Data-dependent Generalization Bounds for Multi-class Classification

Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft

***Abstract*—In this paper, we study *data-dependent* generalization error bounds that exhibit a mild dependency on the number of classes, making them suitable for multi-class learning with a large number of label classes. The bounds generally hold for empirical multi-class risk minimization algorithms using an arbitrary norm as the regularizer. Key to our analysis are new structural results for multi-class Gaussian complexities and empirical ℓ_∞-norm covering numbers, which exploit the Lipschitz continuity of the loss function with respect to the ℓ_2- and ℓ_∞-norm, respectively. We establish data-dependent error bounds in terms of the complexities of a linear function class defined on a finite set induced by training examples, for which we show tight lower and upper bounds. We apply the results to several prominent multi-class learning machines and show a tighter dependency on the number of classes than the state of the art. For instance, for the multi-class SVM of Crammer and Singer (2002), we obtain a data-dependent bound with a logarithmic dependency, which is a significant improvement of the previous square-root dependency. Experimental results are reported to verify the effectiveness of our theoretical findings.**

***Index Terms*—Multi-class classification, Generalization error bounds, Covering numbers, Rademacher complexities, Gaussian complexities.**

I. INTRODUCTION

Multi-class learning is a classic problem in machine learning [1]. The outputs here stem from a finite set of categories (*classes*), and the aim is to classify each input into one of several possible target classes [2–4]. Classic applications of multi-class classification include handwritten optical character recognition, where the system learns to automatically interpret handwritten characters [5], part-of-speech tagging, where each word in a text is annotated with part-of-speech tag [6], and image categorization, where predefined categories are associated with digital images [7, 8].

Y. Lei is with Shenzhen Key Laboratory of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China (e-mail: leiyw@sustc.edu.cn). He was also with Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China.

Ü. Dogan is with Microsoft, 1020 enterprise way, Sunnyvale, CA 94089, USA (e-mail: udogan@microsoft.com).

D.-X. Zhou is with School of Data Science and Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China (e-mail: mazhou@cityu.edu.hk).

M. Kloft is with Department of Computer Science, TU Kaiserslautern, Kaiserslautern, Germany (e-mail: kloft@cs.uni-kl.de). He is also with Department of Computer Science, University of Southern California, Los Angeles, USA.

This paper was presented in part at Advances in Neural Information Processing Systems 28 (2015), 2035–2043.

Providing a theoretical framework of multi-class learning algorithms is a fundamental task in statistical learning theory [1]. Statistical learning theory aims to ensure formal guarantees to safeguard the performance of learning algorithms, often in the form of generalization error bounds [9]. Such bounds may lead to improved understanding of commonly used empirical practices and spur the development of novel learning algorithms ("Nothing is more practical than a good theory" [1]).

Classic generalization bounds for multi-class learning scale rather unfavorably (e.g., quadratic, linear, or square root at best) with the number of classes [9–11]. This may be because the standard theory has been constructed without the need of having a large number of label classes in mind as many classic multi-class learning problems consist of only a small number of classes. For instance, the historically first multi-class dataset—`Iris`—[12]—contains only three classes, the MNIST dataset [13] consists of 10 classes, and most of the datasets in the popular UCI corpus [14] contain up to several dozen classes.

However, with the advent of the big data era, multi-class learning problems—such as text or image classification [7, 15]—can involve tens or hundreds of thousands of classes. Recently, a subarea of machine learning that studies classification problems involving an extremely large number of classes (such as those mentioned above) called *eXtreme Classification* (XC) has emerged [16]. Several algorithms have recently been proposed to speed up the training or improve the prediction accuracy in classification problems with many classes [15, 17–26].

However, a discrepancy remains between *algorithms* and *theory* in classification with many classes, as standard statistical learning theory is void in the large number of classes scenario [27]. With the present paper we want to contribute toward a *better theoretical understanding* of multi-class classification with many classes. This theoretical understanding can provide grounds for the commonly used empirical practices in classification with many classes and lead to insights that may be used to guide the design of new learning algorithms.

Note that the present paper focuses on *multi-class* learning. Recently, there has been a growing interest in *multi-label* learning. The difference in the two scenarios is that each instance is associated with exactly one label class (in the multi-class case) or multiple classes (in the multi-label case), respectively. While the present analysis is tailored to the multi-class learning scenario, it may serve as a starting point for subsequent analysis of the multi-label learning scenario.

A. Summary of Contributions

We build the present journal article upon our previous conference paper published at NIPS 2015 [28], where we propose a multi-class support vector machine (MC-SVM) using block $\ell_{2,p}$-norm regularization, for which we proved data-dependent generalization bounds based on Gaussian complexities (GCs).

