Fine-grained generalization analysis of vector-valued learning

Wu, Liang; Ledent, Antoine; Lei, Yunwen; Kloft, Marius

[Link to publication on Research at Birmingham portal](#)

Fine-grained Generalization Analysis of Vector-valued Learning

Liang Wu [1], Antoine Ledent [2], Yunwen Lei [3, 2] and Marius Kloft [2]

[1]Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China
[2]Department of Computer Science, TU Kaiserslautern, 67653 Kaiserslautern, Germany
[3]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom

Abstract

Many fundamental machine learning tasks can be formulated as a problem of learning with vector-valued functions, where we learn multiple scalar-valued functions together. Although there is some generalization analysis on different specific algorithms under the empirical risk minimization principle, a unifying analysis of vector-valued learning under a regularization framework is still lacking. In this paper, we initiate the generalization analysis of regularized vector-valued learning algorithms by presenting bounds with a mild dependency on the output dimension and a fast rate on the sample size. Our discussions relax the existing assumptions on the restrictive constraint of hypothesis spaces, smoothness of loss functions and low-noise condition. To understand the interaction between optimization and learning, we further use our results to derive the first generalization bounds for stochastic gradient descent with vector-valued functions. We apply our general results to multi-class classification and multi-label classification, which yield the first bounds with a logarithmic dependency on the output dimension for extreme multi-label classification with the Frobenius regularization. As a byproduct, we derive a Rademacher complexity bound for loss function classes defined in terms of a general strongly convex function.

Introduction

In machine learning, we often encounter learning tasks involving vector-valued prediction functions (Álvarez, Rosasco, and Lawrence 2012; Xu et al. 2019). As examples consider the following. In multi-class classification (MCC), we aim to assign each instance to a single label class (Mohri, Rostamizadeh, and Talwalkar 2012; Crammer and Singer 2002). In multi-label classification (MLC), each instance may be annotated with one or multiple class labels (Zhou et al. 2012; Yu et al. 2014). In both cases, we build one scalar-valued prediction function per class. Together these functions form a vector-valued predictor with the output dimension being the total number of label classes. These two important learning tasks have found wide applications in real systems, where they are used, for instance, for image and video annotation, face classification, and query/keyword suggestion (Yu et al. 2014). Another popular learning setting where we encounter vector-valued prediction functions is multi-task learning (MTL). Here we build, for each task, a distinct predictor. The benefit of learning these predictors together is that of exploiting a shared hidden structure in these tasks (Zhang and Yang 2017; Maurer and Pontil 2016; Yousefi et al. 2018; Argyriou et al. 2007; Ciliberto et al. 2017). The more these tasks are related, the larger the structure and the benefit of learning them together. We refer to problems of learning a vector-valued prediction function—such as the above ones—as *vector-valued learning problems* (Micchelli and Pontil 2005; Liu et al. 2019; Xu et al. 2019; Álvarez, Rosasco, and Lawrence 2012; Lei et al. 2015a).

An important measure of the quality of vector-valued learning models is their generalization performance, i.e., their ability to generalize their empirical behavior on training examples to unseen test data. As a central topic in statistical learning theory, generalization analysis has received a lot of attention. Other than providing an intuitive understanding of how different parameters affect the learning performance, generalization analysis is also effective in designing novel learning machines (Cortes, Kloft, and Mohri 2013).

Unlike traditional binary classification problems, a distinguished property of vector-valued learning is that the output dimension plays an important role in the analysis (Reddi et al. 2019). This is especially the case for problems with a huge output dimension, which are becoming more and more ubiquitous in the big data era. One such example is eXtreme Classification (XC; Bengio et al. 2019). Here we deal with multi-class and multi-label problems involving an extremely large total number of potential class labels (Jain et al. 2019). Generalization bounds with an emphasis on the output dimension were developed for specific vector-valued learning problems such as MCC (Zhang 2004; Lei et al. 2015b; Guermeur 2017; Li et al. 2018; Musayeva, Lauer, and Guermeur 19) or MLC (Yu et al. 2014; Liu et al. 2018; Shen et al. 2018; Xu et al. 2016; Khandagale, Xiao, and Babbar 2020).

Recently, Maurer and Pontil (2016); Li, Liu, and Wang (2019) initiated the study of the general framework of vector-valued learning. These studies exhibit the following limitations. First, they exploit the Lipschitz continuity of loss functions with respect to (w.r.t.) the Euclidean norm, while typical loss functions occurring in vector-valued learning problems are Lipschitz continuous w.r.t. the ℓ_∞-norm (Lei et al. 2019), with a comparable Lipschitz constant as for the Euclidean norm (note ℓ_∞-norm can be significantly smaller than the Euclidean norm). This mismatch between

generalization analysis and Lipschitz continuity induces, for standard Euclidean regularization, a square-root dependency on the number of components (that is, the number of classes or tasks in MCC/MLC and MTL, respectively) (Maurer and Pontil 2016; Li, Liu, and Wang 2019). Second, there is the following conceptual mismatch between algorithms and theory. While the theory is developed for empirical risk minimization (ERM) in a constrained space (Maurer and Pontil 2016; Li, Liu, and Wang 2019; Reeve and Kaban 2020), in practice, regularization schemes are used, which are oftentimes easier to solve (Yu et al. 2014; Lei et al. 2015b; Li et al. 2018; Li, Liu, and Wang 2019). Third, the existing generalization analysis fails to take into account the computational properties of the algorithm, which is important to understand the interaction between optimization and learning. Lastly, most existing studies are limited to the classic regime of "slow" rates ($\Omega(n^{-\frac{1}{2}}$, where n is the sample size). In order to achieve so-called *fast rates* (Bartlett, Bousquet, and Mendelson 2005), they require restrictive assumptions, such as the smoothness of loss function (Reeve and Kaban 2020), the capacity assumption on the hypothesis space, or the existence of a model with vanishing errors.

In this paper, we address all the above issues. Our contributions are as follows.

1. We show the first generalization bounds for general vector-valued learning using a *regularization* framework, thus removing the gap between the algorithm that is analyzed in theory and the one that is considered in practice.

2. Not only do we analyze a more realistic model, we dramatically improve the best known dependency of bounds for general vector-valued learning (in either framework). For instance for standard Frobenius regularization, we drop the dependency on the number of components in the model from $\sqrt{c}$ to $\log c$.

3. In MLC, the components are the classes. Thus our result establishes guarantees that scale logarithmic in the number of classes. This is remarkable because the previously best result for MLC scaled square root in the number of classes. Thus, for the first time, we establish non-void theory for extreme multi-label classification (yet, as mentioned in Point 1, our analyzed algorithm is more realistic). Note that MLC is the by far most common scenario in XC.

4. Our results apply also to the fast-rate regime, that is, the learning scenario where bounds enjoy a fast decay on n. In this regime, our analysis improves not only the best known rates in c (see Point 2) over previous work, but lifts also assumptions employed therein on the smoothness of the loss function and hypothesis space.

Related Work

Here we survey the related work on vector-valued learning.

There is a large body of work on the generalization analysis of MCC (Lei et al. 2019; Thrampoulidis, Oymak, and Soltanolkotabi 2020), based on various capacity measures: e.g., covering numbers (Zhang 2004; Lei et al. 2019), the fat-shattering dimension (Guermeur 2017), and (local) Rademacher complexities (Mohri, Rostamizadeh, and Tal-

walkar 2012; Lei et al. 2015b; Maurer 2016; Cortes et al. 2016; Li et al. 2018; Maximov, Amini, and Harchaoui 2018; Musayeva, Lauer, and Guermeur 19, 2018; Deshmukh et al. 2019). Unlike binary classification problems, the number of classes plays here an important role in the generalization performance. Until recently, the coupling among the class components, while exploited by most practical multiclass algorithms (Crammer and Singer 2002), was ignored by generalization bounds. As a result, they exhibited at least a linear dependency on the number of classes (Mohri, Rostamizadeh, and Talwalkar 2012). Subsequently, several works aimed to improve this dependency through structural results on Gaussian and Rademacher complexities that exploit the coupling among classes, first achieving a square root (Guermeur 2017; Maurer 2016; Lei et al. 2015b) and later on a logarithmic dependency (Lei et al. 2019). For the low-noise regime, Li et al. (2018) show a fast-rate generalization bound based on local Rademacher complexities.

There is far less work on the theoretical analysis of MLC. The consistency of MLC with different loss functions was studied in Gao and Zhou (2011). For decomposable loss functions involving the specific least-squares loss, generalization bounds based on the Rademacher complexity were derived in Yu et al. (2014). The best known dependency on the output dimension is square root and was shown in Liu et al. (2018); Wu and Zhu (2020). Xu et al. (2016) bounded the local Rademacher complexity of MLC, which motivated the authors to study a novel MLC algorithm based on the tail-sum of singular values of the predictors.

Generalization analysis for general vector-valued learning algorithms was initiated by Maurer and Pontil (2016) and put forward by Li, Liu, and Wang (2019). However, their analysis implies generalization bounds with a square-root dependency on the output dimension for reasonable regularizers such as those based on the Frobenius norm. This is because they work with the Lipschitz continuity w.r.t. the Euclidean norm. In the present work, we consider the infinity norm instead. Notice that the existing results for MCC, MLC or general vector-valued learning algorithms are mainly established for ERM in a constrained hypothesis space (Maurer and Pontil 2016; Li, Liu, and Wang 2019; Reeve and Kaban 2020). As a comparison, there is scarce work on vector-valued learning based on regularization, while in practice this is the mostly used scheme (Crammer and Singer 2002; Lei et al. 2015b; Yu et al. 2014).

