Generalization error bounds for learning to rank: Does the length of document lists matter?

Ambuj Tewari TEWARIA@UMICH.EDU
University of Michigan, Ann Arbor

Sougata Chaudhuri SOUGATA@UMICH.EDU
University of Michigan, Ann Arbor

Abstract

We consider the generalization ability of algorithms for learning to rank at a query level, a problem also called subset ranking. Existing generalization error bounds necessarily degrade as the size of the document list associated with a query increases. We show that such a degradation is not intrinsic to the problem. For several loss functions, including the cross-entropy loss used in the well known ListNet method, there is *no* degradation in generalization ability as document lists become longer. We also provide novel generalization error bounds under ℓ_1 regularization and faster convergence rates if the loss function is smooth.

1. Introduction

Learning to rank at the query level has emerged as an exciting research area at the intersection of information retrieval and machine learning. Training data in learning to rank consists of queries along with associated documents, where documents are represented as feature vectors. For each query, the documents are labeled with human relevance judgements. The goal at training time is to learn a ranking function that can, for a future query, rank its associated documents in order of their relevance to the query. The performance of ranking functions on test sets is evaluated using a variety of performance measures such as NDCG (Järvelin & Kekäläinen, 2002), ERR (Chapelle et al., 2009) or Average Precision (Yue et al., 2007).

The performance measures used for testing ranking methods cannot be directly optimized during training time as they lead to discontinuous optimization problems. As a result, researchers often minimize *surrogate* loss functions that are easier to optimize. For example, one might consider smoothed versions of, or convex upper bounds on, the target performance measure. However, as soon as one optimizes a surrogate loss, one has to deal with two questions (Chapelle et al., 2011). First, does minimizing the surrogate on finite training data imply small expected surrogate loss on infinite unseen data? Second, does small expected surrogate loss on infinite unseen data imply small *target* loss on infinite unseen data? The first issue is one of *generalization error bounds* for empirical risk minimization (ERM) algorithms that minimize surrogate loss on training data. The second issue is one of *calibration*: does consistency in the surrogate loss imply consistency in the target loss?

This paper deals with the former issue, viz. that of generalization error bounds for surrogate loss minimization. In pioneering works, Lan et al. (2008; 2009) gave generalization error bounds for learning to rank algorithms. However, while the former paper was restricted to analysis of pairwise approach to learning to rank, the later paper was limited to results on just three surrogates: ListMLE, ListNet and RankCosine. To the best of our knowledge, the most generally applicable bound on the generalization error of query-level learning to rank algorithms has been obtained by Chapelle & Wu (2010).

The bound of Chapelle & Wu (2010), while generally applicable, does have an explicit dependence on the *length* of the document list associated with a query. Our investigations begin with this simple question: is an explicit dependence on the length of document lists unavoidable in generalization error bounds for query-level learning to rank algorithms? We focus on the prevalent technique in literature where learning to rank algorithms learn linear scoring functions and obtain ranking by sorting scores in descending order. Our first contribution (Theorem 3) is to show that dimension of linear scoring functions that are *permutation invariant* (a necessary condition for being valid scoring functions for learning to rank) has no dependence on the length of document lists. Our second contribution (Theorems 5,

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Table 1. A comparison of three bounds given in this paper for Lipschitz loss functions. Criteria for comparison: algorithm bound applies to (OGD = Online Gradient Descent, [R]ERM = [Regularized] Empirical Risk Minimization), whether it applies to general (possibly non-convex) losses, and whether the constants involved are tight.

Bound	Applies to	Handles Nonconvex Loss	"Constant" hidden in $O(\cdot)$ notation
Theorem 5	OGD	No	Smallest
Theorem 6	RERM	No	Small
Theorem 9	ERM	Yes	Hides several logarithmic factors

6, 9) is to show that as long as one uses the "right" norm in defining the Lipschitz constant of the surrogate loss, we can derive generalization error bounds that have *no explicit dependence on the length of document lists*. The reason that the second contribution involves three bounds is that they all have different strengths and scopes of application (See Table 1 for a comparison). Our final contribution is to provide novel generalization error bounds for learning to rank in two previously unexplored settings: almost dimension independent bounds when using high dimensional features with ℓ_1 regularization (Theorem 12) and "optimistic" rates (that can be as fast as $O(1/n)$) when the loss function is smooth (Theorem 17). We also apply our results on popular convex and non-convex surrogates. All omitted proofs can be found in the appendix (see supplementary material).

2. Preliminaries

In learning to rank (also called subset ranking to distinguish it from other related problems, e.g., bipartite ranking), a training example is of the form $((q, d_1, \ldots, d_m), y)$. Here q is a search query and $d_1, \ldots, d_m$ are m documents with varying degrees of *relevance* to the query. Human labelers provide the relevance vector $y \in \mathbb{R}^m$ where the entries in y contain the relevance labels for the m individual documents. Typically, y has integer-valued entries in the range $\{0, \ldots, Y_{\max}\}$ where $Y_{\max}$ is often less than 5. For our theoretical analysis, we get rid of some of these details by assuming that some feature map Ψ exists to map a query document pair (q, d) to $\mathbb{R}^d$. As a result, the training example $((q, d_1, \ldots, d_m), y)$ gets converted into (X, y) where $X = [\Psi(q, d_1), \ldots, \Psi(q, d_m)]^\top$ is an $m \times d$ matrix with the m query-document feature vector as rows. With this abstraction, we have an input space $\mathcal{X} \subseteq \mathbb{R}^{m \times d}$ and a label space $\mathcal{Y} \subseteq \mathbb{R}^m$.

A training set consists of iid examples $(X^{(1)}, y^{(1)}), \ldots, (X^{(n)}, y^{(n)})$ drawn from some underlying distribution D. To rank the documents in an instance $X \in \mathcal{X}$, often a score vector $s \in \mathbb{R}^m$ is computed. A ranking of the documents can then be obtained from s by sorting its entries in decreasing order. A common choice for the scoring function is to make it *linear* in the input X and consider the following class of vector-valued functions:

$$\mathcal{F}_{\text{lin}} = \{X \mapsto Xw \;:\; X \in \mathbb{R}^{m \times d}, w \in \mathbb{R}^d\}. \tag{1}$$

Depending upon the regularization, we also consider the following two subclasses of $\mathcal{F}_{\text{lin}}$:

$$\mathcal{F}_2 := \{X \mapsto Xw \;:\; X \in \mathbb{R}^{m \times d}, w \in \mathbb{R}^d, \|w\|_2 \le W_2\},$$
$$\mathcal{F}_1 := \{X \mapsto Xw \;:\; X \in \mathbb{R}^{m \times d}, w \in \mathbb{R}^d, \|w\|_1 \le W_1\}.$$

In the input space $\mathcal{X}$, it is natural for the rows of X to have a bound on the appropriate dual norm. Accordingly, whenever we use $\mathcal{F}_2$, the input space is set to $\mathcal{X} = \{X \in \mathbb{R}^{m \times d} : \forall j \in [m],\ \|X_j\|_2 \le R_X\}$ where X_j denotes jth row of X and $[m] := \{1, \ldots, m\}$. Similarly, when we use $\mathcal{F}_1$, we set $\mathcal{X} = \{X \in \mathbb{R}^{m \times d} \;:\; \forall j \in [m],\ \|X_j\|_\infty \le \bar{R}_X\}$. These are natural counterparts to the following function classes studied in binary classification and regression:

$$\mathcal{G}_2 := \{x \mapsto \langle x, w\rangle \;:\; \|x\|_2 \le R_X, w \in \mathbb{R}^d, \|w\|_2 \le W_2\},$$
$$\mathcal{G}_1 := \{x \mapsto \langle x, w\rangle \;:\; \|x\|_\infty \le \bar{R}_X, w \in \mathbb{R}^d, \|w\|_1 \le W_1\}.$$

A key ingredient in the basic setup of the learning to rank problem is a loss function $\phi : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}_+$ where $\mathbb{R}_+$ denotes the set of non-negative real numbers. Given a class $\mathcal{F}$ of vector-valued functions, a loss ϕ yields a natural loss class: namely the class of real-valued functions that one gets by composing ϕ with functions in $\mathcal{F}$:

$$\phi \circ \mathcal{F} := \{(X, y) \mapsto \phi(f(X), y) \;:\; X \in \mathbb{R}^{m \times d}, f \in \mathcal{F}\}.$$

For vector valued scores, the Lipschitz constant of ϕ depends on the norm $|||\cdot|||$ that we decide to use in the score space ($|||\cdot|||_\star$ is dual of $|||\cdot|||$):

$$\forall y \in \mathcal{Y}, s, s' \in \mathbb{R}^m,\ |\phi(s_1, y) - \phi(s_2, y)| \le G_\phi |||s_1 - s_2|||.$$