While the previous analysis employed margin-based loss, in the present article, we generalize GC-based data-dependent analysis to general loss functions that are Lipschitz continuous with respect to (w.r.t.) a variant of the ℓ_2-norm . Furthermore, we develop a new approach to derive data-dependent bounds based on empirical covering numbers (CNs) to capture the Lipschitz continuity of loss functions w.r.t. the ℓ_∞-norm with a moderate Lipschitz constant, which is *not* studied in the conference version of this article. For both approaches, our data-dependent error bounds can be stated in terms of the complexities of a linear function class defined on only a finite set induced by training examples, for which we give lower and upper bounds matching up to a constant factor. We present examples to show that each of these two approaches has its advantages and may outperform the other by inducing tighter error bounds for specific MC-SVMs.

As applications of our theory, we show error bounds for several prominent multi-class learning algorithms: multinomial logistic regression [29], top-k MC-SVM [30], ℓ_p-norm MC-SVM [28], and several classic MC-SVMs [31–33]. For all these methods, we show error bounds with an improved dependency on the number of classes over the state-of-the-art methods. For instance, the best known bounds for multinomial logistic regression and the MC-SVM by Crammer and Singer [31] scale as the square root of the number of classes. We improve this dependency to be *logarithmic*, which gives strong theoretical grounds for using these methods in classification with many classes.

We develop a novel algorithm to train the ℓ_p-norm MC-SVM [28] and report the experimental results to verify our theoretical findings and their applicability to model selection.

II. Related Work and Contributions

In this section, we discuss related work and outline the main contributions of this paper.

A. Related Work

In this subsection, we recapitulate the state of the art in multi-class learning theory.

1) Related Work on Data-dependent Bounds: The existing error bounds for multi-class learning can be classified into two groups: *data-dependent* and *data-independent* error bounds. Both types of bounds are often based on the assumption that the data are realized from independent and identically distributed random variables. However, this assumption can be relaxed to weakly dependent time series, for which Mohri and Rostamizadeh [34] and Steinwart et al. [35] show data-dependent and data-independent generalization bounds, respectively.

Data-dependent generalization error bounds refer to bounds that can be evaluated on training samples and thus can capture properties of the distribution that has generated the data [9]. Often, these bounds are built on the empirical Rademacher complexity (RC) [36–38], which can be used in model selection and for the construction of new learning algorithms [39].

The investigation of data-dependent error bounds for multi-class learning is initiated, to the best of our knowledge, by Koltchinskii and Panchenko [10], who give the following structural result on RCs: given a set $H = \{h = (h_1, \ldots, h_c)\}$ of vector-valued functions and training examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, it holds

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h\in H} \sum_{i=1}^{n} \epsilon_i \max\big\{h_1(\mathbf{x}_i), \ldots, h_c(\mathbf{x}_i)\big\} \leq \sum_{j=1}^{c} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h\in H} \sum_{i=1}^{n} \epsilon_i h_j(\mathbf{x}_i). \quad (1)$$

Here, $\epsilon_1, \ldots, \epsilon_n$ denote independent Rademacher variables (i.e., taking values $+1$ or -1, with equal probability), and $\mathbb{E}_{\boldsymbol{\epsilon}}$ denotes the conditional expectation operator removing the randomness coming from the variables $\epsilon_1, \ldots, \epsilon_n$.

In much of the subsequent theoretical work on multi-class learning, the above result is used as a starting point, by which the maximum operator involved in multi-class hypothesis classes (Eq. 1, left-hand side) can be removed [9, 31]. Applying this result leads to a simple sum of c RCs (Eq. (1), right-hand side), each of which can be bounded using standard theory [37]. In this way, Koltchinskii and Panchenko [10], Cortes et al. [40], and Mohri et al. [9] derive multi-class generalization error bounds that exhibit a quadratic dependency on the number of classes, which Kuznetsov et al. [41] improve to a linear dependency.

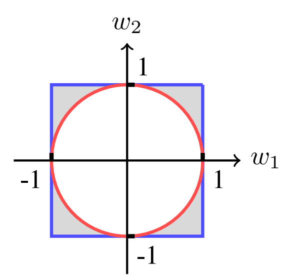


Fig. 1. Illustration of why Eq. (1) is loose. Consider a 1-dimensional binary classification problem with hypothesis class H consisting of functions mapping $x \in \mathbb{R}$ to $\max(h_1(x), h_2(x))$, where $h_j(x) = w_j x$ for $j = 1, 2$. Assume the class is regularized through the constraint $\|(w_1, w_2)\|_2 \leq 1$, so the left-hand side of the inequality (1) involves a supremum over the ℓ_2-norm constraint $\|(w_1, w_2)\|_2 \leq 1$. By contrast, the right-hand side of (1) has individual suprema for w_1 and w_2 (no coupling), resulting in a supremum over the ℓ_∞-norm constraint $\|(w_1, w_2)\|_\infty \leq 1$. Thus applying Eq. (1) enlarges the size of the constraint set by the area that is shaded in the figure, which grows as $O(\sqrt{c})$. In the present paper, we show a proof technique to elevate this problem, resulting in an improved bound (tighter by a factor of $\sqrt{c}$).