Problem Formulation and Results

Problem formulation

We describe here the framework of vector-valued learning with regularization. Let ρ be a probability measure defined over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is an input space and $\mathcal{Y}$ is an output space (d is the input dimension). Let $S = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$ be the training dataset drawn independently from ρ. For vector-valued learning, we aim to build a vector-valued predictor $h : \mathcal{X} \mapsto \mathbb{R}^c$, i.e., $h = (h_1, \ldots, h_c)$ with $h_j : \mathcal{X} \mapsto \mathbb{R}$. We denote by c the output dimension. We consider non-parametric learning in a reproducing kernel Hilbert space $\mathcal{H}_K$ associated with a Mer-

cer kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Let $\phi : \mathcal{X} \mapsto \mathcal{H}_K$ be the corresponding feature map, i.e., $K(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \mathcal{X}$ with $\langle \cdot, \cdot \rangle$ being the inner product. Consider the hypothesis space

$$\mathcal{W} = \{ \mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_c) \in \mathcal{H}_K^c \}.$$

We consider vector-valued predictors of the form

$$h^{\mathbf{w}}(x) = \left(h_1^{\mathbf{w}}(x), \ldots, h_c^{\mathbf{w}}(x) \right),$$

where $\mathbf{w} \in \mathcal{W}$ and $h_j^{\mathbf{w}}(x) = \langle \mathbf{w}_j, \phi(x) \rangle$. Note that if $\phi(x) = x$, then the model $h^{\mathbf{w}}$ can be characterized by a matrix $\mathbf{w}$ in $\mathbb{R}^{d \times c}$ with $\mathbf{w}_j$ being the j-th column. The performance of $h^{\mathbf{w}}$ on a single example z is measured by a loss function $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$. An effective approach to building a model is to learn with regularization, where we build an objective function $F_S : \mathcal{W} \mapsto \mathbb{R}_+$ by

$$F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}; z_i) + r(\mathbf{w}). \tag{1}$$

The first term characterizes the empirical behavior of a model $\mathbf{w}$ on the training examples S, while the second term $r : \mathcal{W} \mapsto \mathbb{R}_+$ is a regularizer. The predictor is then established by minimizing the objective function over the hypothesis space (regularized risk minimization, RRM), i.e., $\mathbf{w}_S = \arg\min_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w})$. The generalization behavior of a model on a testing example can be measured by the regularized risk defined by $F(\mathbf{w}) = \mathbb{E}_Z \left[\ell(\mathbf{w}; Z) \right] + r(\mathbf{w})$, where $\mathbb{E}_Z$ denotes the expectation w.r.t. Z. As we will show in applications, this framework of vector-valued learning covers important learning tasks such as MCC and MLC by considering specific instantiations of the output space, loss functions and regularizers. For any $k \in \mathbb{N}$, let $[k] = \{1, \ldots, k\}$.

To study high-probability bounds for excess regularized risks of $\mathbf{w}_S$, we introduce some necessary assumptions. The first assumption is the Lipschitz continuity of loss functions.

Assumption 1 (Lipschitz continuity). We assume ℓ satisfies a Lipschitz continuity w.r.t. the infinity norm as follows

$$\left| \ell(\mathbf{w}; z) - \ell(\mathbf{w}'; z) \right| \leq L \| h^{\mathbf{w}}(x) - h^{\mathbf{w}'}(x) \|_\infty, \quad (2)$$

where $L > 0$ and $\|\mathbf{t}\|_\infty = \max_{j \in [c]} |t_j|$ for $\mathbf{t} = (t_1, \ldots, t_c)$.

A notable property is that we consider the Lipschitz continuity w.r.t. the infinity-norm instead of the Euclidean norm in the literature (Li et al. 2018; Li, Liu, and Wang 2019). Although these norms are equivalent, the involved Lipschitz constant can differ up to a factor of $\sqrt{c}$ which plays an important role in the generalization behavior if the output dimension is large. Therefore, the L-Lipschitz condition w.r.t. infinity norm is much stronger than that w.r.t. the Euclidean norm. Fortunately, popular loss functions in MLC and MCC actually satisfy the more restrictive condition of the Lipschitz continuity w.r.t. infinity-norm, where the Lipschitz constant is independent of c. This is why we can exploit the restrictive assumption on infinity-norm to develop a bound with a better dependency on c, which is an advantage over the analysis w.r.t. Euclidean norm (Li, Liu, and Wang 2019).

Our second assumption is the (strong) convexity of loss functions and regularizers (Sridharan, Shalev-Shwartz, and Srebro 2009; Kakade, Shalev-Shwartz, and Tewari 2012).

Assumption 2. We assume ℓ is convex w.r.t. the first argument. We also assume r is σ-strongly convex w.r.t. a norm $\| \cdot \|$, i.e., for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$

$$r(\mathbf{w}) \geq r(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', r'(\mathbf{w}') \rangle + \frac{\sigma}{2} \| \mathbf{w} - \mathbf{w}' \|^2,$$

where $r'(\mathbf{w}')$ denotes a subgradient of r at $\mathbf{w}'$.

A popular strongly convex regularizer is $r(\mathbf{w}) = \frac{\sigma}{2} \|\mathbf{w}\|_{2,p}^2$, where $\|\mathbf{w}\|_{2,p} = \left(\sum_{j=1}^c \|\mathbf{w}_j\|_2^p \right)^{\frac{2}{p}}$ is the $\ell_{2,p}$ norm ($p \geq 1$). Here $\| \cdot \|_2$ denotes the norm in $\mathcal{H}_K$ induced by $\langle \cdot, \cdot \rangle$. It is known that this regularizer is $\sigma(p-1)$-strongly convex w.r.t. $\| \cdot \|_{2,p}$ for $p \in (1, 2]$ (Kakade, Shalev-Shwartz, and Tewari 2012). Another popular regularizer is $r(\mathbf{w}) = \frac{\sigma}{2} \|\mathbf{w}\|_{S_p}^2$ for $\mathbf{w} \in \mathbb{R}^{d \times c}$ where $\|\mathbf{w}\|_{S_p}$ is the Shatten-p norm (Kakade, Shalev-Shwartz, and Tewari 2012). This regularizer is $(p-1)\sigma$-strongly convex w.r.t. $\| \cdot \|_{S_p}$ for $p \in (1, 2]$. Furthermore, we always assume $\|\mathbf{w}\| \geq \|\mathbf{w}\|_{2,\infty}$ for all $\mathbf{w} \in \mathcal{W}$. This is a very mild assumption and is satisfied for all the norm considered in this paper. Finally, we assume $\sup_{x \in \mathcal{X}} \|\phi(x)\|_2 \leq \kappa$.

Main Results

We now present our results. We use Rademacher complexity to measure the capacity of hypothesis spaces.

Definition 1 (Rademacher complexity). Let $\mathcal{H}$ be a class of real-valued functions defined over a space $\mathcal{Z}$ and $S = \{z_i\}_{i=1}^n \in \mathcal{Z}^n$. The *empirical Rademacher complexities* of $\mathcal{H}$ with respect to S is defined as

$$\mathfrak{R}_S(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i) \right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher variables, i.e., ϵ_i takes an equal probability of being either 1 or -1.

Our first result is an upper bound on the Rademacher complexity of loss function classes. This result is general in the sense that we define hypothesis spaces in terms of a general strongly convex function $\tau(\mathbf{w}) = F(\mathbf{w}) - F(\mathbf{w}^*)$, where $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w})$. For any $\Lambda > 0$, we denote

$$\mathcal{W}_\Lambda = \{ \mathbf{w} \in \mathcal{W} : F(\mathbf{w}) - F(\mathbf{w}^*) \leq \Lambda \} \tag{3}$$

and the associated class of loss functions

$$\mathcal{F}_\Lambda = \{ \mathbf{w} \mapsto \ell(\mathbf{w}; z) : \mathbf{w} \in \mathcal{W}_\Lambda \}. \tag{4}$$

Theorem 1 (Rademacher complexity bound). *Let Assumptions 1 and 2 hold. Then there exists a constant C_1 independent of n, c, L and Λ such that*

$$\mathfrak{R}_S(\mathcal{F}_\Lambda) \leq \frac{C_1 L \sqrt{2\Lambda} \widetilde{B} \log^2(nc)}{\sqrt{n\sigma}},$$

where $\widetilde{B} = \sup_{(x,j)} \|\tilde{\phi}_j(x)\|_*$ and $\| \cdot \|_*$ is the dual norm of $\| \cdot \|$ in $\mathcal{W}$. Here for any $x \in \mathcal{X}$ and $j \in [c]$ we use the notation $\tilde{\phi}_j(x) := \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(x), \underbrace{0, \ldots, 0}_{c-j} \big) \in \mathcal{H}_K^c$.

With different loss functions and regularizers, Theorem 1 immediately implies specific Rademacher complexity bounds for different vector-valued learning problems. It is worth mentioning that the upper bound enjoys a logarithmic dependency on the output dimension (we treat Λ as a constant which is problem-dependent and should be tuned by cross validation in practice). Therefore, this result is particularly useful for large-scale learning problems. This mild dependency is derived by exploiting the Lipschitz continuity of loss functions w.r.t. $\|\cdot\|_{\infty}$ and a structural result to capture this Lipschitz continuity. The proof of Theorem 1 is given in Section A.1 in the Appendix.

Remark 1. We now compare this result with the existing Rademacher complexity bounds for loss function classes in vector-valued learning problems. Initially, the Rademacher complexity bounds (Lemma 8.1 in Mohri, Rostamizadeh, and Talwalkar (2012)) failed to capture the coupling of different components reflected by the constraint (e.g., $\|\mathbf{w}\| \leq \Lambda$ for some $\|\cdot\|$) and therefore presented a crude dependency on the output dimension. These results are improved in Lei et al. (2015b) and Maurer (2016) by exploiting the Lipschitz continuity of loss functions w.r.t. ℓ_2 norm. Very recently, the Lipschitz continuity w.r.t. ℓ_{∞} norm is also considered in the literature (Lei et al. 2019; Foster and Rakhlin 2019). However, the analysis in Foster and Rakhlin (2019) failed to exploit the coupling among components while the analysis in Lei et al. (2019) only exploited this coupling enforced by some specific norms (case by case investigation is required). For example, they require Khintchine inequalities for vectors and matrices to handle $(2, p)$-norm and Schatten p-norm, respectively. As a comparison, Theorem 1 provides a more general result where $\mathbf{w}$ is constrained via a strongly convex function. A nice property is that Theorem 1 treats $(2, p)$-norm and Schatten p-norm exactly the same, and does not need case-by-case discussions. Moreover, Theorem 1 also applies to other regularization schemes, e.g., learning with entropic regularizer in a probability simplex.

We now present upper bounds on the excess regularized risk for vector-valued learning. We use a variant of the big-O notation $\widetilde{O}$ to hide any logarithmic factor. The proof is given in Section A.2 in the Appendix.

Theorem 2 (Regularized risk bound for RRM). *Let Assumptions 1 and 2 hold. Let $\delta \in (0, 1)$. With probability at least $1 - \delta$ Eq. (5) holds uniformly for all $\mathbf{w} \in \mathcal{W}$*

$$F(\mathbf{w}) - F(\mathbf{w}^*) = \max\left\{ F_S(\mathbf{w}) - F_S(\mathbf{w}^*), 1/(n\sigma) \right\} +$$
$$\widetilde{O}\left(\left(\frac{\left(F(\mathbf{w}) - F(\mathbf{w}^*)\right)\left(\log(1/\delta) + \widetilde{B}\right)}{n\sigma} \right)^{\frac{1}{2}} \right). \quad (5)$$

In particular, the following inequality holds with probability at least $1 - \delta$

$$F(\mathbf{w}_S) - F(\mathbf{w}^*) = \widetilde{O}\left(\frac{\log(1/\delta) + \widetilde{B}}{n\sigma} \right). \quad (6)$$

Eq. (5) applies uniformly to all $\mathbf{w} \in \mathcal{W}$ while (6) applies to the RRM scheme. Both of these upper bounds admit a very mild dependency on the output dimension. Unlike ERM which generally implies slow rates $O(1/\sqrt{n})$ (Lei

et al. 2019), we establish fast rates of the order $\widetilde{O}(1/n)$ for RRM by leveraging the strong convexity of the objective function. A notable property of this result is its generality. It requires only Lipschitz continuity of loss functions and strong convexity of objective functions to develop meaningful generalization bounds for any vector-valued learning methods. It does not impose a capacity assumption on hypothesis spaces or a low-noise condition for fast generalization bounds as was required in the existing treatments (Li et al. 2018; Li, Liu, and Wang 2019).

