Generalization Bounds for Convex Surrogates in Learning to Rank

Sougata Chaudhuri
sougata@umich.edu

Ambuj Tewari
tewaria@umich.edu

Abstract

Target metrics like Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP), used in Learning to Rank, operate on a pair of (permutation, ground truth relevance vector) pertaining to a list of objects. The metrics are discontinuous in the score vector which induces the permutation. Thus, like in classification, the prevalent technique for learning a ranking function during training phase is via optimization of surrogate functions. Generalization ability of functions learnt via minimizing such surrogates is a natural question. To the best of our knowledge, [1] has provided the best known generalization bound for general Lipschitz surrogates with linear ranking functions. However, the complexity term in the established bounds has a $\sqrt{m}$ factor, where m is the number of objects per query. We first provide an intuition as to why the complexity term should be independent of m. We then provide a *sufficient* condition for convex surrogates with linear ranking function to have generalization bound independent of m. We then frame open questions for improved generalization bounds for Lipschitz and smooth surrogates.

1 Introduction

Learning to rank is a supervised learning problem where the output space is the space of *rankings* of a set of objects. The accuracy of a ranked list, in comparison to the actual relevance of the documents, is measured by various ranking performance measures, such as NDCG [2], MAP [3] and others.

All major performance measures are discontinuous in the scores induced by the ranking function. Due to the computational difficulty of optimizing them during the training phase, several existing ranking methods have been developed based on minimizing *surrogate* losses, which are easy to optimize. One of the open questions in learning to rank is the generalization ability of linear ranking functions, learnt via optimizing such surrogates [4].

Since the surrogates operate on vectors, the Ledoux-Talagrand contraction principle cannot be used to calculate the Rademacher complexity of the composite class of surrogates and linear ranking functions. To the best of our knowledge, [1] has provided the best known bound for general Lipschitz in ℓ_2 norm ranking surrogates. The generalization bound has an inherent dependence on m, which is the number of objects (documents) per query. We provide intuition and then give a sufficient condition for convex ranking surrogates to have generalization bound independent of m.

2 Problem Definition

In learning to rank, an instance consists of a query q, along with a list of m documents, and the corresponding label is a relevance vector of length m. Formally, the input space is $\mathcal{X} \subseteq \mathbb{R}^{m \times d}$, consisting of lists of m documents represented as d dimensional feature vectors and supervision space is $\mathcal{Y} \subseteq \mathbb{R}^m$, representing relevance label vectors.

The objective is to learn a ranking function which ranks the documents associated with a query. A common technique in the literature is to learn a scoring function and obtain a ranking by sorting the score vector. For $X \in \mathcal{X}$, a linear scoring function is $f_w(X) = Xw = s^w \in \mathbb{R}^m$, where $w \in \mathbb{R}^d$. The quality of the learnt ranking function is evaluated on an independent test query by comparing the ranks of the documents, associated with the query, according to the scores, and their ranks according to actual relevance labels, using various performance measures.

3 Generalization Error Bound

We start with Theorem.1 of [1] .

Theorem 1. *Let $\phi : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}$ be a loss function such that $\forall\ R \in \mathcal{Y}$, $\phi(\cdot, R)$ is Lipschitz in ℓ_2 norm with constant L: $\forall\ R \in \mathcal{Y}, \forall\ f, f' \in \mathbb{R}^m, |\phi(f, R) - \phi(f', R)| \leq L\|f - f'\|_2$. Then, for any training set of n queries $((x^1, R^1), \ldots, (x^n, R^n))$, drawn independently according to probability distribution on $\mathcal{X} \times \mathcal{Y}$, with probability $\geq\ 1 - \delta$, the following inequality holds $\forall\ w \in \mathbb{R}^d$, such that $\|w\|_2 \leq B$*

$$E\phi(w^\top x_1, \ldots, w^\top x_m, R) \leq \frac{1}{n}\sum_{i=1}^{n} \phi(w^\top x_1^i, \ldots, w^\top x_m^i, R^i) + 3LBR\sqrt{\frac{m}{n}} + \sqrt{\frac{8log(2/\delta)}{n}}$$

where the expectation on the left hand size is taken over the underlying distribution on $\mathcal{X} \times \mathcal{Y}$