If ϕ is differentiable, this is equivalent to: $\forall y \in \mathcal{Y}, s \in \mathbb{R}^m,\ |||\nabla_s \phi(s, y)|||_\star \le G_\phi$. Similarly, the smoothness constant H_ϕ of ϕ defined as: $\forall y \in \mathcal{Y}, s, s' \in \mathbb{R}^m$,

$$|||\nabla_s \phi(s_1, y) - \nabla_s \phi(s_2, y)|||_\star \le H_\phi |||s_1 - s_2|||.$$

also depends on the norm used in the score space. If ϕ is twice differentiable, the above inequality is equivalent to

$$\forall y \in \mathcal{Y}, s \in \mathbb{R}^m,\ |||\nabla_s^2 \phi(s, y)|||_{\text{op}} \le H_\phi$$

where $|||\cdot|||_{\text{op}}$ is the operator norm induced by the pair $|||\cdot|||, |||\cdot|||_\star$ and defined as $|||M|||_{\text{op}} := \sup_{v\neq 0} \frac{|||Mv|||_\star}{|||v|||}$. Define the expected loss of w under the distribution D $L_\phi(w) := \mathbb{E}_{(X,y)\sim D}\left[\phi(Xw, y)\right]$ and its empirical loss on the sample as $\hat{L}_\phi(w) := \frac{1}{n}\sum_{i=1}^n \phi(X^{(i)}w, y^{(i)})$. The minimizer of $L_\phi(w)$ (resp. $\hat{L}_\phi(w)$) over some function class (parameterized by w) will be denoted by $w^\star$ (resp. $\hat{w}$). We may refer to expectations w.r.t. the sample using $\widehat{\mathbb{E}}\left[\cdot\right]$. To reduce notational clutter, we often refer to (X, y) jointly by Z and $\mathcal{X}\times\mathcal{Y}$ by $\mathcal{Z}$. For vectors, $\langle u, v\rangle$ denotes the standard inner product $\sum_i u_i v_i$ and for matrices U, V of the same shape, $\langle U, V\rangle$ means $\text{Tr}(U^\top V) = \sum_{ij} U_{ij}V_{ij}$. The set of $m!$ permutation π of degree m is denoted by S_m. A vector of ones is denoted by $\mathbf{1}$.

3. Application to Specific Losses

To whet the reader's appetite for the technical presentation that follows, we will consider two loss functions, one convex and one non-convex, to illustrate the concrete improvements offered by our new generalization bounds. A generalization bound is of the form: $L_\phi(\hat{w}) \leq L_\phi(w^\star)+$"complexity term". It should be noted that $w^\star$ is not available to the learning algorithm as it needs knowledge of underlying distribution of the data. The complexity term of Chapelle & Wu (2010) is $O(G_\phi^{CW} W_2 R_X \sqrt{m/n})$. The constant G_ϕ^{CW} is the Lipschitz constant of the surrogate ϕ (viewed as a function of the score vector s) w.r.t. ℓ_2 norm. Our bounds will instead be of the form $O(G_\phi W_2 R_X \sqrt{1/n})$, where G_ϕ is the Lipschitz constant of ϕ w.r.t. ℓ_∞ norm. Note that our bounds are free of any explicit m dependence. Also, by definition, $G_\phi \leq G_\phi^{CW}\sqrt{m}$ but the former can be much smaller as the two examples below illustrate. In benchmark datasets (Liu et al., 2007), m can easily be in the 100-1000 range.

3.1. Application to ListNet

The ListNet ranking method (Cao et al., 2007) uses a *convex* surrogate, that is defined in the following way[1]. Define m maps from $\mathbb{R}^m$ to $\mathbb{R}$ as: $P_j(v) = \exp(v_j)/\sum_{i=1}^m \exp(v_i)$ for $j \in [m]$. Then, we have, for $s \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$,

$$\phi_{\text{LN}}(s, y) = -\sum_{j=1}^m P_j(y)\log P_j(s).$$

An easy calculation shows that the Lipschitz (as well as smoothness) constant of ϕ_{LN} is m independent.

Proposition 1. *The Lipschitz (resp. smoothness) constant of ϕ_{LN} w.r.t. $\|\cdot\|_\infty$ satisfies $G_{\phi_{\text{LN}}} \leq 2$ (resp. $H_{\phi_{\text{LN}}} \leq 2$) for any $m \geq 1$.*

Since the bounds above are independent of m, so the generalization bounds resulting from their use in Theorem 9 and Theorem 17 will also be independent of m (up to logarithmic factors). We are not aware of prior generalization bounds for ListNet that do not scale with m. In particular, the results of Lan et al. (2009) have an $m!$ dependence since they consider the top-m version of ListNet. However, even if the top-1 variant above is considered, their proof technique will result in at least a linear dependence on m and does not result in as tight a bound as we get from our general results. It is also easy to see that the Lipschitz constant $G_{\phi_{\text{LN}}}^{CW}$ of ListNet loss w.r.t. ℓ_2 norm is also 2 and hence the bound of Chapelle & Wu (2010) necessarily has a $\sqrt{m}$ dependence in it. Moreover, generalization error bounds for ListNet exploiting its smoothness will interpolate between the pessimistic $1/\sqrt{n}$ and optimistic $1/n$ rates. These have never been provided before.

3.2. Application to Smoothed DCG@1

This example is from the work of Chapelle & Wu (2010). Smoothed DCG@1, a *non-convex* surrogate, is defined as:

$$\phi_{\text{SD}}(s, y) = D(1)\sum_{i=1}^m G(y_i)\frac{\exp(s_i/\sigma)}{\sum_j \exp(s_j/\sigma)},$$

where $D(i) = 1/\log_2(1+i)$ is the "discount" function and $G(i) = 2^i - 1$ is the "gain" function. The amount of smoothing is controlled by the parameter $\sigma > 0$ and the smoothed version approaches DCG@1 as $\sigma \to 0$ (DCG stands for Discounted Cumulative Gain (Järvelin & Kekäläinen, 2002)).

Proposition 2. *The Lipschitz constant of ϕ_{SD} w.r.t. $\|\cdot\|_\infty$ satisfies $G_{\phi_{\text{SD}}} \leq 2D(1)G(Y_{\max})/\sigma$ for any $m \geq 1$. Here $Y_{\max}$ is maximum possible relevance score of a document (usually less than 5).*

As in the ListNet loss case we previously considered, the generalization bound resulting from Theorem 9 will be independent of m. This is intuitively satisfying: DCG@1, whose smoothing we are considering, only depends on the document that is put in the top position by the score vector s (and not on the entire sorted order of s). Our generalization bound does not deteriorate as the total list size m grows. In contrast, the bound of Chapelle & Wu (2010) will necessarily deteriorate as $\sqrt{m}$ since the constant $G_{\phi_{\text{SD}}}^{CW}$ is the same as $G_{\phi_{\text{SD}}}$. Moreover, it should be noted that even in the original SmoothedDCG paper, σ is present in the denominator of $G_{\phi_{\text{SD}}}^{CW}$, so our results are directly comparable. Also note that this example can easily be extended to consider DCG@k for case when document list length $m \gg k$ (a very common scenario in practice).

[1]The ListNet paper actually defines a family of losses based on probability models for top k documents. We use $k = 1$ in our definition since that is the version implemented in their experimental results.

3.3. Application to RankSVM

RankSVM (Joachims, 2002) is another well established ranking method, which minimizes a *convex* surrogate based on pairwise comparisons of documents. A number of studies have shown that ListNet has better empirical performance than RankSVM. One possible reason for the better performance of ListNet over RankSVM is that the Lipschitz constant of RankSVM surrogate w.r.t $\|\cdot\|_\infty$ doe scale with document list size as $O(m^2)$. Due to lack of space, we give the details in the supplement.

4. Does The Length of Document Lists Matter?

Our work is directly motivated by a very interesting generalization bound for learning to rank due to Chapelle & Wu (2010, Theorem 1). They considered a Lipschitz continuous loss ϕ with Lipschitz constant G_ϕ^{CW} w.r.t. *the* ℓ_2 *norm*. They show that, with probability at least $1-\delta$,

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \leq \hat{L}_\phi(w) + 3\, G_\phi^{CW} W_2 R_X \sqrt{\frac{m}{n}} + \sqrt{\frac{8\log(1/\delta)}{n}}.$$

The dominant term on the right is $O(G_\phi^{CW} W_2 R_X \sqrt{m/n})$. In the next three sections, we will derive improved bounds of the form $\tilde{O}(G_\phi W_2 R_X \sqrt{1/n})$ where $G_\phi \leq G_\phi^{CW}\sqrt{m}$ but can be much smaller. Before we do that, let us examine the dimensionality reduction in linear scoring function that is caused by a natural permutation invariance requirement.

4.1. Permutation invariance removes m dependence in dimensionality of linear scoring functions

As stated in Section 2, a ranking is obtained by sorting a score vector obtained via a linear scoring function f. Consider the space of *linear* scoring function that consists of *all* linear maps f that map $\mathbb{R}^{m\times d}$ to $\mathbb{R}^m$:

$$\mathcal{F}_{\text{full}} := \left\{ X \mapsto [\langle X, W_1\rangle, \ldots, \langle X, W_m\rangle]^\top \,:\, W_i \in \mathbb{R}^{m\times d} \right\}.$$

These linear maps are fully parameterized by matrices $W_1, \ldots, W_m$. Thus, a full parameterization of the linear scoring function is of dimension $m^2 d$. Note that the popularly used class of linear scoring functions $\mathcal{F}_{\text{lin}}$ defined in Eq. 1 is actually a low d-dimensional subspace of the full $m^2 d$ dimensional space of all linear maps. It is important to note that the dimension of $\mathcal{F}_{\text{lin}}$ is *independent of* m.