Remark 2. Generalization error bounds with an implicit dependency on the output dimension were recently derived for ERM with vector-valued functions (Li, Liu, and Wang 2019). The discussions there imposes a constraint in terms of either $\|\mathbf{w}\|_{2,1}$ or $\|\mathbf{w}\|_{S_1}$ on the hypothesis space. As a comparison, we get a logarithmic dependency by considering a regularizer involving a much milder norm $\|\cdot\|_{2,2}$, (note $\|\cdot\|_{2,1}$ can be as large as $\sqrt{c}\|\cdot\|_{2,2}$). Indeed, the analysis in Li, Liu, and Wang (2019) would imply a generalization bound with a square-root dependency on c if the constraint $\|\mathbf{w}\|_{2,1} \leq \Lambda$ there is replaced by $\|\mathbf{w}\|_{2,2} \leq \Lambda$. We achieve this improvement by exploiting the Lipschitz continuity of loss functions w.r.t. the infinity-norm instead of w.r.t. the Euclidean norm. Although the Lipschitz contintuity w.r.t. infinity-norm is a much stronger assumption than that w.r.t. Euclidean norm, it is satisfied by popular loss functions with the Lipschitz constant independent of c (cf. Section on Applications). Our analysis is able to fully exploit this stronger assumption and therefore gets a tighter dependency on c. The use of strong convexity allows for a tighter dependency on n.

Theorem 2 establishes excess risk bounds for $\mathbf{w}_S$ under RRM, which is non-constructive in the sense that it does not tell us how the model $\mathbf{w}_S$ is built from the data. In the following theorem, we will consider a constructive algorithm called SGD, which is a very popular optimization algorithm especially useful for large data analysis due to its simplicity and efficiency. To study its excess risk bounds, we need to consider both statistical and computational properties as well as the trade-off realized by early-stopping. We give the proof in Section A.3 (Appendix).

Theorem 3 (Regularized risk bound for SGD). *Assume $\ell(\mathbf{w}; z)$ takes the form $\ell(\mathbf{w}; z) = \psi(h^{\mathbf{w}}(x), y)$ with $\psi : \mathbb{R}^c \times \mathcal{Y} \mapsto \mathbb{R}$ being a convex function w.r.t. the first argument. Let Assumption 1 hold and $r(\mathbf{w}) = \frac{\sigma}{2}\|\mathbf{w}\|_{2,2}^2$. Let $f(\mathbf{w}; z) = \ell(\mathbf{w}; z) + r(\mathbf{w})$. Let $\mathbf{w}_1 = 0 \in \mathcal{W}$ and $\{\mathbf{w}_t\}_{t \in \mathbb{N}}$ be the sequence produced by SGD, i.e.,*

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t f'(\mathbf{w}_t; z_{i_t}), \quad (7)$$

where $\eta_t = 1/(t\sigma)$ and $\{i_t\}_t$ is independently drawn from the uniform distribution over $[n]$. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ the following holds:

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) = \widetilde{O}\left(\frac{\log(1/\delta)}{\min\{n, T\}\sigma} \right). \quad (8)$$

If T is of the order of n then with probability at least $1 - \delta$,

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) = \widetilde{O}\left(\frac{\log(1/\delta)}{n\sigma} \right). \quad (9)$$

Table 1: Generalization bounds for MCC in a Frobenius learning framework (either constraint $\|W\|_F \leq 1$ or regularizer $\|W\|_F^2$)

Bound	Lipschitz Continuity	Additional Assumption	Method	Reference
$c/\sqrt{n}$	ℓ_1-norm	No	ERM	Kuznetsov, Mohri, and Syed (2014)
$\sqrt{c/n}$	ℓ_2-norm	No	ERM	Lei et al. (2015b)
$\sqrt{c}\log^3(n)/n$	ℓ_2-norm	smoothness and low noise	ERM	Li et al. (2018)
$\log^2(nc)/\sqrt{n}$	ℓ_∞-norm	No	ERM	Lei et al. (2019)
$\sqrt{c/n}$	ℓ_2-norm	decay of singular values	ERM	Li, Liu, and Wang (2019)
$\log^3(nc)/n$	ℓ_∞-norm	strong convexity	RRM/SGD	This work

Remark 3. According to Theorem 3, there is no need to run more iterations once T is of the order of n since further computation would not further improve the generalization performance. In other words, a computationally efficient method is to early-stop SGD after a constant number of passes over the data. To our best knowledge, this is the first result on the generalization analysis of SGD for vector-valued learning algorithms.

Applications

In this section, we apply our general results to specific vector-valued learning tasks, including MCC and MLC. Let $\tilde{\ell} : \mathbb{R} \mapsto \mathbb{R}$ be convex, decreasing and $(L/2)$-Lipschitz continuous. Popular choices of $\tilde{\ell}$ includes the hinge-loss $\tilde{\ell}(t) = \max\{0, 1-t\}$ and the logistic loss $\tilde{\ell}(t) = \log(1+\exp(-t))$. We only consider applications to RRM here. It should be mentioned that our results directly apply to SGD.

Multi-class classification

MCC is a classic learning problem which aims to assign a *single* class label to each training example, i.e., $\mathcal{Y} = [c]$. In this problem setting, the j-th component of a model $h : \mathcal{X} \mapsto \mathbb{R}^c$ measures the likelihood of the output label being j. The prediction is made by taking the component with the largest value as the predicted label, i.e., $x \mapsto \arg\max_{j \in [c]} h_j(x)$. We apply our result to three classical MCC models: the multi-class SVM, the multinomial logistic regression and top-k SVM. In Table 1, we compare generalization bounds for MCC in a Frobenius learning framework, i.e., either there is a constraint $\|W\|_F \leq 1$ or a regularizer $\|W\|_F^2$.

Example 1 (Multi-class SVM). Let $\mathcal{Y} = [c]$. Consider the objective function F_S in (1) with
$$\ell(\mathbf{w}; z) = \max_{y' \in [c]: y' \neq y} \tilde{\ell}\big(\langle \mathbf{w}_y - \mathbf{w}_{y'}, \phi(x)\rangle\big)$$
and $r(\mathbf{w}) = \frac{\sigma}{2}\sum_{j \in [c]} \|\mathbf{w}_j\|_2^2$. This is a margin-based loss function and $\min_{y' \in [c]: y' \neq y} \langle \mathbf{w}_y - \mathbf{w}_{y'}, x\rangle$ is called the margin of $h^{\mathbf{w}}$ at $z = (x, y)$. It is clear that the loss function is designed to favor a model with a large margin. This objective function recovers the multi-class SVM in Crammer and Singer (2002) by taking $\tilde{\ell}$ as the hinge loss. To apply Theorem 2, it suffices to verify Assumptions 1, 2. It was shown that ℓ is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ (Lei et al. 2019). The convexity of ℓ follows from the convexity of $\tilde{\ell}$ and the linearity of hypothesis. The σ-strong convexity of r w.r.t.

$\|\cdot\|_{2,2}$ is clear in the literature (Kakade, Shalev-Shwartz, and Tewari 2012). The dual norm is $\|\cdot\|_{2,2}$, and therefore $\widetilde{B} \leq \kappa$. Therefore, we can apply Theorem 2 to develop the generalization bound $F(\mathbf{w}_S) - F(\mathbf{w}^*) = \widetilde{O}\big(\frac{\log(1/\delta)}{n\sigma}\big)$ with high probability.

Example 2 (Multinomial logistic regression). Let $\mathcal{Y} = [c]$. Consider the loss function
$$\ell(\mathbf{w}; z) = \log\Big(\sum_{j \in [c]} \exp\big(\langle \mathbf{w}_j - \mathbf{w}_y, \phi(x)\rangle\big)\Big).$$
Consider the objective function F_S (1) with the above loss function and the regularizer $r(\mathbf{w}) = \frac{\sigma}{2}\sum_{j \in [c]} \|\mathbf{w}_j\|_2^2$. This learning scheme recovers the popular multinomial logistic regression. It was shown that the above loss function is 2-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ (Lei et al. 2019). Assumption 2 also holds. Therefore, we can apply Theorem 2 to develop the generalization bound $F(\mathbf{w}_S) - F(\mathbf{w}^*) = \widetilde{O}\big(\frac{\log(1/\delta)}{n\sigma}\big)$ with probability at least $1 - \delta$.

Example 3 (Top-k SVM). Let $\mathcal{Y} = [c]$. Consider the loss
$$\ell(\mathbf{w}; z) = \max\Big\{0, \frac{1}{k}\sum_{j=1}^c \big(\mathbb{I}_{y \neq 1} + \langle \mathbf{w}_1 - \mathbf{w}_y, \phi(x)\rangle, \cdots,$$
$$\mathbb{I}_{y \neq c} + \langle \mathbf{w}_c - \mathbf{w}_y, \phi(x)\rangle\big)_{\{j\}}\Big\},$$
where $\mathbb{I}$ is the indicator function and for any $\mathbf{t} \in \mathbb{R}^c$ the notation $\{\cdot\}$ denotes a permutation such that $\{j\}$ is the index of the j-th largest score, i.e., $t_{\{1\}} \geq t_{\{2\}} \geq \cdots \geq t_{\{c\}}$. If we use the above loss function and the regularizer $r(\mathbf{w}) = \frac{\sigma}{2}\sum_{j \in [c]} \|\mathbf{w}_j\|_2^2$ in (1), we recover the top-k SVM useful to tackle the ambiguity in class labels (Lapin, Hein, and Schiele 2015). It was shown that the above loss function is 2-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ (Lei et al. 2019). Assumption 2 also holds. Therefore, we can apply Theorem 2 to develop the generalization bound $F(\mathbf{w}_S) - F(\mathbf{w}^*) = \widetilde{O}\big(\frac{\log(1/\delta)}{n\sigma}\big)$ with probability at least $1 - \delta$.