Discussion: The generalization error bound of Theorem.1 is established by using Slepian lemma en route to calculation of gaussian complexity of ϕ. The use of Slepian lemma necessiates the Lipschitz in ℓ_2 norm property of ϕ and introduces the dependence on m.

Intuition behind m independent generalization bound: As stated in Sec.2 , ranking is obtained by sorting a score vector obtained via a linear scoring function f_w. The space of linear scoring function consists of linear maps $f : \mathbb{R}^{m \times d} \to \mathbb{R}^m$. The linear function space can be fully parameterized by matrices $(W_1, \ldots, W_m)$, where $W_i \in \mathbb{R}^{m \times d}$. The representation will be of the form $f(X) = [\langle X, W_1\rangle, \ldots, \langle X, W_m\rangle]^\top \in \mathbb{R}^m$, where $\langle X, W\rangle := \mathrm{Tr}(X^\top W) = \mathrm{Tr}(W^\top X)$. Thus, a full parameterization of the linear scoring function is of dimension $m^2 \times d$.

The popularly used form of linear scoring function, viz. $f(X) = Xw \in \mathbb{R}^m$, with $w \in \mathbb{R}^d$ is actually a low d-dimensional subspace of the full $m^2 d$ dimensional space of linear maps. *Most importantly, the dimension 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 parameterization of the linear ranking function is of dimension independent of m, intuition would suggest that, under some conditions, ranking surrogates with linear ranking function should have an m independent complexity term in the generalization bound.

Before we establish our generalization error bounds, we analyze why the linear scoring function $f(X) = Xw \in \mathbb{R}^m$, with $w \in \mathbb{R}^d$ is *not* the *only* correct choice in the learning to rank setting.

Lemma 2. *The maximum dimension of the linear function space $\mathcal{F}$, consisting of valid scoring functions for Learning to Rank problems, is $2 * d$. Moreover, each W_i matrix can have one of the following 2 structures: a) ith row will be a vector of maximum dimension d and all other rows be 0 vectors and the non-zero row will be same across each matrix. b) ith row will be vector of maximum dimension d and every other row is same and of maximum dimension d. Moreover, the same rows are permuted across the matrices.*

Proof. An important property in ranking is permutation invariance. This means that score assigned to documents should be independent of the order in which documents are listed. Formally, a linear scoring function can be used for ranking if it satisifies the permutation invariance property: $\{\forall\ \pi \in S_m, \forall\ X \in \mathbb{R}^{m \times d} |\ \pi f(X) = f(\pi X)\}$. We prove that the space of linear functions which satisifies this property has dimension no more than d.

Using the full parameterization model, the permutation invariance property translates into: $P[\langle X, W_1\rangle, \ldots, \langle X, W_m\rangle] = [\langle PX, W_1\rangle, \ldots, \langle PX, W_m\rangle], \forall\ P$, where P is permutation matrix of order m. We assume that none of matrices W_i are 0 matrices.

Let $\rho_1 = \{P : \pi_P(1) = 1\}$, where $\pi_P(i)$ denotes the index of the element in the ith position of the permutation induced by P. Then, $\forall\ P \in \rho_1$, $\langle X, W_1\rangle = \langle PX, W_1\rangle \implies \mathrm{Tr}({W_1}^\top X) =$

$\mathrm{Tr}({W_1}^\top PX)$. Using the fact that $\mathrm{Tr}(AC) = \mathrm{Tr}(BC) \implies A = B$, this shows that ${W_1}^\top = {W_1}^\top P$. This clearly shows that there are 2 possible structures of ${W_1}^\top$, i.e, a) All but the 1st column of ${W_1}^\top$ are 0 vectors, or b) all but possibly the 1st column of ${W_1}^\top$ are same. The same arguement can be repeated to show the same for ith column of ${W_i}^\top$. That is, for ${W_i}^\top$, i column is non-zero and all other columns are zero or all are same and possibly different from ith column.