In learning theory, one of the factors influencing the generalization error bound is the richness of the class of hypothesis functions. Since the linear function class $\mathcal{F}_{\text{lin}}$ has dimension independent of m, we intuitively expect that, at least under some conditions, algorithms that minimize ranking losses using linear scoring functions should have an m independent complexity term in the generalization bound. The reader might wonder whether the dimension reduction from $m^2 d$ to d in going from $\mathcal{F}_{\text{full}}$ to $\mathcal{F}_{\text{lin}}$ is arbitrary. To dispel this doubt, we prove the lower dimensional class $\mathcal{F}_{\text{lin}}$ is the *only sensible choice* of linear scoring functions in the learning to rank setting. This is because scoring functions should satisfy a permutation invariance property. That is, if we apply a permutation $\pi \in S_m$ to the rows of X to get a matrix πX then the scores should also simply get permuted by π. That is, we should only consider scoring functions in the following class:

$$\mathcal{F}_{\text{perminv}} = \{f : \forall \pi \in S_m, \forall X \in \mathbb{R}^{m\times d}, \pi f(X) = f(\pi X)\}.$$

The permutation invariance requirement, in turn, forces a reduction from dimension $m^2 d$ to just $2d$ (which has no dependence on m).

Theorem 3. *The intersection of the function classes* $\mathcal{F}_{\text{full}}$ *and* $\mathcal{F}_{\text{perminv}}$ *is the* $2d$*-dimensional class:*

$$\mathcal{F}'_{\text{lin}} = \{X \mapsto Xw + (\mathbf{1}^\top Xv)\mathbf{1} \,:\, w, v \in \mathbb{R}^d\}. \quad (2)$$

Note that the extra degree of freedom provided by the v parameter in Eq. 2 is useless for ranking purposes since adding a constant vector (i.e., a multiple of $\mathbf{1}$) to a score vector has no effect on the sorted order. This is why we said that $\mathcal{F}_{\text{lin}}$ is the only sensible choice of linear scoring functions.

5. Online to Batch Conversion

In this section, we build some intuition as to why it is natural to use $\|\cdot\|_\infty$ in defining the Lipschitz constant of the loss ϕ. To this end, consider the following well known online gradient descent (OGD) regret guarantee. Recall that OGD refers to the simple online algorithm that makes the update $w_{i+1} \leftarrow w_i - \eta \nabla_{w_i} f_i(w_i)$ at time i. If we run OGD to generate w_i's, we have, for all $\|w\|_2 \leq W_2$:

$$\sum_{i=1}^n f_i(w_i) - \sum_{i=1}^n f_i(w) \leq \frac{W_2^2}{2\eta} + \eta G^2 n$$

where G is a bound on the maximum ℓ_2-norm of the gradients $\nabla_{w_i} f_i(w_i)$ and f_i's have to be *convex*. If $(X^{(1)}, y^{(1)}), \ldots, (X^{(n)}, y^{(n)})$ are iid then by setting $f_i(w) = \phi(X^{(i)}w, y^{(i)})$, $1 \leq i \leq n$ we can do an "online to batch conversion". That is, we optimize over η, take expectations and use Jensen's inequality to get the following excess risk bound:

$$\forall \|w\|_2 \leq W_2,\ \mathbb{E}\left[L_\phi(\hat{w}_{\text{OGD}})\right] - L_\phi(w) \leq W_2 G \sqrt{\frac{2}{n}}$$

where $\hat{w}_{\text{OGD}} = \frac{1}{n}\sum_{i=1}^n w_i$ and G has to satisfy (noting that $s = X^{(i)} w_i$)

$$G \geq \|\nabla_{w_i} f_i(w_i)\|_2 = \|(X^{(i)})^\top \nabla_s \phi(X^{(i)} w_i, y^{(i)})\|_2$$

where we use the chain rule to express ∇_w in terms of ∇_s. Finally, we can upper bound

$$\begin{aligned}&\|(X^{(i)})^\top \nabla_s \phi(X^{(i)} w_i, y^{(i)})\|_2 \\ &\le \|(X^{(i)})^\top\|_{1\to 2} \cdot \|\nabla_s \phi(X^{(i)} w_i, y^{(i)})\|_1 \\ &\le R_X \|\nabla_s \phi(X^{(i)} w_i, y^{(i)})\|_1\end{aligned}$$

as $R_X \ge \max_{j=1}^m \|X_j\|_2$ and because of the following lemma.

Lemma 4. *For any $1 \le p \le \infty$,*

$$\begin{aligned}\|X\|_{p\to q} &= \sup_{v\neq 0} \frac{\|Xv\|_q}{\|v\|_p} \\ \|X^\top\|_{1\to p} &= \|X\|_{q\to\infty} = \max_{j=1}^m \|X_j\|_p ,\end{aligned}$$

where q is the dual exponent of p (i.e., $\frac{1}{q} + \frac{1}{p} = 1$).

Thus, we have shown the following result.

Theorem 5. *Let ϕ be convex and have Lipschitz constant G_ϕ w.r.t. $\|\cdot\|_\infty$. Suppose we run online gradient descent (with appropriate step size η) on $f_i(w) = \phi(X^{(i)} w, y^{(i)})$ and return $\hat{w}_{\mathrm{OGD}} = \frac{1}{T}\sum_{i=1}^n w_i$. Then we have,*

$$\forall \|w\|_2 \le W_2,\ \mathbb{E}\left[L_\phi(\hat{w}_{\mathrm{OGD}})\right] - L_\phi(w) \le G_\phi\, W_2\, R_X \sqrt{\frac{2}{n}}.$$

The above excess risk bound has no explicit m dependence. This is encouraging but there are two deficiencies of this approach based on online regret bounds. First, the result applies to the output of a specific algorithm that may not be the method of choice for practitioners. For example, the above argument does not yield uniform convergence bounds that could lead to excess risk bounds for ERM (or regularized versions of it). Second, there is no way to generalize the result to Lipschitz, but *non-convex* loss functions. It may noted here that the original motivation for Chapelle & Wu (2010) to prove their generalization bound was to consider the non-convex loss used in their SmoothRank method. We will address these issues in the next two sections.

6. Stochastic Convex Optimization

We first define the regularized empirical risk minimizer:

$$\hat{w}_\lambda = \operatorname*{argmin}_{\|w\|_2 \le W_2} \frac{\lambda}{2}\|w\|_2^2 + \hat{L}_\phi(w). \tag{3}$$

We now state the main result of this section.

Theorem 6. *Let the loss function ϕ be convex and have Lipschitz constant G_ϕ w.r.t. $\|\cdot\|_\infty$. Then, for an appropriate choice of $\lambda = O(1/\sqrt{n})$, we have*

$$\mathbb{E}\left[L_\phi(\hat{w}_\lambda)\right] \le L_\phi(w^\star) + 2\, G_\phi\, R_X\, W_2 \left(\frac{8}{n} + \sqrt{\frac{2}{n}}\right).$$

This result applies to a batch algorithm (regularized ERM) but unfortunately requires the regularization parameter λ to be set in a particular way. Also, it does not apply to non-convex losses and does not yield uniform convergence bounds. In the next section, we will address these deficiencies. However, we will incur some extra logarithmic factors that are absent in the clean bound above.

7. Bounds for Non-convex Losses

The above discussion suggests that we have a possibility of deriving tighter, possibly m-independent, generalization error bounds by assuming that ϕ is Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. The standard approach in binary classification is to appeal to the Ledoux-Talagrand contraction principle for establishing Rademacher complexity (Bartlett & Mendelson, 2003). It gets rid of the loss function and incurs a factor equal to the Lipschitz constant of the loss in the Rademacher complexity bound. Since the loss function takes scalar argument, the Lipschitz constant is defined for only one norm, i.e., the absolute value norm. It is not immediately clear how such an approach would work when the loss takes vector valued arguments and is Lipschitz w.r.t. $\|\cdot\|_\infty$. We are not aware of an appropriate extension of the Ledoux-Talagrand contraction principle. Note that Lipschitz continuity w.r.t. the Euclidean norm $\|\cdot\|_2$ does not pose a significant challenge since Slepian's lemma can be applied to get rid of the loss. Several authors have already exploited Slepian's lemma in this context (Bartlett & Mendelson, 2003; Chapelle & Wu, 2010). We take a route involving covering numbers and define the data-dependent (pseudo-)metric:

$$d_\infty^{Z^{(1:n)}}(w, w') := \max_{i=1}^n \left|\phi(X^{(i)} w, y^{(i)}) - \phi(X^{(i)} w', y^{(i)})\right|$$

Let $\mathcal{N}_\infty(\epsilon, \phi\circ\mathcal{F}, Z^{(1:n)})$ be the covering number at scale ϵ of the composite class $\phi\circ\mathcal{F} = \phi\circ\mathcal{F}_1$ or $\phi\circ\mathcal{F}_2$ w.r.t. the above metric. Also define

$$\mathcal{N}_\infty(\epsilon, \phi\circ\mathcal{F}, n) := \max_{Z^{(1:n)}} \mathcal{N}_\infty(\epsilon, \phi\circ\mathcal{F}, Z^{(1:n)}).$$

With these definitions in place, we can state our first result on covering numbers.