Remark 4. The existing generalization analysis of MCC considers ERM in a constrained hypothesis space (Lei et al. 2019). However, RRM is more popular in MCC. For example, the multi-class SVM, multinomial logistic regression and top-k SVM in the above three examples are proposed in a regularization setting (Lapin, Hein, and Schiele 2015; Crammer and Singer 2002). Furthermore, the existing discussions generally imply a slow generalization bound

$O(1/\sqrt{n})$ (Lei et al. 2015b, 2019) or require restrictive assumptions such as low-noise condition and capacity assumptions to achieve a fast generalization bound $O(1/n)$ (Li, Liu, and Wang 2019). As a comparison, our discussion implies a fast bound $O(1/n)$ without these assumptions.

Multi-label classification

For MLC, each training example can be associated with one or more class labels (Yu et al. 2014; Zhang and Zhou 2013; Xu, Liu, and Geng 2020; Dembczyński et al. 2012; Wu and Zhu 2020). This can be realized by setting $\mathcal{Y} = \{-1, +1\}^c$, i.e., each $y = (y^{(1)}, \ldots, y^{(c)})$ is a binary vector where $y^{(j)} = 1$ if the j-th label is relevant and $y^{(j)} = -1$ if the j-th label is irrelevant. A popular approach to MLC is to learn a vector-valued function $h : \mathcal{X} \mapsto \mathbb{R}^c$ and predict the output $\hat{y}$ for z according to its sign, i.e., $\hat{y}^{(j)} = \text{sgn}(h_j(x))$, where $\text{sgn}(a)$ denotes the sign of $a \in \mathbb{R}$. There are various performance measures to quantify the performance of a multi-label predictor, including the subset accuracy and the ranking loss. In Table B.1, we compare generalization bounds for MLC in a Frobenius learning framework, i.e., either a constraint $\|W\|_F \leq 1$ or a regularizer $\|W\|_F^2$.

Example 4 (Learning with subset loss)**.** Let $\mathcal{Y} = \{-1, +1\}^c$. Consider the loss function

$$\ell(\mathbf{w}; z) = \max_{j \in [c]} \tilde{\ell}\big(y^{(j)} \langle \mathbf{w}_j, \phi(x) \rangle\big). \quad (10)$$

This loss function is called the subset loss (Zhang and Zhou 2013). Note that $\tilde{\ell}(y^{(j)} \langle \mathbf{w}_j, \phi(x) \rangle)$ is a standard loss if we consider the prediction of the j-th label as a standard binary classification problem. Then this loss function encourages us to predict all labels correctly. As we will show in Proposition 4, the subset loss is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. If we consider the objective function (1) with the subset loss and the regularizer $r(\mathbf{w}) = \frac{\sigma}{2} \sum_{j \in [c]} \|\mathbf{w}_j\|_2^2$, then we can apply Theorem 2 to derive with probability $1 - \delta$ that $F(\mathbf{w}_S) - F(\mathbf{w}^*) = \widetilde{O}\left(\frac{\log(1/\delta)}{n\sigma}\right)$.

Example 5 (Learning with ranking loss)**.** Let $\mathcal{Y} = \{-1, +1\}^c$. For each $y \in \{-1, +1\}^c$ we denote $y_+ = \{j \in [c] : y^{(j)} = +1\}$ and $y_- = \{j \in [c] : y^{(j)} = -1\}$ as the set of relevant and irrelevant labels, respectively. Consider the following ranking loss (Zhang and Zhou 2013)

$$\ell(\mathbf{w}; z) = \frac{1}{|y_+||y_-|} \sum_{j_+ \in y_+} \sum_{j_- \in y_-} \tilde{\ell}\big(\langle \mathbf{w}_{j_+} - \mathbf{w}_{j_-}, \phi(x) \rangle\big), \quad (11)$$

where $|A|$ denotes the cardinality of a set A. Intuitively, the ranking loss encourages predictors with larger function values for a relevant label than a irrelevant label. As we will show in Proposition 4, the ranking loss is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. If we consider the objective function (1) with the ranking loss and the regularizer $r(\mathbf{w}) = \frac{\sigma}{2} \sum_{j \in [c]} \|\mathbf{w}_j\|_2^2$, then we can apply Theorem 2 to show with probability at least $1 - \delta$ that $F(\mathbf{w}_S) - F(\mathbf{w}^*) = \widetilde{O}\left(\frac{\log(1/\delta)}{n\sigma}\right)$.

Remark 5. Generalization bounds of the order $O(1/\sqrt{n})$ were established for MLC with the decomposable loss $\ell(\mathbf{w}; z) = \frac{1}{c} \sum_{j \in [c]} \big(y^{(j)} - \langle \mathbf{w}_j, x \rangle\big)^2$ under ERM in a constrained space $\{\mathbf{w} \in \mathbb{R}^{d \times c} : \|\mathbf{w}\|_{S_1} \leq \Lambda\}$ (Yu et al. 2014). Although this bound has no dependency on c, it requires a constraint in terms of $\|\cdot\|_{S_1}$, which can be as large as $\sqrt{c}\|\cdot\|_{2,2}$ (i.e., a constraint $\|\mathbf{w}\|_{S_1} \leq \Lambda$ corresponds to the constraint $\|\cdot\|_{2,2,} \leq \sqrt{c}\Lambda$). As a comparison, we consider a regularization scheme involving the much milder norm $\|\cdot\|_{2,2}$. Also, there is a gap between the theoretical analysis and the algorithm design: the generalization analysis there is for ERM, while the algorithm is designed based on a regularization scheme (Yu et al. 2014). Furthermore, our analysis for regularization implies a fast rate $O(1/n)$.

The proof of Proposition 4 is given in Appendix A.4.

Proposition 4. *1. If $\tilde{\ell}$ is L-Lipschitz continuous, then the subset loss* (10) *is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$.*

2. If $\tilde{\ell}$ is $(L/2)$-Lipschitz continuous, then the ranking loss (11) *is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$.*

Experimental Verification

In this section, we present experimental results to verify our theoretical analysis. We consider a specific vector-valued learning problem called multinomial logistic regression, where the aim is to predict a class label for each training example. We apply SGD (7) to solve the optimization problem with the objective function

$$F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \log \Big(\sum_{j \in [c]} \exp \big(\langle \mathbf{w}_j - \mathbf{w}_{y_i}, x_i \rangle\big)\Big) + \frac{\lambda}{2} \|\mathbf{w}\|_{2,2}^2, \quad (12)$$

where $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_c) \in \mathbb{R}^{d \times c}$ and $\lambda = 0.01$. We set the initial point $\mathbf{w} = 0$ and the step size $\eta_t = 1/(\lambda t + 1)$. We consider four real-world datasets available from the LIBSVM homepage (Chang and Lin 2011), whose information is summarized in Table B.2. We repeat experiments 50 times and report the average as well as standard deviation of the experimental results. We call F_S and F the training error and testing error, respectively. We consider two experiments.

For the first experiment, we aim to verify the generalization bound in (8). We randomly use 80% of data for training and reserve the remaining 20% for testing. We plot the testing errors $F(\mathbf{w}_t)$ of the SGD sequence versus the number of passes, i.e., t/n in Figure 1. It is clear that the testing error decreases initially as we run more SGD iterations. After sufficient number of iterations, the generalization performance of SGD iterates no long improves. This is well consistent to the generalization bound in (8) which shows that the generalization bound would be dominated by the effect due to sample size if $T \geq n$.

For the second experiment, our aim is to show whether the dependency of the generalization bound (9) on the sample size can be really captured in practice. For each dataset, we randomly select different number of examples for training and reserve the remaining for testing. For each fixed number of training examples, we train a model by solving (12) with SGD and stop it after 5 passes over the data. We compute
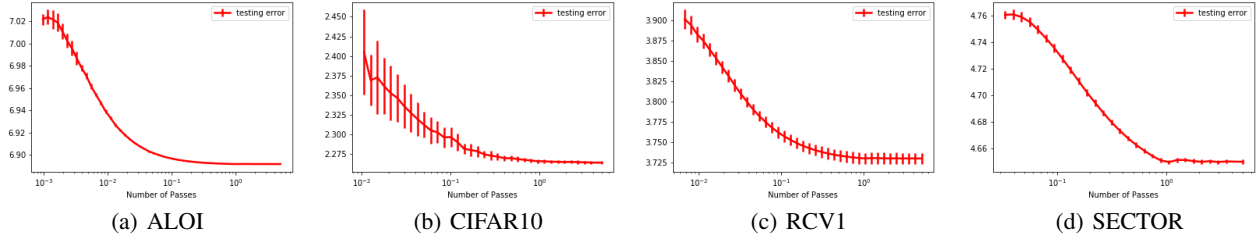
| (a) ALOI | (b) CIFAR10 | (c) RCV1 | (d) SECTOR |

Figure 1: Testing errors versus the number of passes.

| (a) ALOI | (b) CIFAR10 | (c) RCV1 | (d) SECTOR |
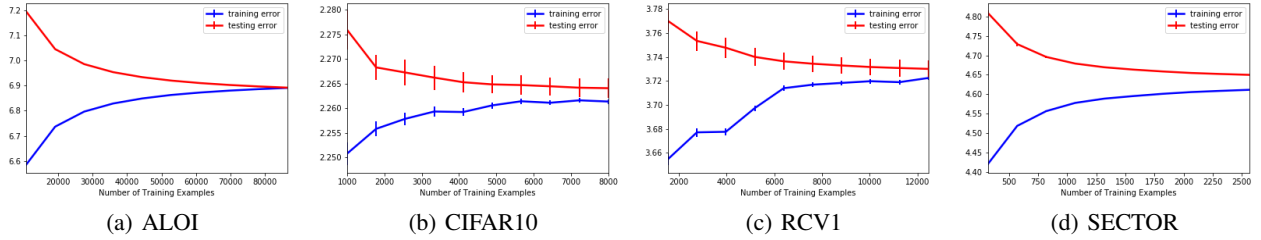
Figure 2: Training and testing errors versus the training data size.