Let $\rho_2 = \{P : \pi_P(1) = 2\}$. Then, $\forall\ P \in \rho_2$, $\langle X, W_2\rangle = \langle PX, W_1\rangle \implies W_2^\top = W_1^\top P$. P will put the 1st column of W_1 in 2nd position and create any other permutation of the other columns. Hence, the 2nd column of $W_2^\top$ will match the 1st column of $W_1^\top$, and all other columns of $W_2^\top$, other than possibly 2nd column, will match all other columns of $W_1^\top$, other than possibly 3rd column. This arguement can be extended to see thaat ith column of ${W_i}^\top$ matches across matrices, and all other columns of ${W_i}^\top$, other than possibly ith column matches all other columns of ${W_j}^\top$, other than possibly jth columns, $\forall\ i, j$.

Assuming all matrices are not same, rank 1 matrices, we conclude the proof. □

However, it needs to be noted that for the permissible rank 1 matrices, the matrices can be collapsed into a single vector. But for rank 2 matrices, this cannot be done. Since in practise, m, the number of documents per query is not fixed, permissible rank 2 matrices cannot be set up for learning.

Our main theorem on generalization error is applicable to *any* convex ranking surrogate with linear ranking function.

Before we state our main theorem on generalization error bound, we need some notations. For input matrix $X \in \mathcal{X}$, relevance vector $R \in \mathcal{Y}$, weight vector $w \in \mathbb{R}^d$, score vector $s^w = Xw$ and any convex (in first argument) surrogate loss function $\ell(s^w, R)$, we denote

$$L(w) = \mathbb{E}\left[\ell(s^w, R)\right] \tag{1}$$

The expectation is taken over the underlying joint distribution on $\mathcal{X} \times \mathcal{Y}$. We also define

$$w^\star = \operatorname*{argmin}_w\ L(w) \quad and \quad \hat{w}\ = \operatorname*{argmin}_w \left\{ \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \ell((s^w)^{(i)}, R^{(i)}) \right\} \tag{2}$$

where $((X^{(1)}, R^{(1)}), \ldots, (X^{(n)}, R^{(n)}))$ are iid samples from the underlying joint distribution. We now have our main theorem on generalization error bound.

Theorem 3. *Let $w \mapsto \ell(s^w, R)$ be convex and Lipschitz continuous w.r.t. w in the l_2 norm with constant L_2. Let $\hat{w}$ and $w^\star$ be defined as in Eq. 2, with the further restriction that $\|w\|_2 \leq B$. Then, with a sample size of n, and with $\lambda = O(1/\sqrt{n})$, we have*

$$\mathbb{E}\left[L(\hat{w})\right] \leq L(w^\star) + 2\,L_2\,B\left(\frac{8}{n} + \sqrt{\frac{2}{n}}\right) \tag{3}$$

where the expectation is taken over input sample $((X^{(1)}, R^{(1)}), \ldots, (X^{(n)}, R^{(n)}))$.

Lipschitz continuity of $\ell(s^w, R)$ w.r.t w in l_2 norm means that there is a constant L_2 such that $|\ell(s^w, R) - \ell(s^{w'}, R)| \leq L_2\|w - w'\|_2$, for all $w, w' \in \mathbb{R}^m$. By duality, it follows that $L_2 \geq \sup_w \|\nabla_w \ell(s^w, R)\|_2$. Now, by chain rule, we have $\|\nabla_w \ell(s^w, R)\|_2 = \|X^\top \nabla_{s^w} \ell(s^w, R)\|_2$. It turns out that if each row of X is bounded in l_2 norm by R_X and $\|\nabla_{s^w} \ell(s^w, R)\|_1 \leq \widetilde{L}_1$ then $\|X^\top \nabla_{s^w} \ell(s^w, R)\|_2 \leq R_X \widetilde{L}_1$ and the bound in Theorem 3 becomes $O(\widetilde{L}_1 B R_X/\sqrt{n})$. This immediately gives the following corollary, which provides a sufficient condition for an m independent generalization bound to hold.