Proposition 7. *Let the loss ϕ be Lipschitz w.r.t. $\|\cdot\|_\infty$ with constant G_ϕ. Then following covering number bound holds:*

$$\log_2 \mathcal{N}_\infty(\epsilon, \phi\circ\mathcal{F}_2, n) \le \left\lceil \frac{G_\phi^2\, W_2^2\, R_X^2}{\epsilon^2} \right\rceil \log_2(2mn+1).$$

Proof. Note that

$$\begin{aligned}&\max_{i=1}^n \left|\phi(X^{(i)} w, y^{(i)}) - \phi(X^{(i)} w', y^{(i)})\right| \\ &\le G_\phi \cdot \max_{i=1}^n \max_{j=1}^m \left|\left\langle X_j^{(i)}, w\right\rangle - \left\langle X_j^{(i)}, w'\right\rangle\right|.\end{aligned}$$

This immediately implies that if we have a cover of the class $\mathcal{G}_2$ (Sec.2) at scale ϵ/G_ϕ w.r.t. the metric

$$\max_{i=1}^{n} \max_{j=1}^{m} \left| \left\langle X_j^{(i)}, w \right\rangle - \left\langle X_j^{(i)}, w' \right\rangle \right|$$

then it is also a cover of $\phi \circ \mathcal{F}_2$ w.r.t. $d_\infty^{Z^{(1:n)}}$, at scale ϵ. Now comes a simple, but crucial observation: *from the point of view of the scalar valued function class $\mathcal{G}_2$, the vectors $(X_j^{(i)})_{j=1:m}^{i=1:n}$ constitute a data set of size mn.* Therefore,

$$\mathcal{N}_\infty(\epsilon, \phi \circ \mathcal{F}_2, n) \le \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_2, mn). \tag{4}$$

Now we appeal to the following bound due to Zhang (2002, Corollary 3) (and plug the result into (4)):

$$\log_2 \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_2, mn) \le \left\lceil \frac{G_\phi^2\, W_2^2\, R_X^2}{\epsilon^2} \right\rceil \log_2(2mn+1)$$

□

Covering number $\mathcal{N}_2(\epsilon, \phi \circ \mathcal{F}, Z^{(1:n)})$ uses pseudo-metric:

$$d_2^{Z^{(1:n)}}(w, w') := \left(\sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)}w, y^{(i)}) - \phi(X^{(i)}w', y^{(i)}) \right)^2 \right)^{1/2}$$

It is well known that a control on $\mathcal{N}_2(\epsilon, \phi \circ \mathcal{F}, Z^{(1:n)})$ provides control on the empirical Rademacher complexity and that $\mathcal{N}_2$ covering numbers are smaller than $\mathcal{N}_\infty$ ones. For us, it will be convenient to use a more refined version[2] due to Mendelson (2002). Let $\mathcal{H}$ be a class of functions, with $\mathcal{H} : \mathcal{Z} \mapsto \mathbb{R}$, uniformly bounded by B. Then, we have following bound on empirical Rademacher complexity

$$\begin{aligned} &\widehat{\mathfrak{R}}_n(\mathcal{H}) \\ &\le \inf_{\alpha > 0} \left(4\alpha + 10 \int_\alpha^{\sup_{h \in \mathcal{H}} \sqrt{\widehat{\mathbb{E}}[h^2]}} \sqrt{\frac{\log_2 \mathcal{N}_2(\epsilon, \mathcal{H}, Z^{(1:n)})}{n}} d\epsilon \right) \qquad (5) \\ &\le \inf_{\alpha > 0} \left(4\alpha + 10 \int_\alpha^{B} \sqrt{\frac{\log_2 \mathcal{N}_2(\epsilon, \mathcal{H}, Z^{(1:n)})}{n}} d\epsilon \right). \qquad (6) \end{aligned}$$

Here $\widehat{\mathfrak{R}}_n(\mathcal{H})$ is the empirical Rademacher complexity of the class $\mathcal{H}$ defined as

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) := \mathbb{E}_{\sigma_{1:n}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(Z_i) \right],$$

where $\sigma_{1:n} = (\sigma_1, \ldots, \sigma_n)$ are iid Rademacher (symmetric Bernoulli) random variables.

[2]We use a further refinement due to Srebro and Sridharan available at http://ttic.uchicago.edu/~karthik/dudley.pdf

Corollary 8. *Let ϕ be Lipschitz w.r.t. $\|\cdot\|_\infty$ and uniformly bounded[3] by B for $w \in \mathcal{F}_2$. Then the empirical Rademacher complexities of the class $\phi \circ \mathcal{F}_2$ is bounded as*

$$\begin{aligned} \widehat{\mathfrak{R}}_n(\phi \circ \mathcal{F}_2) \le 10 G_\phi W_2 R_X \sqrt{\frac{\log_2(3mn)}{n}} \\ \times \log \frac{6B\sqrt{n}}{5G_\phi W_2 R_X \sqrt{\log_2(3mn)}}. \end{aligned}$$

Proof. This follows by simply plugging in estimates from Proposition 7 into (6) and choosing α optimally. □

Control on the Rademacher complexity immediately leads to uniform convergence bounds and generalization error bounds for ERM. The informal $\tilde{O}$ notation hides factors logarithmic in $m, n, B, G_\phi, R_X, W_1$. Note that all hidden factors are small and computable from the results above.

Theorem 9. *Suppose ϕ is Lipschitz w.r.t. $\|\cdot\|_\infty$ with constant G_ϕ and is uniformly bounded by B as w varies over $\mathcal{F}_2$. With probability at least $1-\delta$,*

$$\begin{aligned} &\forall w \in \mathcal{F}_2,\ L_\phi(w) \le \hat{L}_\phi(w) \\ &\qquad + \tilde{O}\left(G_\phi W_2 R_X \sqrt{\frac{1}{n}} + B\sqrt{\frac{\log(1/\delta)}{n}} \right) \end{aligned}$$

and therefore with probability at least $1-2\delta$,

$$L_\phi(\hat{w}) \le L_\phi(w^\star) + \tilde{O}\left(G_\phi W_2 R_X \sqrt{\frac{1}{n}} + B\sqrt{\frac{\log(1/\delta)}{n}} \right).$$

where $\hat{w}$ is an empirical risk minimizer over $\mathcal{F}_2$.

Proof. Follows from standard bounds using Rademacher complexity. See, for example, Bartlett & Mendelson (2003). □

As we said before, ignoring logarithmic factors, the bound for $\mathcal{F}_2$ is an improvement over the bound of Chapelle & Wu (2010).

8. Extensions

We extend the generalization bounds above to two settings: a) high dimensional features and b) smooth losses.

8.1. High-dimensional features

In learning to rank situations involving high dimensional features, it may not be appropriate to use the class $\mathcal{F}_2$ of ℓ_2 bounded predictors. Instead, we would like to consider the class $\mathcal{F}_1$ of ℓ_1 bounded predictors. In this case, it is

[3]A uniform bound on the loss easily follows under the (very reasonable) assumption that $\forall y, \exists s_y$ s.t. $\phi(s_y, y) = 0$. Then $\phi(Xw, y) \le G_\phi \|Xw - s_y\|_\infty \le G_\phi(W_2 R_X + \max_{y \in \mathcal{Y}} \|s_y\|_\infty) \le G_\phi(2W_2 R_X)$.

natural to measure size of the input matrix X in terms of a bound $\bar{R}_X$ on the maximum ℓ_∞ norm of each of its row. The following analogue of Proposition 7 can be shown.

Proposition 10. *Let the loss ϕ be Lipschitz w.r.t. $\|\cdot\|_\infty$ with constant G_ϕ. Then the following covering number bound holds:*

$$\log_2 \mathcal{N}_\infty(\epsilon, \phi \circ \mathcal{F}_1, n) \le \left\lceil \frac{288\, G_\phi^2\, W_1^2\, \bar{R}_X^2\, (2 + \log d)}{\epsilon^2} \right\rceil \times \log_2 \left(2 \left\lceil \frac{8 G_\phi W_1 \bar{R}_X}{\epsilon} \right\rceil mn + 1 \right).$$

Using the above result to control the Rademacher complexity of $\phi \circ \mathcal{F}_1$ gives the following bound.

Corollary 11. *Let ϕ be Lipschitz w.r.t. $\|\cdot\|_\infty$ and uniformly bounded by B for $w \in \mathcal{F}_1$. Then the empirical Rademacher complexities of the class $\phi \circ \mathcal{F}_1$ is bounded as*

$$\widehat{\mathfrak{R}}_n(\phi \circ \mathcal{F}_1) \le 120\sqrt{2} G_\phi W_1 \bar{R}_X \sqrt{\frac{\log(d)\, \log_2(24mnG_\phi W_1 \bar{R}_X)}{n}} \times \log^2 \frac{B + 24mnG_\phi W_1 \bar{R}_X}{40\sqrt{2} G_\phi W_1 \bar{R}_X \sqrt{\log(d)\, \log_2(24mnG_\phi W_1 \bar{R}_X)}}.$$

As in the previous section, control of Rademacher complexity immediately yields uniform convergence and ERM generalization error bounds.