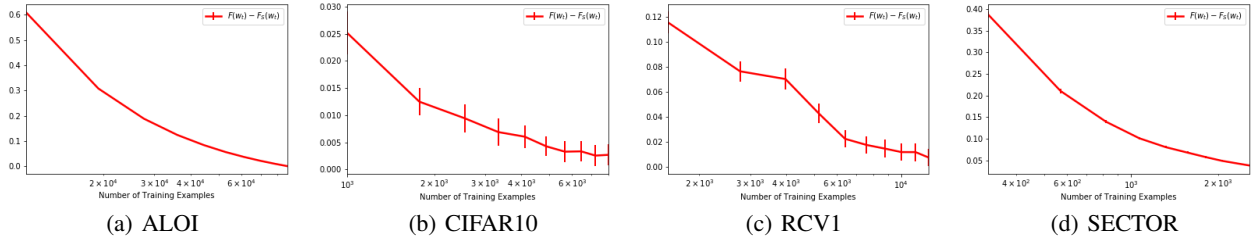

| (a) ALOI | (b) CIFAR10 | (c) RCV1 | (d) SECTOR |

Figure 3: $F(\mathbf{w}_T) - F_S(\mathbf{w}_T)$ versus the training data size.

both the training error $F_S(\mathbf{w}_T)$ and testing error $F(\mathbf{w}_T)$ of the output model $\mathbf{w}_T$ (last iterate). In this way we get a training error and a testing error for each considered sample size. We then plot the relative behavior of these errors versus the number of training examples in Figure 2. According to Figure 2, it is clear that training errors increase as the increase of training examples. This phenomenon is due to the increasing difficulty of the optimization problem as the increase of training examples. Note the fit of a model on $10,000$ examples would be significantly harder than that on 100 examples. It is also clear that the generalization behavior improves as we have more training examples. This matches well the generalization bound (9). In Figure 3 we further plot the difference between testing error and training error ($F(\mathbf{w}_T) - F_S(\mathbf{w}_T)$) versus the sample size. It is clear this difference is a decreasing function of the sample size.

Conclusion

We present a unifying generalization analysis for the regularization framework of learning with vector-valued functions. Our generalization bounds admit a mild dependency on the output dimension and decay fast w.r.t. the sample size. For instance for MLC, they improve the best known dependency on the number of classes from $O(\sqrt{c})$ to a $O(\log(c))$, making them suitable for extreme classification. Further-

more, our analysis relax the existing restrictive assumptions such as a low noise condition or a smoothness requirement on loss functions. We develop the first generalization analysis for SGD to learn vector-valued functions, which is important to understand both statistical properties of models and the convergence of the algorithm. We present applications to specific learning machines and conduct experiments to verify our theory.

It would be interesting to extend our discussion to a distributed learning setting (Lin and Zhou 2018; Hu, Wu, and Zhou 2020). It would be also interesting to study non-strongly convex regularizers such as ℓ_1 regularizers which is useful to learn sparse models (Guo et al. 2017).

References

Álvarez, M. A.; Rosasco, L.; and Lawrence, N. D. 2012. Kernels for Vector-Valued Functions: A Review. *Foundations and Trends® in Machine Learning* 4(3): 195–266.

Argyriou, A.; Pontil, M.; Ying, Y.; and Micchelli, C. 2007. A spectral regularization framework for multi-task structure learning. *Advances in neural information processing systems* 20: 25–32.

Bartlett, P.; Bousquet, O.; and Mendelson, S. 2005. Local

Rademacher complexities. *Annals of Statistics* 33(4): 1497–1537.

Bartlett, P.; and Mendelson, S. 2002. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research* 3: 463–482.

Bengio, S.; Dembczynski, K.; Joachims, T.; Kloft, M.; and Varma, M. 2019. Extreme classification (dagstuhl seminar 18291). In *Dagstuhl Reports*, volume 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3): 27.

Ciliberto, C.; Rudi, A.; Rosasco, L.; and Pontil, M. 2017. Consistent multitask learning with nonlinear output relations. In *Advances in Neural Information Processing Systems*, 1986–1996.

Cortes, C.; Kloft, M.; and Mohri, M. 2013. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems*, 2760–2768.

Cortes, C.; Kuznetsov, V.; Mohri, M.; and Yang, S. 2016. Structured Prediction Theory Based on Factor Graph Complexity. In *Advances in Neural Information Processing Systems*, 2514–2522.

Crammer, K.; and Singer, Y. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* 2: 265–292.

Dembczyński, K.; Waegeman, W.; Cheng, W.; and Hüllermeier, E. 2012. On label dependence and loss minimization in multi-label classification. *Machine Learning* 88(1-2): 5–45.

Deshmukh, A. A.; Lei, Y.; Sharma, S.; Dogan, U.; Cutler, J. W.; and Scott, C. 2019. A Generalization Error Bound for Multi-class Domain Generalization. *arXiv preprint arXiv:1905.10392* .

Foster, D. J.; and Rakhlin, A. 2019. ℓ_∞ Vector Contraction for Rademacher Complexity. *arXiv preprint arXiv:1911.06468* .

Gao, W.; and Zhou, Z.-H. 2011. On the Consistency of Multi-Label Learning. In *Conference on Learning Theory*, volume 19, 341–358.

Guermeur, Y. 2017. Lp-norm Sauer–Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences* 89: 450–473.

Guo, Z.-C.; Xiang, D.-H.; Guo, X.; and Zhou, D.-X. 2017. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications* 15(03): 433–455.

Harvey, N. J. A.; Liaw, C.; Plan, Y.; and Randhawa, S. 2019. Tight analyses for non-smooth stochastic gradient descent. In Beygelzimer, A.; and Hsu, D., eds., *Conference on Learning Theory*, 1579–1613.

Hu, T.; Wu, Q.; and Zhou, D.-X. 2020. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Applied and Computational Harmonic Analysis* 49(1): 229–256.

Jain, H.; Balasubramanian, V.; Chunduri, B.; and Varma, M. 2019. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 528–536.

Kakade, S. M.; Shalev-Shwartz, S.; and Tewari, A. 2012. Regularization techniques for learning with matrices. *Journal of Machine Learning Research* 13: 1865–1890.

Khandagale, S.; Xiao, H.; and Babbar, R. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning* 109(11): 2099–2119.

Kuznetsov, V.; Mohri, M.; and Syed, U. 2014. Multi-Class Deep Boosting. In *Advances in Neural Information Processing Systems*, 2501–2509.

Lapin, M.; Hein, M.; and Schiele, B. 2015. Top-k Multiclass SVM. In *Advances in Neural Information Processing Systems*, 325–333.

Lei, Y.; Binder, A.; Dogan, Ü.; and Kloft, M. 2015a. Theory and algorithms for the localized setting of learning kernels. In *Feature Extraction: Modern Questions and Challenges*, 173–195. PMLR.

Lei, Y.; Dogan, U.; Binder, A.; and Kloft, M. 2015b. Multiclass SVMs: From Tighter Data-Dependent Generalization Bounds to Novel Algorithms. In *Advances in Neural Information Processing Systems*, 2026–2034.

Lei, Y.; Dogan, Ü.; Zhou, D.-X.; and Kloft, M. 2019. Data-dependent Generalization Bounds for Multi-class Classification. *IEEE Transactions on Information Theory* 65(5): 2995–3021.

Li, J.; Liu, Y.; and Wang, W. 2019. Learning Vector-valued Functions with Local Rademacher Complexity. *arXiv preprint arXiv:1909.04883* .

Li, J.; Liu, Y.; Yin, R.; Zhang, H.; Ding, L.; and Wang, W. 2018. Multi-class learning: from theory to algorithm. In *Advances in Neural Information Processing Systems*, 1586–1595.

Lin, S.-B.; and Zhou, D.-X. 2018. Distributed kernel-based gradient descent algorithms. *Constructive Approximation* 47(2): 249–276.

Liu, C.; Zhao, P.; Huang, S.-J.; Jiang, Y.; and Zhou, Z.-H. 2018. Dual set multi-label learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, W.; Xu, D.; Tsang, I. W.; and Zhang, W. 2019. Metric Learning for Multi-Output Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 408–422.

Maurer, A. 2016. A vector-contraction inequality for Rademacher complexities. In *International Conference on Algorithmic Learning Theory*, 3–17.

Maurer, A.; and Pontil, M. 2016. Bounds for vector-valued function estimation. *arXiv preprint arXiv:1606.01487* .

Maximov, Y.; Amini, M.-R.; and Harchaoui, Z. 2018. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research* 61: 761–786.

Micchelli, C. A.; and Pontil, M. 2005. On learning vector-valued functions. *Neural computation* 17(1): 177–204.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of Machine Learning*. MIT press.

Musayeva, K.; Lauer, F.; and Guermeur, Y. 19. Rademacher complexity and generalization performance of multi-category margin classifiers. *Neurocomputing* 342: 6–15.

Musayeva, K.; Lauer, F.; and Guermeur, Y. 2018. A sharper bound on the Rademacher complexity of margin multi-category classifiers. In *ESANN*.

Reddi, S. J.; Kale, S.; Yu, F.; Holtmann-Rice, D.; Chen, J.; and Kumar, S. 2019. Stochastic negative mining for learning with large output spaces. In *International Conference on Artificial Intelligence and Statistics*, 1940–1949.

Reeve, H. W.; and Kaban, A. 2020. Optimistic bounds for multi-output prediction. *arXiv preprint arXiv:2002.09769* .

Shen, X.; Liu, W.; Tsang, I. W.; Sun, Q.-S.; and Ong, Y.-S. 2018. Compact multi-label learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sridharan, K.; Shalev-Shwartz, S.; and Srebro, N. 2009. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, 1545–1552.

Thrampoulidis, C.; Oymak, S.; and Soltanolkotabi, M. 2020. Theoretical Insights Into Multiclass Classification: A High-dimensional Asymptotic View. *Advances in Neural Information Processing Systems* 33.

Wu, G.; and Zhu, J. 2020. Multi-label classification: do Hamming loss and subset accuracy really conflict with each other? *Advances in Neural Information Processing Systems* 33.

Xu, C.; Liu, T.; Tao, D.; and Xu, C. 2016. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing* 25(3): 1495–1507.

Xu, D.; Shi, Y.; Tsang, I. W.; Ong, Y.-S.; Gong, C.; and Shen, X. 2019. Survey on multi-output learning. *IEEE Transactions on Neural Networks and Learning Systems* .

Xu, N.; Liu, Y.-P.; and Geng, X. 2020. Partial Multi-Label Learning with Label Distribution. In *AAAI*, 6510–6517.

Yousefi, N.; Lei, Y.; Kloft, M.; Mollaghasemi, M.; and Anagnostopoulos, G. 2018. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research* 19: 1385–1431.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. 2014. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, 593–601.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8): 1819–1837.

Zhang, T. 2004. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research* 5: 1225–1251.

Zhang, Y.; and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* .

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2012. Multi-instance multi-label learning. *Artificial Intelligence* 176(1): 2291–2320.

A Proofs

A.1 Proof of Theorem 1

The following lemma relates the Rademacher complexity of loss function classes to that of hypothesis spaces (Lei et al. 2019). The idea of considering a function space $\widetilde{\mathcal{H}}_\Lambda$ defined over an extended input argument is to fully exploit the Lipschitz continuity of loss functions w.r.t. the infinity norm, i.e.,

$$\max_z \left|\ell(\mathbf{w}; z) - \ell(\mathbf{w}'; z)\right| \leq L \max_x \max_j \left|h^{\mathbf{w}_j}(x) - h^{\mathbf{w}'_j}(x)\right|.$$

We define the *worst-case* Rademacher complexity as $\mathfrak{R}_n(\mathcal{H}) = \sup_{S \in \mathcal{Z}^n} \mathfrak{R}_S(\mathcal{H})$.

