector is bounded by above by the Gaussian surface area of a particular face $F_1$ times a logarithmic blow-up factor $\sqrt{\log(1/\Gamma(F_1))} + 1$.

**Claim 3.** Let $v_1, \ldots, v_k \in \mathbb{R}^d$ and $r \in \mathbb{R}^d$ with $\theta(v_1 \cdot r) \leq \theta$ for some sufficiently small $\theta \in (0, \pi/2)$. Let $F_1$ be the face with $v_1 \cdot x = 0$ of the cone $R(Vx)$ and $c > 0$ be some universal constant. Then,

$$
\Pr_{x \sim \mathcal{N}_d} [R(v_1 \cdot x, \ldots, v_k \cdot x) \neq R(r \cdot x, v_2 \cdot x, \ldots, v_k \cdot x)] \leq c \cdot \theta \cdot \Gamma(F_1) \sqrt{\log \left( \frac{1}{\Gamma(F_1)} + 1 \right)}.
$$

**Proof Sketch of Claim 3.** Since the constraints $v_2 : x \geq 0, \ldots, v_k : x \geq 0$ are common in the two cones, we have that $R(v_1 \cdot x, \ldots, v_k \cdot x) \neq R(r \cdot x, v_2 \cdot x, \ldots, v_k \cdot x)$ only when the first “halfspaces” disagree, i.e., when $\theta(v_1 \cdot x)(r \cdot x) < 0$. Thus, we have that the LHS probability of Claim 3 is equal to

$$
\mathbb{E}_{x \sim \mathcal{N}_d} [R(v_2 \cdot x, \ldots, v_k \cdot x) \cdot \mathbf{1}\{|v_1 \cdot x| < 2\theta |q \cdot x|\}].
$$

This expectation contains two terms: the term $R(v_2 \cdot x, \ldots, v_k \cdot x)$ that contains the last $k - 1$ common constraints of the two cones and the region where the first two halfspaces disagree, i.e., the set $\{x : (v_1 \cdot x)(r \cdot x) < 0\}$. In order to upper bound this integral in terms of the angle $\theta$, we observe that (for $\theta$ sufficiently small) it is not hard to show (see Appendix B) that the disagreement region, which is itself a (non-convex) cone, is a subset of the region $\{x : |v_1 \cdot x| \leq 2\theta |q \cdot x|\}$, where $q$ is the normalized projection of $r$ onto the orthogonal complement of $v_1$, i.e., $q = \text{proj}_{v_1} r / \|\text{proj}_{v_1} r\|_2$. Therefore, we have that the integral of Eq. (3) is at most

$$
2 \mathbb{E}_{x \sim \mathcal{N}_d} [R(v_2 \cdot x, \ldots, v_k \cdot x) \cdot \mathbf{1}\{|v_1 \cdot x| \leq 2\theta |q \cdot x|\}] .
$$

This is where the definition of the Gaussian surface area appears. In fact, we have to compute the derivative of the above expression (which is a function of $\theta$) with respect to $\theta$ and evaluate it at $\theta = 0$. The idea behind this computation is that we can upper bound probability mass of the cone disagreement, i.e., the term $\Pr_{x \sim \mathcal{N}_d} [R(v_1 \cdot x, \ldots, v_k \cdot x) \neq R(r \cdot x, v_2 \cdot x, \ldots, v_k \cdot x)]$ by its derivative with respect to $\theta$ (evaluated at 0) times $\theta$ by introducing $o(\theta)$ error. Hence, it suffices to upper bound the value of this derivative at 0, which is:

$$
2 \mathbb{E}_{x \sim \mathcal{N}_d} [R(v_2 \cdot x, \ldots, v_k \cdot x) \cdot |q \cdot x| \cdot \delta(|v_1 \cdot x|)] ,
$$

where $\delta$ is the Dirac delta function. Notice that, if we did not have the term $|q \cdot x|$, the above expression would be exactly equal to two times the Gaussian surface area of the face with $v_1 \cdot x = 0$, i.e., it would be equal to $2\Gamma(F_1)$. We now show that this extra term of $|q \cdot x|$ can only increase the above surface integral by at most a logarithmic factor. For some $\xi$ to be decided, we have that

$$
\mathbb{E}_{x \sim \mathcal{N}_d} [R(v_2 \cdot x, \ldots, v_k \cdot x) \cdot |q \cdot x| \cdot \delta(|v_1 \cdot x|)] = \int_{x \in F_1} \phi_d(x) |q \cdot x| d\mu(x)
$$

$$
\leq \int_{x \in F_1} \phi_d(x) |q \cdot x| \mathbf{1}\{|q \cdot x| \leq \xi\} d\mu(x) + \int_{x \in F_1} \phi_d(x) |q \cdot x| \mathbf{1}\{|q \cdot x| \geq \xi\} d\mu(x)
$$

$$
\leq \xi \int_{x \in F_1} \phi_d(x) d\mu(x) + \int_{x \in F_1} \phi_d(x) |q \cdot x| \mathbf{1}\{|q \cdot x| \geq \xi\} d\mu(x) ,
$$

where $d\mu(x)$ is the standard surface measure in $\mathbb{R}^d$. The first integral above is exactly equal to the Gaussian surface area of the face $F_1$. To bound from above the second term we can use the next claim showing that not a lot of mass of the face $F_1$ can concentrate on the region where $|q \cdot x|$ is very large. Its proof relies on standard Gaussian concentration arguments, and is provided in Appendix B.
Claim 4. It holds that
\[ \int_{x \in F_1} \phi_d(x) |q \cdot x| 1 \{ |q \cdot x| \geq \xi \} d\mu(x) \leq O(\exp(-\xi^2/2)). \]

Using the above result, we get that
\[ \frac{d}{d\theta} \left( \mathbb{E}_{x \sim N_d} [R(v_2 \cdot x, \ldots, v_k \cdot x) 1 \{ |v_1 \cdot x| \leq 2\theta |q \cdot x| \}] \right) \bigg|_{\theta=0} \leq O(\xi) \Gamma(F_1) + O(\exp(-\xi^2/2)). \]

By picking \( \xi = \Theta(\sqrt{\log(1 + 1/\Gamma(F_1))}) \), the result follows since, up to introducing \( o(\theta) \) error, we can bound the term \( \mathbb{P}_{x \sim N_d} [R(v_1 \cdot x, \ldots, v_k \cdot x) \neq R(r \cdot x, v_2 \cdot x, \ldots, v_k \cdot x)] \) by its derivative with respect to \( \theta \), evaluated at 0, times \( \theta \).

\[ \square \]

Conclusion. Our work presents the first theoretical guarantees for (linear) LR with noise and settles interesting directions for future work, as mentioned in Section 1. This paper is theoretical and does not have any negative social impact.

Acknowledgments and Disclosure of Funding

Dimitris Fotakis and Alkis Kalavasis were supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant”, project BALSAM, HFRIFM17-1424.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We have presented the assumptions of our models and we have explained future research directions in Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We have discussed that in the Conclusion paragraph at the end of the main body.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We have specified the theoretical models we are working on, see e.g., Section 1.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space constraints we have included the full proofs of the results in the Supp. Material.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]