Theorem 12. *Suppose ϕ is Lipschitz w.r.t. $\|\cdot\|_\infty$ with constant G_ϕ and is uniformly bounded by B as w varies over $\mathcal{F}_1$. With probability at least $1-\delta$,*

$$\forall w \in \mathcal{F}_1,\ L_\phi(w) \le \hat{L}_\phi(w) + \tilde{O}\left(G_\phi W_1 \bar{R}_X \sqrt{\frac{\log d}{n}} + B\sqrt{\frac{\log(1/\delta)}{n}} \right)$$

and therefore with probability at least $1-2\delta$,

$$L_\phi(\hat{w}) \le L_\phi(w^\star) + \tilde{O}\left(G_\phi W_1 \bar{R}_X \sqrt{\frac{\log d}{n}} + B\sqrt{\frac{\log(1/\delta)}{n}} \right)$$

where $\hat{w}$ is an empirical risk minimizer over $\mathcal{F}_1$.

As can be easily seen from Theorem. 12, the generalization bound is *almost* independent of the dimension of the document feature vectors. We are not aware of existence of such a result in learning to rank literature.

8.2. Smooth losses

We will again use online regret bounds to explain why we should expect "optimistic" rates for smooth losses before giving more general results for smooth but possibly nonconvex losses.

8.3. Online regret bounds under smoothness

Let us go back to OGD guarantee, this time presented in a slightly more refined version. If we run OGD with learning rate η then, for all $\|w\|_2 \le W_2$:

$$\sum_{i=1}^n f_i(w_i) - \sum_{i=1}^n f_i(w) \le \frac{W_2^2}{2\eta} + \eta \sum_{i=1}^n \|g_i\|_2^2$$

where $g_i = \nabla_{w_i} f_i(w_i)$ (if f_i is not differentiable at w_i then we can set g_i to be an arbitrary subgradient of f_i at w_i). Now assume that all f_i's are non-negative functions and are smooth w.r.t. $\|\cdot\|_2$ with constant H. Lemma 3.1 of Srebro et al. (2010) tells us that any non-negative, smooth function $f(w)$ enjoy an important *self-bounding* property for the gradient:

$$\|\nabla_w f_i(w)\|_2 \le \sqrt{4Hf_i(w)}$$

which bounds the magnitude of the gradient of f at a point in terms of the value of the function itself at that point. This means that $\|g_i\|_2^2 \le 4Hf_i(w_i)$ which, when plugged into the OGD guarantee, gives:

$$\sum_{i=1}^n f_i(w_i) - \sum_{i=1}^n f_i(w) \le \frac{W_2^2}{2\eta} + 4\eta H \sum_{i=1}^n f_i(w_i)$$

Again, setting $f_i(w) = \phi(X^{(i)}w, y^{(i)})$, $1 \le t \le n$, and using the online to batch conversion technique, we can arrive at the bound: for all $\|w\|_2 \le W_2$:

$$\mathbb{E}\left[L_\phi(\hat{w})\right] \le \frac{L_\phi(w)}{(1-4\eta H)} + \frac{W_2^2}{2\eta(1-4\eta H)n}$$

At this stage, we can fix $w = w^\star$, the optimal ℓ_2-norm bounded predictor and get optimal η as:

$$\eta = \frac{W_2}{4HW_2 + 2\sqrt{4H^2W_2^2 + 2HL_\phi(w^\star)n}}. \tag{7}$$

After plugging this value of η in the bound above and some algebra (see Section H), we get the upper bound

$$\mathbb{E}\left[L_\phi(\hat{w})\right] \le L_\phi(w^\star) + 2\sqrt{\frac{2HW_2^2 L_\phi(w^\star)}{n}} + \frac{8HW_2^2}{n}. \tag{8}$$

Such a rate interpolates between a $1/\sqrt{n}$ rate in the "pessimistic" case ($L_\phi(w^\star) > 0$) and the $1/n$ rate in the "optimistic" case ($L_\phi(w^\star) = 0$) (this terminology is due to Panchenko (2002)).

Now, assuming ϕ to be twice differentiable, we need H such that

$$H \ge \|\nabla_w^2 \phi(X^{(i)}w, y^{(i)})\|_{2\to 2} = \|X^\top \nabla_s^2 \phi(X^{(i)}w, y^{(i)}) X\|_{2\to 2}$$

where we used the chain rule to express ∇_w^2 in terms of ∇_s^2. Note that, for OGD, we need smoothness in w w.r.t.

$\|\cdot\|_2$ which is why the matrix norm above is the operator norm corresponding to the pair $\|\cdot\|_2, \|\cdot\|_2$. In fact, when we say "operator norm" without mentioning the pair of norms involved, it is this norm that is usually meant. It is well known that this norm is equal to the largest singular value of the matrix. But, just as before, we can bound this in terms of the smoothness constant of ϕ w.r.t. $\|\cdot\|_\infty$ (see Section I in the appendix):

$$\|(X^{(i)})^\top \nabla_s^2 \phi(X^{(i)}w, y^{(i)}) X^{(i)}\|_{2\to 2} \\ \leq R_X^2 \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1}.$$

where we used Lemma 4 once again. This result using online regret bounds is great for building intuition but suffers from the two defects we mentioned at the end of Section 5. In the smoothness case, it additionally suffers from a more serious defect: the correct choice of the learning rate η requires knowledge of $L_\phi(w^\star)$ which is seldom available.

8.4. Generalization error bounds under smoothness

Once again, to prove a general result for possibly non-convex smooth losses, we will adopt an approach based on covering numbers. To begin, we will need a useful lemma from Srebro et al. (2010, Lemma A.1 in the Supplementary Material). Note that, for functions over real valued predictions, we do not need to talk about the norm when dealing with smoothness since essentially the only norm available is the absolute value.

Lemma 13. *For any h-smooth non-negative function $f : \mathbb{R} \to \mathbb{R}_+$ and any $t, r \in \mathbb{R}$ we have*

$$(f(t) - f(r))^2 \leq 6h(f(t) + f(r))(t - r)^2.$$

We first provide an extension of this lemma to the vector case.

Lemma 14. *If $\phi : \mathbb{R}^m \to \mathbb{R}_+$ is a non-negative function with smoothness constant H_ϕ w.r.t. a norm $|||\cdot|||$ then for any $s_1, s_2 \in \mathbb{R}^m$ we have*

$$(\phi(s_1) - \phi(s_2))^2 \leq 6H_\phi \cdot (\phi(s_1) + \phi(s_2)) \cdot |||s_1 - s_2|||^2.$$

Using the basic idea behind local Rademacher complexity analysis, we define the following loss class:

$$\mathcal{F}_{\phi,2}(r) := \{(X, y) \mapsto \phi(Xw, y) : \|w\|_2 \leq W_2, \hat{L}_\phi(w) \leq r\}.$$

Note that this is a random subclass of functions since $\hat{L}_\phi(w)$ is a random variable.

Proposition 15. *Let ϕ be smooth w.r.t. $\|\cdot\|_\infty$ with constant H_ϕ. The covering numbers of $\mathcal{F}_{\phi,2}(r)$ in the $d_2^{Z^{(1:n)}}$ metric defined above are bounded as follows:*

$$\log_2 \mathcal{N}_2(\epsilon, \mathcal{F}_{\phi,2}(r), Z^{(1:n)}) \leq \left\lceil \frac{12 H_\phi\, W_2^2\, R_X^2\, r}{\epsilon^2} \right\rceil \log_2(2mn+1).$$

Control of covering numbers easily gives a control on the Rademacher complexity of the random subclass $\mathcal{F}_{\phi,2}(r)$.

Corollary 16. *Let ϕ be smooth w.r.t. $\|\cdot\|_\infty$ with constant H_ϕ and uniformly bounded by B for $w \in \mathcal{F}_2$. Then the empirical Rademacher complexity of the class $\mathcal{F}_{\phi,2}(r)$ is bounded as*

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_{\phi,2}(r)) \leq 4\sqrt{r} C \log \frac{3\sqrt{B}}{C}$$

where $C = 5\sqrt{3} W_2 R_X \sqrt{\frac{H_\phi \log_2(3mn)}{n}}$.

With the above corollary in place we can now prove our second key result.

Theorem 17. *Suppose ϕ is smooth w.r.t. $\|\cdot\|_\infty$ with constant H_ϕ and is uniformly bounded by B over $\mathcal{F}_2$. With probability at least $1 - \delta$,*

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \leq \hat{L}_\phi(w) + \tilde{O}\left(\sqrt{\frac{L_\phi(w) D_0}{n}} + \frac{D_0}{n}\right)$$

where $D_0 = B \log(1/\delta) + W_2^2 R_X^2 H_\phi$. Moreover, with probability at least $1 - 2\delta$,

$$L_\phi(\hat{w}) \leq L_\phi(w^\star) + \tilde{O}\left(\sqrt{\frac{L_\phi(w^\star) D_0}{n}} + \frac{D_0}{n}\right)$$

where $\hat{w}, w^\star$ are minimizers of $\hat{L}_\phi(w)$ and $L_\phi(w)$ respectively (over $w \in \mathcal{F}_2$).