Lemma A.1 (Lei et al.,2019). *Let Assumption 1 hold. For any $\Lambda > 0$, define*

$$\widetilde{\mathcal{H}}_\Lambda = \left\{ (x, j) \mapsto h^{\mathbf{w}_j}(x) : x \in \mathcal{X}, j \in [c], \mathbf{w} \in \mathcal{W}_\Lambda \right\}. \tag{A.1}$$

Then there exists a constant C_1 independent of n, c, L and Λ such that

$$\mathfrak{R}_S(\mathcal{F}_\Lambda) \leq C_1 L \sqrt{c} \mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_\Lambda) \log^2(nc).$$

To apply the above lemma, we need to estimate $\mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_\Lambda)$. This is achieved in the following lemma. Note that the upper and lower bound matches up to a constant factor. Note $\widetilde{\mathcal{H}}_\Lambda$ takes the form $\widetilde{\mathcal{H}}_\Lambda = \left\{ (x, j) \mapsto \langle \mathbf{w}_j, x \rangle : \mathbf{w} \in \mathcal{W}_\Lambda \right\}$ in this paper. Recall that $\tilde{\phi}_j(x) := \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(x), \underbrace{0, \ldots, 0}_{c-j}\big) \in \mathcal{H}_K^c$.

Lemma A.2. *If Assumption 2 holds, then the worst-case Rademacher complexity of $\widetilde{\mathcal{H}}_\Lambda$ can be bounded by*

$$\sqrt{\frac{1}{2nc}} \sup_{\mathbf{w} \in \mathcal{W}_\Lambda} \left\langle \mathbf{w}, \tilde{\phi}_j(x) \right\rangle \leq \mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_\Lambda) \leq \sqrt{\frac{2\Lambda}{nc\sigma}} \widetilde{B}.$$

For a convex function f, we denote by f^* its Fenchel conjugate, i.e., $f^*(\mathbf{v}) := \sup_{\mathbf{w}}[\langle \mathbf{w}, \mathbf{v} \rangle - f(\mathbf{w})]$. The following lemma is due to Kakade, Shalev-Shwartz, and Tewari (2012).

Lemma A.3. *If τ is σ-strongly convex w.r.t. $\|\cdot\|$ and $\tau^*(0) = 0$, then for any sequence $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and any $\mathbf{w}$ we have*

$$\sum_{i=1}^n \langle \mathbf{w}, \mathbf{v}_i \rangle \leq \tau(\mathbf{w}) + \sum_{i=1}^n \langle \nabla \tau^*(\mathbf{v}_{1:i-1}), \mathbf{v}_i \rangle + \frac{1}{2\sigma} \sum_{i=1}^n \|\mathbf{v}_i\|_*^2,$$

where $\mathbf{v}_{1:i-1}$ denotes the sum $\sum_{j=1}^{i-1} \mathbf{v}_j$.

Proof of Lemma A.2. Define $\tau : \mathcal{W} \mapsto \mathbb{R}$ by $\tau(\mathbf{w}) = F(\mathbf{w}) - F(\mathbf{w}^*)$. It is clear that τ is σ-strongly convex w.r.t. $\|\cdot\|$. Furthermore, it follows from the definition of $\mathbf{w}^*$ that

$$\tau^*(0) = \sup_{\mathbf{v}} \left(\langle 0, \mathbf{v} \rangle - \tau(\mathbf{v}) \right) = -\inf_{\mathbf{v}} \tau(\mathbf{v}) = 0.$$

According to the definition of Rademacher complexity (Bartlett and Mendelson 2002) and $\widetilde{\mathcal{H}}_\Lambda$ in (A.1), we know

$$nc\mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_\Lambda) = \sup_{(x_i, j_i): i \in [n]} \mathbb{E}_\epsilon \sup_{\mathbf{w} \in \mathcal{W}_\Lambda} \sum_{i=1}^{nc} \epsilon_i \langle \mathbf{w}_{j_i}, \phi(x_i) \rangle = \sup_{(x_i, j_i): i \in [n]} \mathbb{E}_\epsilon \sup_{\mathbf{w} \in \mathcal{W}_\Lambda} \sum_{i=1}^{nc} \epsilon_i \langle \mathbf{w}, \tilde{\phi}_{j_i}(x_i) \rangle, \tag{A.2}$$

where we have used the definition of $\tilde{\phi}_{j_i}$ in the last identity. We apply Lemma A.3 with $\mathbf{v}_i = \lambda \epsilon_i \tilde{\phi}_{j_i}(x_i)$ and derive

$$\sum_{i=1}^{nc} \epsilon_i \langle \mathbf{w}, \lambda \tilde{\phi}_{j_i}(x_i) \rangle \leq \tau(\mathbf{w}) + \sum_{i=1}^{nc} \langle \nabla \tau^*(\mathbf{v}_{1:i-1}), \mathbf{v}_i \rangle + \frac{\lambda^2}{2\sigma} \sum_{i=1}^{nc} \|\tilde{\phi}_{j_i}(x_i)\|_*^2.$$

It is clear that

$$\mathbb{E}_\epsilon \langle \nabla \tau^*(\mathbf{v}_{1:i-1}), \mathbf{v}_i \rangle = \mathbb{E}_\epsilon \left\langle \nabla \tau^* \left(\lambda \sum_{k=1}^{i-1} \epsilon_k \tilde{\phi}_{j_k}(x_k) \right), \lambda \epsilon_i \tilde{\phi}_{j_i}(x_i) \right\rangle = 0$$

and therefore

$$\mathbb{E}_\epsilon \sup_{\mathbf{w} \in \mathcal{W}_\Lambda} \sum_{i=1}^{nc} \epsilon_i \langle \mathbf{w}, \tilde{\phi}_{j_i}(x_i) \rangle \leq \frac{\sup_{\mathbf{w} \in \mathcal{W}_\Lambda} \tau(\mathbf{w})}{\lambda} + \frac{\lambda}{2\sigma} \sum_{i=1}^{nc} \|\tilde{\phi}_{j_i}(x_i)\|_*^2.$$

We can combine the above inequality and (A.2) together and get

$$nc\mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_\Lambda) \leq \frac{\Lambda}{\lambda} + \frac{\lambda}{2\sigma} \sup_{(x_i,j_i):i\in[n]} \sum_{i=1}^{nc} \|\tilde{\phi}_{j_i}(x_i)\|_*^2 = \frac{\Lambda}{\lambda} + \frac{\lambda nc}{2\sigma}\widetilde{B}^2.$$

With the choice $\lambda = \left(\frac{2\Lambda\sigma}{nc\widetilde{B}^2}\right)^{\frac{1}{2}}$, we derive

$$nc\mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_\Lambda) \leq \sqrt{2\Lambda nc\sigma^{-1}}\widetilde{B}.$$

This establishes the upper bounds. We now turn to the lower bounds. By taking $x_1 = \cdots = x_n = x, j_1 = \cdots = j_n = j$, we know ($\mathcal{W}_\Lambda$ is symmetric)

$$\sup_{\mathbf{w}\in\mathcal{W}_\Lambda} \left\langle \mathbf{w}, \sum_{i=1}^{nc} \epsilon_i \tilde{\phi}_{j_i}(x_i)\right\rangle = \sup_{\mathbf{w}\in\mathcal{W}_\Lambda} \left\langle \mathbf{w}, \sum_{i=1}^{nc} \epsilon_i \tilde{\phi}_j(x)\right\rangle = \left|\sum_{i=1}^{nc} \epsilon_i\right| \sup_{\mathbf{w}\in\mathcal{W}_\Lambda} \left\langle \mathbf{w}, \tilde{\phi}_j(x)\right\rangle.$$

It then follows from (A.2) and Khitchine's inequality $\mathbb{E}_\epsilon|\sum_{i=1}^n \epsilon_i| \geq \sqrt{n/2}$ (Mohri, Rostamizadeh, and Talwalkar 2012) that

$$\sup_{(x_i,j_i):i\in[n]} \mathbb{E}_\epsilon \sup_{\mathbf{w}\in\mathcal{W}_\Lambda} \sum_{i=1}^{nc} \epsilon_i \langle \mathbf{w}_{j_i}, \phi(x_i)\rangle \geq \sup_{(x,j)} \mathbb{E}_\epsilon \left|\sum_{i=1}^{nc} \epsilon_i\right| \sup_{\mathbf{w}\in\mathcal{W}_\Lambda} \left\langle \mathbf{w}, \tilde{\phi}_j(x)\right\rangle$$

$$\geq \sqrt{nc/2} \sup_{(x,j)} \sup_{\mathbf{w}\in\mathcal{W}_\Lambda} \left\langle \mathbf{w}, \tilde{\phi}_j(x)\right\rangle.$$

This establishes the stated lower bound and finishes the proof. $\qquad\square$

Proof of Theorem 1. Theorem follows directly by combining Lemma A.1 and Lemma A.2 together. $\qquad\square$

A.2   Proof of Theorem 2

To prove Theorem 2, we first introduce some necessary notations. Let $\rho_0 > 0$ and $\delta_0 \in (0,1)$ be two numbers to be fixed. We construct two sequences $\rho_k = 2^k\rho_0, \delta_k = 2^{-k}\delta_0$ for $k \in \mathbb{N}$. For brevity we set $\rho_{-1} = 0$. We decompose $\mathcal{W}$ into a sequence of disjoint sets according to the function value

$$\mathcal{W}_k' = \left\{\mathbf{w} \in \mathcal{W} : \rho_{k-1} < F(\mathbf{w}) - F(\mathbf{w}^*) \leq \rho_k\right\}.$$

For brevity, for any $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ we denote $\mathbb{E}[f(\mathbf{w};\cdot)]$ the expectation $\mathbb{E}_z[f(\mathbf{w};z)]$ and $\hat{\mathbb{E}}_S[f(\mathbf{w};\cdot)] = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w};z_i)$ the empirical average. For any $\mathbf{w}$, we define $s(\mathbf{w},\cdot)$ the shifted loss relative to $\mathbf{w}^*$

$$s(\mathbf{w};z) = \ell(\mathbf{w};z) - \ell(\mathbf{w}^*;z).$$

Our proof of Theorem 2 requires to use a concentration inequality called McDiarmid's inequality (Mohri, Rostamizadeh, and Talwalkar 2012).