Corollary 4. *A sufficient condition for the ranking surrogate $\ell(s^w, R)$ to have m independent generalization bound is for it have m independent Lipschitz bound, w.r.t s^w, in l_∞ norm. That is, there is a constant $\widetilde{L}_1$, independent of m, such that $\widetilde{L}_1 \geq \sup_{s^w} \|\nabla_{s^w} \ell(s^w, R)\|_1$*

We point out that the generalization bound in Theorem 3 depends on the Lipschitz constant of $\ell(\cdot, R)$ w.r.t w. However, the condition in Corollary 4 depends on the Lipschitz constant of $\ell(\cdot, R)$ w.r.t s^w (the tilde in $\widetilde{L}$ serves a reminder that Lipschitz continuity is meant w.r.t. s^w, not w).

Our generalization bound is always better than the bound of Theorem.1, which is $O(\widetilde{L}_2 B R_X \sqrt{m/n})$, where $\widetilde{L}_2$ is the Lipschitz constant of the surrogate w.r.t s^w in l_2-norm. This is because $\widetilde{L}_1 \leq \sqrt{m}\widetilde{L}_2$ and their bound has inherent dependence on m. However, the price we pay is that our result only holds for *convex* surrogates whereas that of [1] holds for any Lipschitz surrogate.

Existing convex surrogates: It is easy to calculate that cross entropy surrogate used in ListNet has m independent generalization bound which the hinge like surrogate used in RankSVM has m dependent generalization bound.

4 Conclusion

While we have established a condition for convex ranking surrogates to have m independent generalization bounds, the question for general Lipschitz surrogates remain open. It is well known that in classification, Lipschitz continuity and smoothness affect the generalization error rate of surrogates. In learning to rank, it would be interesting to investigate the same properties. We do note that some preliminary work has been done in that regard [5], which introduces a $log(m)$ dependent generalization bound for Lipschitz losses.

We acknowledge the support of NSF under grant IIS-1319810.

References

[1] O. Chapelle and M. Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

[2] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, pages 422–446, 2002.

[3] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York., 1999.

[4] Olivier Chapelle, Yi Chang, and Tie-Yan Liu. Future directions in learning to rank. In *JMLR Workshop and Conference Proceedings*, pages 91–100, 2011.

[5] Ambuj Tewari and Sougata Chaudhuri. On lipschitz continuity and smoothness of loss functions in learning to rank. *arXiv preprint arXiv:1405.0586*, 2014.

[6] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

5 Generalization Bound Proofs

Our theorem is developed from the expected version of Theorem 6. in [6], which is originally given in probabilistic form. The expected version is as follows:

Let $f(w, z)$ be a λ strongly convex and L-Lipschitz (in $\|\cdot\|_2$) function in w. We define $F(w) = E_z f(w, z)$ and $w^* = \operatorname{argmin}_w F(w)$. Let $z_1, .., z_n$ be i.i.d sample and $\hat{w} = \operatorname{argmin}_w \frac{1}{n}\sum_{i=1}^n f(w, z_i)$. Then $E[F(\hat{w}) - F(w^*)] \leq \frac{4L^2}{\lambda n}$, where the expectation is taken over the sample. The above equation can be proved by carefully going through the proof of Theorem 6. in [6].