9. Conclusion

We showed that it is not necessary for generalization error bounds for query-level learning to rank algorithms to deteriorate with increasing length of document lists associated with queries. The key idea behind our improved bounds was defining Lipschitz constants w.r.t. ℓ_∞ norm instead of the "standard" ℓ_2 norm. As a result, we were able to derive much tighter guarantees for popular loss functions such as ListNet and Smoothed DCG@1 than previously available.

Our generalization analysis of learning to rank algorithms paves the way for further interesting work. One possibility is to use these bounds to design active learning algorithms for learning to rank with formal label complexity guarantees. Another interesting possibility is to consider other problems, such as multi-label learning, where functions with vector-valued outputs are learned by optimizing a joint function of those outputs.

Acknowledgement

We gratefully acknowledge the support of NSF under grant IIS-1319810. Thanks to Prateek Jain for discussions that led us to Theorem 3.

References

Bartlett, Peter L. and Mendelson, Shahar. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3: 463–482, 2003.

Bousquet, Olivier. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Ecole Polytechnique, 2002.

Cao, Zhe, Qin, Tao, Liu, Tie-Yan, Tsai, Ming-Feng, and Li, Hang. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007.

Chapelle, Olivier and Wu, Mingrui. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

Chapelle, Olivier, Metlzer, Donald, Zhang, Ya, and Grinspan, Pierre. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 621–630. ACM, 2009.

Chapelle, Olivier, Chang, Yi, and Liu, Tie-Yan. Future directions in learning to rank. In *Proceedings of the Yahoo! Learning to Rank Challenge June 25, 2010, Haifa, Israel*, Journal of Machine Learning Research Workshop and Conference Proceedings, pp. 91–100, 2011.

Järvelin, Kalervo and Kekäläinen, Jaana. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Joachims, Thorsten. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142. ACM, 2002.

Lan, Yanyan, Liu, Tie-Yan, Qin, Tao, Ma, Zhiming, and Li, Hang. Query-level stability and generalization in learning to rank. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 512–519. ACM, 2008.

Lan, Yanyan, Liu, Tie-Yan, Ma, Zhiming, and Li, Hang. Generalization analysis of listwise learning-to-rank algorithms. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 577–584, 2009.

Liu, Tie-yan, Xu, Jun, Qin, Tao, Xiong, Wenying, and Li, Hang. LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pp. 3–10, 2007.

Mendelson, Shahar. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

Panchenko, Dmitriy. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

Shalev-Shwartz, Shai, Shamir, Ohad, Srebro, Nathan, and Sridharan, Karthik. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Srebro, Nathan, Sridharan, Karthik, and Tewari, Ambuj. Smoothness, low noise, and fast rates. In *Advances in Neural Information Processing Systems 23*, pp. 2199–2207, 2010.

Yue, Yisong, Finley, Thomas, Radlinski, Filip, and Joachims, Thorsten. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 271–278, 2007.

Zhang, Tong. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

A. Proof of Proposition 1

Proof. Let e_j's denote standard basis vectors. We have

$$\nabla_s \phi_{\mathrm{LN}}(s,y) = -\sum_{j=1}^{m} P_j(y) e_j + \sum_{j=1}^{m} \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} e_j$$

Therefore,

$$\begin{aligned}\|\nabla_s \phi_{\mathrm{LN}}(s,y)\|_1 &\le \sum_{j=1}^{m} P_j(y)\|e_j\|_1 + \sum_{j=1}^{m} \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})}\|e_j\|_1 \\ &= 2.\end{aligned}$$

We also have

$$[\nabla_s^2 \phi_{\mathrm{LN}}(s,y)]_{j,k} = \begin{cases} -\frac{\exp(2s_j)}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} + \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} & \text{if } j = k \\ -\frac{\exp(s_j+s_k)}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} & \text{if } j \neq k\ . \end{cases}$$

Moreover,

$$\begin{aligned}\|\nabla_s^2 \phi_{\mathrm{LN}}(s,y)\|_{\infty\to 1} &\le \sum_{j=1}^{m}\sum_{k=1}^{m} |[\nabla_s^2 \phi_{\mathrm{LN}}(s,y)]_{j,k}| \\ &\le \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{\exp(s_j+s_k)}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} + \sum_{j=1}^{m}\frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} \\ &= \frac{(\sum_{j=1}^{m}\exp(s_j))^2}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} + \frac{\sum_{j=1}^{m}\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} \\ &= 2\end{aligned}$$

□

B. Proof of Proposition 2

Proof. Let $1_{(\text{condition})}$ denote an indicator variable. We have

$$[\nabla_s \phi_{\mathrm{SD}}(s,y)]_j = D(1)\left(\sum_{i=1}^{m} G(r_i)\left[\frac{1}{\sigma}\frac{\exp(s_i/\sigma)}{\sum_{j'}\exp(s_{j'}/\sigma)} 1_{(i=j)} - \frac{1}{\sigma}\frac{\exp((s_i+s_j)/\sigma)}{(\sum_{j'}\exp(s_{j'}/\sigma))^2}\right]\right)$$

Therefore,

$$\begin{aligned}\frac{\|\nabla_s \phi_{\mathrm{SD}}(s,y)\|_1}{D(1)G(Y_{\max})} &\le \sum_{j=1}^{m}\left(\sum_{i=1}^{m}\left[\frac{1}{\sigma}\frac{\exp(s_i/\sigma)}{\sum_{j'}\exp(s_{j'}/\sigma)} 1_{(i=j)} + \frac{1}{\sigma}\frac{\exp((s_i+s_j)/\sigma)}{(\sum_{j'}\exp(s_{j'}/\sigma))^2}\right]\right) \\ &= \frac{1}{\sigma}\left(\frac{\sum_j \exp(s_j/\sigma)}{\sum_{j'}\exp(s_{j'}/\sigma)} + \frac{(\sum_j \exp(s_j/\sigma))^2}{(\sum_{j'}\exp(s_{j'}/\sigma))^2}\right) \\ &= \frac{2}{\sigma}.\end{aligned}$$

□

C. RankSVM

The RankSVM surrogate is defined as:

$$\phi_{RS}(s,y) = \sum_{i=1}^{m}\sum_{j=1}^{m} \max(0, 1_{(y_i > y_j)}(1 + s_j - s_i))$$

It is easy to see that $\nabla_s \phi_{RS}(s,y) = \sum_{i=1}^m \sum_{j=1}^m \max(0, 1_{(y_i > y_j)}(1 + s_j - s_i))(e_j - e_i)$. Thus, the ℓ_1 norm of gradient is $O(m^2)$.

D. Proof of Theorem 3

Proof. It is straightforward to check that $\mathcal{F}'_{\text{lin}}$ is contained in both $\mathcal{F}_{\text{full}}$ as well as $\mathcal{F}_{\text{perminv}}$. So, we just need to prove that any f that is in both $\mathcal{F}_{\text{full}}$ and $\mathcal{F}_{\text{perminv}}$ has to be in $\mathcal{F}'_{\text{lin}}$ as well.

Let P_π denote the $m \times m$ permutation matrix corresponding to a permutation π. Consider the full linear class $\mathcal{F}_{\text{full}}$. In matrix notation, the permutation invariance property means that, for any π, X, we have $P_\pi[\langle X, W_1 \rangle, \ldots, \langle X, W_m \rangle)]^\top = [\langle P_\pi X, W_1 \rangle, \ldots, \langle P_\pi X, W_m \rangle]^\top$.

Let $\rho_1 = \{P_\pi : \pi(1) = 1\}$, where $\pi(i)$ denotes the index of the element in the ith position according to permutation π. Fix any $P \in \rho_1$. Then, for any X, $\langle X, W_1 \rangle = \langle PX, W_1 \rangle$. This implies that, for all X, $\text{Tr}({W_1}^\top X) = \text{Tr}({W_1}^\top PX)$. Using the fact that $\text{Tr}(A^\top X) = \text{Tr}(B^\top X), \forall X$ implies $A = B$, we have that ${W_1}^\top = {W_1}^\top P$. Because $P^\top = P^{-1}$, this means $PW_1 = W_1$. This shows that all rows of W_1, other than 1st row, are the same but perhaps different from 1st row. By considering $\rho_i = \{P_\pi : \pi(i) = i\}$ for $i > 1$, the same reasoning shows that, for each i, all rows of W_i, other than ith row, are the same but possibly different from ith row.