Lemma A.4 (McDiarmid's inequality). *Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $f : \mathcal{Z}^n \mapsto \mathbb{R}$ satisfies*

$$\sup_{z_1,\ldots,z_n,\bar{z}_i\in\mathcal{Z}} |f(z_1,\cdots,z_n) - f(z_1,\cdots,z_{i-1},\bar{z}_i,z_{i+1},\cdots,z_n)| \leq c_i \qquad (A.3)$$

for $1 \leq i \leq n$. Then, for any $0 < \delta < 1$, with probability of at least $1 - \delta$, we have

$$f(Z_1,\ldots,Z_n) \leq \mathbb{E}f(Z_1,\ldots,Z_n) + \sqrt{\frac{\sum_{i=1}^n c_i^2 \log(1/\delta)}{2}}.$$

Proof of Theorem 2. By the σ-strong convexity of F, we know $F(\mathbf{w}) - F(\mathbf{w}^*) \geq \frac{\sigma}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$, from which we derive the following inequality for any $\mathbf{w} \in \mathcal{W}_k'$

$$\|\mathbf{w} - \mathbf{w}^*\| \leq \left(2\sigma^{-1}\big(F(\mathbf{w}) - F(\mathbf{w}^*)\big)\right)^{\frac{1}{2}} \leq (2\sigma^{-1}\rho_k)^{\frac{1}{2}}.$$

For any $\mathbf{w} \in \mathcal{W}_k'$ and $z = (x,y)$ we know

$$|s(\mathbf{w};z)| = |\ell(\mathbf{w};z) - \ell(\mathbf{w}^*;z)| \leq L\|h^{\mathbf{w}}(x) - h^{\mathbf{w}^*}(x)\|_\infty$$

$$= L\left\|\left(\langle\mathbf{w}_1 - \mathbf{w}_1^*, \phi(x)\rangle, \ldots, \langle\mathbf{w}_c - \mathbf{w}_c^*, \phi(x)\rangle\right)\right\|_\infty = L \max_{j=1,\ldots,c} |\langle\mathbf{w}_j - \mathbf{w}_j^*, \phi(x)\rangle|$$

$$\leq L \max_{j=1,\ldots,c} \|\mathbf{w}_j - \mathbf{w}_j^*\|_2 \|\phi(x)\|_2 \leq L\kappa\|\mathbf{w} - \mathbf{w}^*\|_{2,\infty} \leq L\kappa(2\rho_k/\sigma)^{\frac{1}{2}},$$

where we have used the mild assumption $\|\cdot\|_{2,\infty} \leq \|\cdot\|$.

For any $S = \{z_1, \ldots, z_{k-1}, z_k, z_{k+1}, \ldots, z_n\}$ and $S' = \{z_1, \ldots, z_{k-1}, z'_k, z_{k+1}, \ldots, z_n\}$, we have

$$\left| \sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big) - \sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_{S'}[s(\mathbf{w}; \cdot)] \big) \right|$$

$$\leq \sup_{\mathbf{w} \in \mathcal{W}'_k} \left| \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_{S'}[s(\mathbf{w}; \cdot)] \right| = \frac{1}{n} \sup_{\mathbf{w} \in \mathcal{W}'_k} \left| s(\mathbf{w}; z_k) - s(\mathbf{w}; z'_k) \right|$$

$$\leq \frac{2}{n} \sup_{\mathbf{w} \in \mathcal{W}'_k} \sup_{z \in \mathcal{Z}} |s(\mathbf{w}; z)| \leq \frac{2L\kappa}{n} \left(2\rho_k / \sigma \right)^{\frac{1}{2}}.$$

By the McDiarmid's inequality with increments bounded by $\frac{2L\kappa}{n}\left(2\rho_k/\sigma\right)^{\frac{1}{2}}$, we get the following inequality with probability at least $1 - \delta_k$

$$\sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big) \leq \mathbb{E}_S \sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big) + 2L\kappa \left(\frac{\rho_k \log(1/\delta_k)}{n\sigma} \right)^{\frac{1}{2}}. \tag{A.4}$$

It follows from the symmetry trick that ($\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ is drawn i.i.d. and independently of S)

$$\mathbb{E}_S \sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big) = \mathbb{E}_S \sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}_{\widetilde{S}} \hat{\mathbb{E}}_{\widetilde{S}}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big)$$

$$\leq \mathbb{E}_{S,\widetilde{S}} \sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\hat{\mathbb{E}}_{\widetilde{S}}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big) = \frac{1}{n} \mathbb{E}_{S,\widetilde{S},\epsilon} \sup_{\mathbf{w} \in \mathcal{W}'_k} \left[\sum_{i=1}^{n} \epsilon_i \big(s(\mathbf{w}; z_i) - s(\mathbf{w}; \tilde{z}_i) \big) \right]$$

$$\leq \frac{2}{n} \mathbb{E}_{S,\epsilon} \sup_{\mathbf{w} \in \mathcal{W}'_k} \left[\sum_{i=1}^{n} \epsilon_i s(\mathbf{w}; z_i) \right] = \frac{2}{n} \mathbb{E}_{S,\epsilon} \sup_{\mathbf{w} \in \mathcal{W}'_k} \sum_{i=1}^{n} \epsilon_i \ell(\mathbf{w}; z_i) = 2\mathbb{E}_S \mathfrak{R}_S(\mathcal{F}_{\rho_k}), \tag{A.5}$$

where in the last second step we remove $\ell(\mathbf{w}^*; z_i)$ since $\mathbf{w}^*$ is fixed. According to Lemma A.1 and Lemma A.2, we know

$$\mathbb{E}_S \mathfrak{R}_S(\mathcal{F}_{\rho_k}) \leq C_1 L \sqrt{c} \mathfrak{R}_{nc}(\widetilde{\mathcal{H}}_{\rho_k}) \log^2(nc) \leq \frac{\sqrt{2\rho_k} C_1 L \log^2(nc) \widetilde{B}}{\sqrt{n\sigma}}.$$

Combining (A.4), (A.5) and the above inequality together, we derive the following inequality with probability at least $1 - \delta_k$

$$\sup_{\mathbf{w} \in \mathcal{W}'_k} \big(\mathbb{E}[s(\mathbf{w}; \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}; \cdot)] \big) \leq \frac{2\sqrt{2\rho_k} C_1 L \log^2(nc) \widetilde{B}}{\sqrt{n\sigma}} + 2L\kappa \left(\frac{\rho_k \log(1/\delta_k)}{n\sigma} \right)^{\frac{1}{2}}. \tag{A.6}$$

For any $\mathbf{w} \in \mathcal{W}'_k$, we have

$$\frac{\rho_k}{2} = \rho_{k-1} \leq F(\mathbf{w}) - F(\mathbf{w}^*)$$

and therefore

$$\rho_k \leq \max \big\{ \rho_0, 2(F(\mathbf{w}) - F(\mathbf{w}^*)) \big\}.$$

Furthermore, it is clear

$$\frac{1}{\delta_k} = \frac{\rho_0 2^k}{\delta_0 \rho_0} \leq \frac{\max \big\{ \rho_0, 2(F(\mathbf{w}) - F(\mathbf{w}^*)) \big\}}{\delta_0 \rho_0}.$$

By the definition of F, F_S and $s(\mathbf{w})$, we know

$$\big(F(\mathbf{w}) - F(\mathbf{w}^*) \big) - \big(F_S(\mathbf{w}) - F_S(\mathbf{w}^*) \big) = \mathbb{E}[s(\mathbf{w}, \cdot)] - \hat{\mathbb{E}}_S[s(\mathbf{w}, \cdot)].$$

Combining the above discussions and (A.6) together, with probability $1 - \delta_k$ the following inequality holds uniformly for all $\mathbf{w} \in \mathcal{W}'_k$

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \big(F_S(\mathbf{w}) - F_S(\mathbf{w}^*) \big) + 2L \sqrt{\frac{\max \big\{ \rho_0, 2(F(\mathbf{w}) - F(\mathbf{w}^*)) \big\}}{n\sigma}} \times$$

$$\left(\sqrt{2} C_1 \log^2(nc) \widetilde{B} + \kappa \sqrt{\log(1/\delta_0) + \log \max \big\{ 1, 2\rho_0^{-1}(F(\mathbf{w}) - F(\mathbf{w}^*)) \big\}} \right). \tag{A.7}$$

Noticing $\sum_{k=0}^{\infty} \delta_k = 2\delta_0$, the above inequality holds with probability at least $1 - 2\delta_0$ uniformly for all $\mathbf{w} \in \mathcal{W}$. The stated inequality (5) then follows by setting $\rho_0 = 1/(n\sigma)$.

We now prove (6). We prove instead

$$F(\mathbf{w}_S) - F(\mathbf{w}^*) \leq \frac{16L^2}{n\sigma}\Big(2C_1^2 \log^3(nc)\widetilde{B}^2 + 4\kappa^2 \log(1/\delta_0)\Big). \tag{A.8}$$

Since $F_S(\mathbf{w}_S) \leq F_S(\mathbf{w}^*)$, it follows from (A.7) that the following inequality holds with probability $1 - 2\delta_0$

$$F(\mathbf{w}_S) - F(\mathbf{w}^*) \leq 2L\sqrt{\frac{\max\{\rho_0, 2(F(\mathbf{w}_S) - F(\mathbf{w}^*))\}}{n\sigma}} \times$$
$$\Big(\sqrt{2}C_1 \log^2(nc)\widetilde{B} + \kappa\sqrt{\log(1/\delta_0) + \log \max\{1, 2\rho_0^{-1}(F(\mathbf{w}_S) - F(\mathbf{w}^*))\}}\Big). \tag{A.9}$$

Note that the above inequality holds for any $\rho_0 > 0$. We can take $\rho_0 = 64L^2\kappa^2/(n\sigma)$. We now consider two cases. For the case $F(\mathbf{w}_S) - F(\mathbf{w}^*)) < \rho_0/2$, the inequality (A.8) is trivial. We now consider the case $F(\mathbf{w}_S) - F(\mathbf{w}^*)) \geq \rho_0/2$. In this case, it follows from (A.9) and the elementary inequality $(a + b)^2 \leq 2(a^2 + b^2)$ that

$$F(\mathbf{w}_S) - F(\mathbf{w}^*) \leq \frac{16L^2}{n\sigma}\Big(2C_1^2 \log^3(nc)\widetilde{B}^2 + \kappa^2\Big(\log(1/\delta_0) + \log\big(2\rho_0^{-1}(F(\mathbf{w}_S) - F(\mathbf{w}^*)))\big)\Big)\Big).$$

By the elementary inequality $\log a \leq ab + \log(1/b) - 1$ for all $a, b > 0$, we know

$$\log\big(2\rho_0^{-1}(F(\mathbf{w}_S) - F(\mathbf{w}^*))\big) \leq \frac{n\sigma}{32L^2\kappa^2}(F(\mathbf{w}_S) - F(\mathbf{w}^*)) + \log(64L^2\kappa^2/(\rho_0 n\sigma)) - 1.$$

It then follows that

$$F(\mathbf{w}_S) - F(\mathbf{w}^*) \leq \frac{16L^2}{n\sigma}\Big(2C_1^2 \log^3(nc)\widetilde{B}^2 + \kappa^2 \log(1/\delta_0)\Big) + \frac{F(\mathbf{w}_S) - F(\mathbf{w}^*)}{2} + \frac{16L^2\kappa^2 \log(64L^2\kappa^2/(\rho_0 n\sigma))}{n\sigma}.$$

Since $\rho_0 = 64L^2\kappa^2/(n\sigma)$ we then get the stated inequality (A.8). The proof is complete. $\qquad\square$

A.3 Proof of Theorem 3

To prove Theorem 3, we first introduce a high-probability bound on the convergence rate of SGD.