We now derive the expected version of Theorem 7 in [6]. We start with some definitions. Let $f(w, z)$ be a convex function in w. We define $R(w) = E_z f(w, z)$. For i.i.d random sample $z_1, ..., z_n$, the population and regularized empirical minimizers are defined as follows

$$w^* = \operatorname*{argmin}_w R(w), \quad \hat{w}_\lambda = \frac{\lambda}{2}\|w\|_2^2 + \operatorname*{argmin}_w \frac{1}{n}\sum_{i=1}^n f(w, z_i) \tag{4}$$

Theorem 5. *Let $f : W \times Z \to \mathbb{R}$ be such that W is bounded by B in $\|\cdot\|_2$, and $f(w, z)$ is convex and L-Lipschitz in $\|\cdot\|_2$ with respect to w. Let $z_1, ..., z_n$ be an i.i.d. sample and let $\lambda = \sqrt{\frac{\frac{4L^2}{n}}{\frac{B^2}{2} + \frac{4B^2}{n}}}$. Then for $\hat{w}_\lambda$ and w^* in Eq.4, we have*

$$E[R(\hat{w}) - R(w^*)] \leq 2LB\left(\frac{8}{n} + \sqrt{\frac{2}{n}}\right) \tag{5}$$

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 $\lambda B + L$ in $\|\cdot\|_2$. Applying expected version of Theorem 6 in [6], we get

$$E(\tfrac{\lambda}{2}\|\hat{w}_\lambda\|_2^2 + R(\hat{w}_\lambda)) \leq \inf_w \{\tfrac{\lambda}{2}\|w\|_2^2 + R(w) + \tfrac{4(\lambda B + L)^2}{\lambda n}\} \leq \tfrac{\lambda}{2}\|w^*\|_2^2 + R(w^*) + \tfrac{4(\lambda B+L)^2}{\lambda n}$$

$$\Rightarrow E(R(\hat{w}_\lambda) - R(w^*)) \leq \tfrac{\lambda B^2}{2} + \tfrac{4(\lambda B+L)^2}{\lambda n}$$

Minimizing the upper bound w.r.t λ, we get $\lambda = \sqrt{\frac{4L^2}{n}}\sqrt{\frac{1}{\frac{B^2}{2} + \frac{4B^2}{n}}}$. Plugging it back in the equation and using the relation $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we get Theorem 5. □

Taking $f(\cdot, \cdot) = \ell(\cdot, \cdot)$ and $Z = (X, R)$, we get the proof of Theorem.3 .

Proof of Corollary 4 : From the definitions preceding Theorem 3, we have $f(w, z) = \ell(s^w, R)$, where $z = (X, R)$ and $s^w = Xw$. Thus, we get the following relation

$$L_2 = \|\nabla_w \ell(s^w, R)\|_2 \leq \|X^\top \nabla_{s^w} \ell(s^w, R)\|_2 \leq \|X^\top\|_{p\to 2}\|\nabla_{s^w}\ell(s^w, R)\|_p \leq R_X^p \widetilde{L}_p,$$

where $R_X^p \geq \sup_{X\in\mathcal{X}} \|X^\top\|_{p\to 2}$

Now putting $p = 1$, we get $\|X^\top\|_{1\to 2} = \max_{u=1} \frac{\|X^\top u\|_2}{\|u\|_1}$. Denoting $X_i^\top$ as the ith column of $X^\top$, we have $\|X^\top u\|_2 = \|\sum_{i=1}^m X_i^\top u_i\|_2 \leq \sum_{i=1}^m |u_i| \|X_i^\top\|_2 \leq \|u\|_1 \max_{i=1,..,m} \|X_i^\top\|_2$. Since $X_i^\top$ is the d dimensional vector representation of a document, assuming bound R_X on l_2 norm of each feature vector, we get $\|X^\top\|_{1\to 2} \leq R_X$.

Thus, we need $\widetilde{L}_1 = \sup_{s^w} \|\nabla_{s^w}\ell(s^w, R)\|_1$ to be m independent constant.