Let $\rho_{1\leftrightarrow 2} = \{P_\pi : \pi(1) = 2, \pi(2) = 1\}$. Fix any $P \in \rho_{1\leftrightarrow 2}$. Then, for any X, $\langle X, W_2 \rangle = \langle PX, W_1 \rangle$ and $\langle X, W_1 \rangle = \langle PX, W_2 \rangle$. Thus, we have $W_2^\top = W_1^\top P$ as well as $W_1^\top = W_2^\top P$ which means $PW_2 = W_1, PW_1 = W_2$. This shows that row 1 of W_1 and row 2 of W_2 are the same. Moreover, row 2 of W_1 and row 1 of W_2 are the same. Thus, for some $u, u' \in \mathbb{R}^d$, W_1 is of the form $[u|u'|u'|\ldots|u']^\top$ and W_2 is of the form $[u'|u|u'|\ldots|u']^\top$. Repeating this argument by considering $\rho_{1\leftrightarrow i}$ for $i > 2$ shows that W_i is of the same form (u in row i and u' elsewhere).

Therefore, we have proved that any linear map that is permutation invariant has to be of the form:

$$X \mapsto \left(u^\top X_i + (u')^\top \sum_{j \neq i} X_j \right)_{i=1}^m .$$

We can reparameterize above using $w = u - u'$ and $v = u'$ which proves the result. $\square$

E. Proof of Lemma 4

Proof. The first equality is true because

$$\begin{aligned} \|X^\top\|_{1\to p} &= \sup_{v\neq 0} \frac{\|X^\top v\|_p}{\|v\|_1} = \sup_{v\neq 0}\sup_{u\neq 0} \frac{\langle X^\top v, u\rangle}{\|v\|_1\|u\|_q} \\ &= \sup_{u\neq 0}\sup_{v\neq 0} \frac{\langle v, Xu\rangle}{\|v\|_1\|u\|_q} = \sup_{u\neq 0} \frac{\|Xu\|_\infty}{\|u\|_q} = \|X\|_{q\to\infty}. \end{aligned}$$

The second is true because

$$\begin{aligned} \|X\|_{q\to\infty} &= \sup_{u\neq 0} \frac{\|Xu\|_\infty}{\|u\|_q} = \sup_{u\neq 0} \max_{j=1}^m \frac{|\langle X_j, u\rangle|}{\|u\|_q} \\ &= \max_{j=1}^m \sup_{u\neq 0} \frac{|\langle X_j, u\rangle|}{\|u\|_q} = \max_{j=1}^m \|X_j\|_p. \end{aligned}$$

$\square$

F. Proof of Theorem 6

Our theorem is developed from the "expectation version" of Theorem 6 of Shalev-Shwartz et al. (2009) that was originally given in probabilistic form. The expected version is as follows.

Let $\mathcal{Z}$ be a space endowed with a probability distribution generating iid draws $Z_1, \ldots, Z_n$. Let $\mathcal{W} \subseteq \mathbb{R}^d$ and $f : \mathcal{W} \times \mathcal{Z} \to$

$\mathbb{R}$ be λ-strongly convex[4] and G-Lipschitz (w.r.t. $\|\cdot\|_2$) in w for every z. We define $F(w) = \mathbb{E}\left[f(w,Z)\right]$ and let

$$w^\star = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\ F(w),$$

$$\hat{w} = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} f(w,Z_i).$$

Then $\mathbb{E}\left[F(\hat{w}) - F(w^\star)\right] \le \frac{4G^2}{\lambda n}$, where the expectation is taken over the sample. The above inequality can be proved by carefully going through the proof of Theorem 6 proved by Shalev-Shwartz et al. (2009).

We now derive the "expectation version" of Theorem 7 of Shalev-Shwartz et al. (2009). Define the regularized empirical risk minimizer as follows:

$$\hat{w}_\lambda = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\ \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} f(w,Z_i). \tag{9}$$

The following result gives optimality guarantees for the regularized empirical risk minimizer.

Theorem 18. *Let $\mathcal{W} = \{w : \|w\|_2 \le W_2\}$ and let $f(w,z)$ be convex and G-Lipschitz (w.r.t. $\|\cdot\|_2$) in w for every z. Let $Z_1,\ldots,Z_n$ be iid samples and let $\lambda = \sqrt{\frac{\frac{4G^2}{n}}{\frac{W_2^2}{2}+\frac{4W_2^2}{n}}}$. Then for $\hat{w}_\lambda$ and $w^\star$ as defined above, we have*

$$\mathbb{E}\left[F(\hat{w}_\lambda) - F(w^\star)\right] \le 2\,G\,W_2\left(\frac{8}{n} + \sqrt{\frac{2}{n}}\right). \tag{10}$$

Proof. Let $r_\lambda(w,z) = \frac{\lambda}{2}\|w\|_2^2 + f(w,z)$. Then r_λ is λ-strongly convex with Lipschitz constant $\lambda W_2 + G$ in $\|\cdot\|_2$. Applying "expectation version" of Theorem 6 of Shalev-Shwartz et al. (2009) to r_λ, we get

$$\mathbb{E}\left[\frac{\lambda}{2}\|\hat{w}_\lambda\|_2^2 + F(\hat{w}_\lambda)\right] \le \min_{w\in\mathcal{W}}\left\{\frac{\lambda}{2}\|w\|_2^2 + F(w)\right\} + \frac{4(\lambda W_2 + G)^2}{\lambda n} \le \frac{\lambda}{2}\|w^\star\|_2^2 + F(w^*) + \frac{4(\lambda W_2+G)^2}{\lambda n}.$$

Thus, we get

$$\mathbb{E}\left[F(\hat{w}_\lambda) - F(w^\star)\right] \le \frac{\lambda W_2^2}{2} + \frac{4(\lambda W_2 + G)^2}{\lambda n}.$$

Minimizing the upper bound w.r.t. λ, we get $\lambda = \sqrt{\frac{4G^2}{n}}\sqrt{\frac{1}{\frac{W_2^2}{2}+\frac{4W_2^2}{n}}}$. Plugging this choice back in the equation above and using the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ finishes the proof of Theorem 18. □

We now have all ingredients to prove Theorem 6.

Proof of Theorem 6. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $f(w,z) = \phi(Xw,y)$ and apply Theorem 18. Finally note that if ϕ is G_ϕ-Lipschitz w.r.t. $\|\cdot\|_\infty$ and every row of $X \in \mathbb{R}^{m\times d}$ has Euclidean norm bounded by R_X then $f(\cdot,z)$ is $G_\phi R_X$-Lipschitz w.r.t. $\|\cdot\|_2$ in w. □

G. Proof of Theorem 12

Proof. Following exactly the same line of reasoning (reducing a sample of size n, where each prediction is $\mathbb{R}^m$-valued, to an sample of size mn, where each prediction is real valued) as in the beginning of proof of Proposition 7, we have

$$\mathcal{N}_\infty(\epsilon, \phi\circ\mathcal{F}_1, n) \le \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_1, mn). \tag{11}$$

Plugging in the following bound due to Zhang (2002, Corollary 5):

$$\log_2 \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_1, mn) \le \left\lceil \frac{288\,G_\phi^2\,W_1^2\,\bar{R}_X^2\,(2+\ln d)}{\epsilon^2} \right\rceil \times \log_2\left(2\lceil 8G_\phi W_1 \bar{R}_X/\epsilon\rceil mn + 1\right)$$

into (11) respectively proves the result. □

[4] Recall that a function is called λ-strongly convex (w.r.t. $\|\cdot\|_2$) iff $f - \frac{\lambda}{2}\|\cdot\|_2^2$ is convex.

H. Calculations involved in deriving Equation (8)

Plugging in the value of η from (7) into the expression

$$\frac{L_\phi(w^\star)}{(1-4\eta H)} + \frac{W_2^2}{2\eta(1-4\eta H)n}$$

yields (using the shorthand $L^\star$ for $L_\phi(w^\star)$)

$$L^\star + \frac{2HW_2L^\star}{\sqrt{4H^2W_2^2+2HL^\star n}} + \frac{W_2}{n}\left[\frac{4H^2W_2^2}{\sqrt{4H^2W_2^2+2HL^\star n}} + \sqrt{4H^2W_2^2+2HL^\star n} + 4HW_2\right]$$

Denoting HW_2^2/n by x, this simplifies to

$$L^\star + \frac{2\sqrt{x}L^\star + 4x\sqrt{x}}{\sqrt{4x+2L^\star}} + \sqrt{x}\sqrt{4x+2L^\star} + 4x.$$

Using the arithmetic mean-geometric mean inequality to upper bound the middle two terms gives

$$L^\star + 2\sqrt{2xL^\star + 4x^2} + 4x.$$

Finally, using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we get our final upper bound

$$L^\star + 2\sqrt{2xL^\star} + 8x.$$

I. Calculation of smoothness constant

$$\begin{aligned}
&\|(X^{(i)})^\top \nabla_s^2 \phi(X^{(i)}w, y^{(i)}) X^{(i)}\|_{2\to 2} = \sup_{v\neq 0} \frac{\|(X^{(i)})^\top \nabla_s^2 \phi(X^{(i)}w, y^{(i)}) X^{(i)} v\|_2}{\|v\|_2} \\
&\le \sup_{v\neq 0} \frac{\|(X^{(i)})^\top\|_{1\to 2} \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)}) X^{(i)} v\|_1}{\|v\|_2} \le \sup_{v\neq 0} \frac{\|(X^{(i)})^\top\|_{1\to 2} \cdot \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1} \cdot \|X^{(i)} v\|_\infty}{\|v\|_2} \\
&\le \sup_{v\neq 0} \frac{\|(X^{(i)})^\top\|_{1\to 2} \cdot \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1} \cdot \|X^{(i)}\|_{2\to\infty} \cdot \|v\|_2}{\|v\|_2} \\
&\le \left(\max_{j=1}^{m} \|X_j^{(i)}\|\right)^2 \cdot \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1} \\
&\le R_X^2 \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1}.
\end{aligned}$$