Lemma A.5 (Harvey et al.,2019). *Suppose $F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z)$ is σ-strongly convex w.r.t. $\|\cdot\|_{2,2}$. Assume for all $\mathbf{w} \in \mathcal{W}$ and $z \in \mathcal{Z}$ there holds $\|f'(\mathbf{w}; z)\|_{2,2} \leq \widetilde{L}$ for some $\widetilde{L} > 0$. Let $\{\mathbf{w}_t\}$ be obtained by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t f'(\mathbf{w}_t; z_{i_t})$, where $\eta_t = 1/(\sigma t)$ and $\{i_t\}_{t\in\mathbb{N}}$ is independently drawn from the uniform distribution over $[n]$. Then with probability at least $1 - \delta$ we have*

$$F_S(\mathbf{w}_T) - \inf_{\mathbf{w}} F_S(\mathbf{w}) = O\Big(\frac{\log(T)\log(1/\delta)}{T\sigma}\Big).$$

Proof of Theorem 3. For any $j \in [c]$ denote $\ell^{(j)}(\mathbf{w}; z) = \frac{\partial \ell(\mathbf{w};z)}{\partial \mathbf{w}_j}$ the partial (sub)gradient w.r.t. the $\mathbf{w}_j$. Then $\ell^{(j)}(\mathbf{w}; z) = \psi^{(j)}(h^{\mathbf{w}}(x), y)\phi(x)$, where $\psi^{(j)}$ denotes the partial (sub)gradient w.r.t. the j-th coordinate. Assumption 1 means that $\big\|\big(\psi^{(j)}(h^{\mathbf{w}}(x), y)\big)_{j=1}^{c}\big\|_1 \leq L$ for all x, y. It then follows that

$$\|\ell'(\mathbf{w}; z)\|_{2,2}^2 = \sum_{j=1}^{c} \big\|\ell^{(j)}(\mathbf{w}; z)\big\|_2^2 = \sum_{j=1}^{c} \big|\psi^{(j)}(h^{\mathbf{w}}(x), y)\big|^2 \|\phi(x)\|_2^2$$
$$\leq \kappa^2 \sum_{j=1}^{c} \big|\psi^{(j)}(h^{\mathbf{w}}(x), y)\big|^2 \leq \kappa^2\Big(\sum_{j=1}^{c} \big|\psi^{(j)}(h^{\mathbf{w}}(x), y)\big|\Big)^2 \leq L^2\kappa^2. \tag{A.10}$$

According to the SGD update, we know

$$\big\|\mathbf{w}_{t+1}\big\|_{2,2} = \Big\|\mathbf{w}_t - \eta_t\big(\ell'(\mathbf{w}_t; z_{i_t}) + \sigma\mathbf{w}_t\big)\Big\|_{2,2}$$
$$\leq \big(1 - \eta_t\sigma\big)\|\mathbf{w}_t\|_{2,2} + \eta_t\|\ell'(\mathbf{w}_t; z_{i_t})\|_{2,2}$$
$$\leq \big(1 - \eta_t\sigma\big)\|\mathbf{w}_t\|_{2,2} + \eta_t\sigma(L\kappa/\sigma)$$
$$\leq \max\{\|\mathbf{w}_t\|_{2,2}, L\kappa/\sigma\},$$

where we have used (A.10) and $\eta_t \leq 1/\sigma$. Applying this inequality recursively, we know by induction that

$$\|\mathbf{w}_{t+1}\|_{2,2} \leq \big\{\|\mathbf{w}_1\|_{2,2}, \ldots, \|\mathbf{w}_t\|_{2,2}, L\kappa/\sigma\big\} = L\kappa/\sigma.$$

This together with (A.10) shows that

$$\|f'(\mathbf{w}_t; z)\|_{2,2} \le \|\ell'(\mathbf{w}_t; z)\|_{2,2} + \sigma\|\mathbf{w}_t\|_{2,2} \le 2L\kappa.$$

Note $r(\mathbf{w})$ is σ-strongly convex w.r.t. $\|\cdot\|_{2,2}$. Therefore, assumptions in Lemma A.5 hold. We can apply Lemma A.5 to show the following inequality with probability at least $1 - \delta/2$

$$F_S(\mathbf{w}_T) - \inf_{\mathbf{w}} F_S(\mathbf{w}) = O\Big(\frac{\log(T)\log(1/\delta)}{T\sigma}\Big). \tag{A.11}$$

According to (5), with probability at least $1 - \delta$ there holds

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) = \max\{F_S(\mathbf{w}_T) - F_S(\mathbf{w}^*), 1/(n\sigma)\} + \widetilde{O}\Big(\Big(\frac{(F(\mathbf{w}_T) - F(\mathbf{w}^*))(\log(1/\delta) + \widetilde{B})}{n\sigma}\Big)^{\frac{1}{2}}\Big)$$

$$= O\Big(\frac{\log(T)\log(1/\delta)}{\min\{T,n\}\sigma}\Big) + \widetilde{O}\Big(\Big(\frac{(F(\mathbf{w}_T) - F(\mathbf{w}^*))(\log(1/\delta) + \widetilde{B})}{n\sigma}\Big)^{\frac{1}{2}}\Big)$$

where we have used $F_S(\mathbf{w}^*) \ge \inf_{\mathbf{w}} F_S(\mathbf{w})$ and (A.11) (It can be shown $\widetilde{B} \le \kappa$). Solving the above quadratic inequality of $F(\mathbf{w}_T) - F(\mathbf{w}^*)$ yields the stated bound (8) with probability at least $1 - \delta$. The proof is complete. $\qquad\square$

A.4 Proof of Proposition 4

Proof of Proposition 4. We first prove the first part. For any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, by the elementary inequality

$$\big|\max\{a_1, \ldots, a_c\} - \max\{b_1, \ldots, b_c\}\big| \le \max\{|a_1 - b_1|, \ldots, |a_c - b_c|\}$$

we have

$$\big|\ell(\mathbf{w}; z) - \ell(\mathbf{w}'; z)\big| = \Big|\max_{j\in[c]} \tilde{\ell}\big(y^{(j)}\langle\mathbf{w}_j, \phi(x)\rangle\big) - \max_{j\in[c]} \tilde{\ell}\big(y^{(j)}\langle\mathbf{w}'_j, \phi(x)\rangle\big)\Big|$$

$$\le \max_{j\in[c]} \big|\tilde{\ell}\big(y^{(j)}\langle\mathbf{w}_j, \phi(x)\rangle\big) - \tilde{\ell}\big(y^{(j)}\langle\mathbf{w}'_j, \phi(x)\rangle\big)\big|$$

$$\le L\max_{j\in[c]} \big|\langle\mathbf{w} - \mathbf{w}'_j, \phi(x)\rangle\big| = L\big\|h^{\mathbf{w}}(x) - h^{\mathbf{w}'}(x)\big\|_{\infty}.$$

This proves the first part.

We now turn to the second part. For any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$

$$\big|\ell(\mathbf{w}; z) - \ell(\mathbf{w}'; z)\big| = \frac{1}{|y_+||y_-|}\Big|\sum_{j_+\in y_+}\sum_{j_-\in y_-}\Big(\tilde{\ell}\big(\langle\mathbf{w}_{j_+} - \mathbf{w}_{j_-}, \phi(x)\rangle\big) - \tilde{\ell}\big(\langle\mathbf{w}'_{j_+} - \mathbf{w}'_{j_-}, \phi(x)\rangle\big)\Big)\Big|$$

$$\le \max_{j_+\in y_+, j_-\in y_-} \Big|\tilde{\ell}\big(\langle\mathbf{w}_{j_+} - \mathbf{w}_{j_-}, \phi(x)\rangle\big) - \tilde{\ell}\big(\mathbf{w}'_{j_+} - \mathbf{w}'_{j_-}, \phi(x)\rangle\big)\Big|$$

$$\le \frac{L}{2}\max_{j_+\in y_+, j_-\in y_-} \Big|\langle\mathbf{w}_{j_+} - \mathbf{w}_{j_-}, \phi(x)\rangle - \langle\mathbf{w}'_{j_+} - \mathbf{w}'_{j_-}, \phi(x)\rangle\Big|$$

$$\le \frac{L}{2}\Big(\max_{j_+\in y_+} \big|\langle\mathbf{w}_{j_+} - \mathbf{w}'_{j_+}, \phi(x)\rangle\big| + \max_{j_-\in y_-} \big|\langle\mathbf{w}_{j_-} - \mathbf{w}'_{j_-}, \phi(x)\rangle\big|\Big)$$

$$\le L\|h^{\mathbf{w}}(x) - h^{\mathbf{w}'}(x)\|_{\infty}.$$

This proves the second part and finishes the proof. $\qquad\square$

B Some Tables

Table B.1: Generalization bounds for MLC in a Frobenius learning framework (either constraint $\|W\|_F \leq 1$ or regularizer $\|W\|_F^2$)

Bound	Lipschitz Continuity	Additional Assumption	Method	Reference
$\sqrt{c/n}$	ℓ_1-norm	sub-gaussian distribution	ERM	Yu et al. (2014)
$\sqrt{c/n}$	ℓ_2-norm	No	ERM	Liu et al. (2018); Wu and Zhu (2020)
$\sqrt{c/n}$	ℓ_2-norm	decay of singular values	ERM	Li, Liu, and Wang (2019)
$\log^3(nc)/n$	ℓ_∞-norm	strong convexity	RRM/SGD	This work

Table B.2: Description of the datasets. c is the number of classes, n is the sample size and d is the input dimension.

Dataset	c	n	d
CIFAR10	10	$10,000$	$3,072$
RCV1	53	$15,564$	$47,236$
SECTOR	105	$3,207$	$55,197$
ALOI	$1,000$	$108,000$	128