J. Proof of Lemma 14

Proof. Consider the function

$$f(t) = \phi((1-t)s_1 + ts_2).$$

It is clearly non-negative. Moreover

$$\begin{aligned}
|f'(t_1) - f'(t_2)| &= |\langle \nabla_s \phi(s_1 + t_1(s_2 - s_1)) - \nabla_s \phi(s_1 + t_2(s_2 - s_1)), s_2 - s_1 \rangle| \\
&\le |||\nabla_s \phi(s_1 + t_1(s_2 - s_1)) - \nabla_s \phi(s_1 + t_2(s_2 - s_1))|||_\star \cdot |||s_2 - s_1||| \\
&\le H_\phi \, |t_1 - t_2| \, |||s_2 - s_1|||^2
\end{aligned}$$

and therefore it is smooth with constant $h = H_\phi |||s_2 - s_1|||^2$. Appealing to Lemma 13 now gives

$$(f(1) - f(0))^2 \le 6H_\phi |||s_2 - s_1|||^2 (f(1) + f(0))(1-0)^2$$

which proves the lemma since $f(0) = \phi(s_1)$ and $f(1) = \phi(s_2)$. □

K. Proof of Proposition 15

Proof. Let $w, w' \in \mathcal{F}_{\phi,2}(r)$. Using Lemma 14

$$
\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) - \phi(X^{(i)} w', y^{(i)}) \right)^2 \\
&\le 6 H_\phi \sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) + \phi(X^{(i)} w', y^{(i)}) \right) \\
&\quad \cdot \| X^{(i)} w - X^{(i)} w' \|_\infty^2 \\
&\le 6 H_\phi \cdot \max_{i=1}^{n} \| X^{(i)} w - X^{(i)} w' \|_\infty^2 \\
&\quad \cdot \sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) + \phi(X^{(i)} w', y^{(i)}) \right) \\
&= 6 H_\phi \cdot \max_{i=1}^{n} \| X^{(i)} w - X^{(i)} w' \|_\infty^2 \cdot \left(\hat{L}_\phi(w) + \hat{L}_\phi(w') \right) \\
&\le 12 H_\phi r \cdot \max_{i=1}^{n} \| X^{(i)} w - X^{(i)} w' \|_\infty^2 .
\end{aligned}
$$

where the last inequality follows because $\hat{L}_\phi(w) + \hat{L}_\phi(w') \le 2r$.
This immediately implies that if we have a cover of the class $\mathcal{G}_2$ at scale $\epsilon / \sqrt{12 H_\phi r}$ w.r.t. the metric

$$
\max_{i=1}^{n} \max_{j=1}^{m} \left| \left\langle X_j^{(i)}, w \right\rangle - \left\langle X_j^{(i)}, w' \right\rangle \right|
$$

then it is also a cover of $\mathcal{F}_{\phi,2}(r)$ w.r.t. $d_2^{Z^{(1:n)}}$. Therefore, we have

$$
\mathcal{N}_2(\epsilon, \mathcal{F}_{\phi,2}(r), Z^{(1:n)}) \le \mathcal{N}_\infty(\epsilon / \sqrt{12 H_\phi r}, \mathcal{G}_2, mn). \tag{12}
$$

Appealing once again to a result by Zhang (2002, Corollary 3), we get

$$
\begin{aligned}
\log_2 \mathcal{N}_\infty(\epsilon / \sqrt{12 H_\phi r}, \mathcal{G}_2, mn) \le &\left\lceil \frac{12 H_\phi \, W_2^2 \, R_X^2 \, r}{\epsilon^2} \right\rceil \\
&\times \log_2(2mn + 1)
\end{aligned}
$$

which finishes the proof. □

L. Proof of Corollary 16

Proof. We plug in Proposition 15's estimate into (5):

$$
\begin{aligned}
\widehat{\mathfrak{R}}_n \left(\mathcal{F}_{\phi,2}(r) \right) &\le \inf_{\alpha > 0} \left(4\alpha + 10 \int_\alpha^{\sqrt{Br}} \sqrt{ \frac{ \left\lceil \frac{12 H_\phi \, W_2^2 \, R_X^2 \, r}{\epsilon^2} \right\rceil \log_2(2mn+1) }{n} } d\epsilon \right) \\
&\le \inf_{\alpha > 0} \left(4\alpha + 20\sqrt{3} W_2 R_X \sqrt{ \frac{ r H_\phi \log_2(3mn) }{n} } \int_\alpha^{\sqrt{Br}} \frac{1}{\epsilon} d\epsilon \right) .
\end{aligned}
$$

Now choosing $\alpha = C\sqrt{r}$ where $C = 5\sqrt{3} W_2 R_X \sqrt{\frac{H_\phi \log_2(3mn)}{n}}$ gives us the upper bound

$$
\widehat{\mathfrak{R}}_n \left(\mathcal{F}_{\phi,2}(r) \right) \le 4\sqrt{r} C \left(1 + \log \frac{\sqrt{B}}{C} \right) \le 4\sqrt{r} C \log \frac{3\sqrt{B}}{C}.
$$

□

M. Proof of Theorem 17

Proof. We appeal to Theorem 6.1 of Bousquet (2002) that assumes there exists an upper bound

$$\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{2,\phi}(r)\right) \leq \psi_n(r)$$

where $\psi_n : [0, \infty) \to \mathbb{R}_+$ is a non-negative, non-decreasing, non-zero function such that $\psi_n(r)/\sqrt{r}$ is non-increasing. The upper bound in Corollary 16 above satisfies these conditions and therefore we set $\psi_n(r) = 4\sqrt{r}C \log \frac{3\sqrt{B}}{C}$ with C as defined in Corollary 16. From Bousquet's result, we know that, with probability at least $1 - \delta$,

$$\begin{aligned}\forall w \in \mathcal{F}_2,\ L_\phi(w) \leq \hat{L}_\phi(w) + 45r_n^\star + \sqrt{8r_n^\star L_\phi(w)} \\ + \sqrt{4r_0 L_\phi(w)} + 20r_0\end{aligned}$$

where $r_0 = B(\log(1/\delta) + \log\log n)/n$ and $r_n^\star$ is the largest solution to the equation $r = \psi_n(r)$. In our case, $r_n^\star = \left(4C \log \frac{3\sqrt{B}}{C}\right)^2$. This proves the first inequality.

Now, using the above inequality with $w = \hat{w}$, the empirical risk minimizer and noting that $\hat{L}_\phi(\hat{w}) \leq \hat{L}_\phi(w^\star)$, we get

$$\begin{aligned}L_\phi(\hat{w}) \leq \hat{L}_\phi(w^\star) + 45r_n^\star + \sqrt{8r_n^\star L_\phi(\hat{w})} \\ + \sqrt{4r_0 L_\phi(\hat{w})} + 20r_0\end{aligned}$$

The second inequality now follows after some elementary calculations detailed below. □

M.1. Details of some calculations in the proof of Theorem 17

Using Bernstein's inequality, we have, with probability at least $1 - \delta$,

$$\begin{aligned}\hat{L}_\phi(w^\star) &\leq L_\phi(w^\star) + \sqrt{\frac{4\mathrm{Var}[\phi(Xw^\star, y)]\log(1/\delta)}{n}} + \frac{4B\log(1/\delta)}{n} \\ &\leq L_\phi(w^\star) + \sqrt{\frac{4BL_\phi(w^\star)\log(1/\delta)}{n}} + \frac{4B\log(1/\delta)}{n} \\ &\leq L_\phi(w^\star) + \sqrt{4r_0 L_\phi(w^\star)} + 4r_0.\end{aligned}$$

Set $D_0 = 45r_n^\star + 20r_0$. Putting the two bounds together and using some simple upper bounds, we have, with probability at least $1 - 2\delta$,

$$\begin{aligned}L_\phi(\hat{w}) &\leq \sqrt{D_0 \hat{L}_\phi(w^\star)} + D_0, \\ \hat{L}_\phi(w^\star) &\leq \sqrt{D_0 L_\phi(w^\star)} + D_0.\end{aligned}$$

which implies that

$$L_\phi(\hat{w}) \leq \sqrt{D_0}\sqrt{\sqrt{D_0 L_\phi(w^\star)} + D_0} + D_0.$$

Using $\sqrt{ab} \leq (a + b)/2$ to simplify the first term on the right gives us

$$L_\phi(\hat{w}) \leq \frac{D_0}{2} + \frac{\sqrt{D_0 L_\phi(w^\star)} + D_0}{2} + D_0 = \frac{\sqrt{D_0 L_\phi(w^\star)}}{2} + 2D_0\ .$$