However, the reduction (1) comes at the expense of at least a linear dependency on the number of classes c, due to the sum in Eq. (1) (right-hand side), which consists of c terms. We show that this linear dependency can often be suboptimal because (1) does not take into account coupling among the classes. To understand why, we consider the example of MC-

SVM by Crammer and Singer [31], which uses an ℓ_2-norm constraint

$$\big\|(h_1,\ldots,h_c)\big\|_2 \leq \Lambda \tag{2}$$

to couple the components $h_1,\ldots,h_c$. The problem with Eq. (1) is that it decouples the components, resulting in the constraint $\big\|(h_1,\ldots,h_c)\big\|_\infty \leq \Lambda$, which—as illustrated in Fig. 1—is a poor approximation of (2).

In our previous work [28], we give a structural result addressing this shortcoming and tightly preserving the constraint defining the hypothesis class. Our result is based on the so-called GC [37], a notion similar to the RC. The difference in the two notions is that RC and GC are the suprema of a Rademacher and Gaussian process, respectively.

The core idea of our analysis is that we exploit a comparison inequality for the suprema of Gaussian processes known as *Slepian's Lemma* [42], by which we can remove, from the GC, the maximum operator that occurs in the definition of the hypothesis class, thus preserving the above mentioned coupling—we call the supremum of the resulting Gaussian process the *multi-class Gaussian complexity*.

On the basis of our structural result, we obtain in [28] a data-dependent error bound for [31] that exhibits—for the first time—a sublinear (square-root) dependency on the number of classes. When using a block $\ell_{2,p}$-norm constraint (with p close to 1), rather than an ℓ_2-norm constraint, one can reduce this dependency to be *logarithmic*, making the analysis appealing for classification with many classes.

We note that, addressing the same need, the following structural result [43, 44] has appeared since the publication of our previous work [28]:

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h\in H} \sum_{i=1}^n \epsilon_i f_i(h(\mathbf{x}_i)) \leq \sqrt{2} L \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h\in H} \sum_{i=1}^n \sum_{j=1}^c \epsilon_{ij} h_j(\mathbf{x}_i), \tag{3}$$

where $f_1,\ldots,f_n$ are L-Lipschitz continuous w.r.t. the ℓ_2-norm.

For the MC-SVM of Crammer and Singer [31], the above result leads to the same favorable square-root dependency on the number of classes as that of our previous result in [28]. We note, however, that the structural result (3) requires f_i to be Lipschitz continuous w.r.t. the ℓ_2-norm, while some multi-class loss functions [30, 32, 45] are Lipschitz continuous with a moderate Lipschitz constant, when choosing a more appropriate norm. In these cases, the analysis given in the present paper improves not only the classical results obtained through (1), but also the results obtained through (3).

2) Related Work on Data-independent Bounds: Data-independent generalization bounds refer to classical theoretical bounds that hold for any sample, with a certain probability over the draw of the samples [1, 46]. In their seminal contribution *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, Vapnik and Chervonenkis [47] propose one of the first bounds of that type—introducing the notion of VC dimension.

Several authors consider data-independent bounds for multi-class learning. By controlling the entropy numbers of linear operators with Maurey's theorem, Guermeur [11] derives generalization error bounds with a linear dependency on the number of classes. This is improved to a square-root dependency by Zhang [48] using ℓ_∞-norm CNs without considering the correlation among class-wise components. Pan et al. [49] consider a multi-class Parzen window classifier and derive an error bound with a quadratic dependency on the number of classes. Several authors present data-independent generalization bounds based on combinatorial dimensions, including the graph dimension, the Natarajan dimension d_{nat}, and its scale-sensitive analog $d_{\text{nat},\gamma}$ for margin γ [50–54].

Guermeur [50, 51] presents a generalization bound decaying as $O\big(\log c\sqrt{\frac{d_{\text{nat},\gamma}\log n}{n}}\big)$. When using an ℓ_∞-norm regularizer $d_{\text{nat},\gamma}$ is bounded by $O(c^2\gamma^{-2})$, and the generalization bound reduces to $O\big(\frac{c\log c}{\gamma}\sqrt{\frac{\log n}{n}}\big)$. The author does not give a bound for an ℓ_2-norm regularizer, which is more challenging due to the above mentioned coupling of the hypothesis components.

Daniely et al. [52] give a bound decaying as $O\big(\sqrt{\frac{d_{\text{nat}}(H)\log c}{n}}\big)$, which changes to $O\big(\sqrt{\frac{dc\log c}{n}}\big)$ for multi-class linear classifiers since the associated Natarajan dimension grows as $O(dc)$ [53].

Guermeur [55] has recently established an ℓ_p-norm Sauer-Shelah lemma for large-margin multi-class classifiers, based on which error bounds with a square-root dependency on the number of classes are derived. This setting comprises the MC-SVM by Crammer and Singer [31].

What is common in all the above mentioned data-independent bounds is their super logarithmic dependency (square root at best) on the number of classes. As a notable exception, Kontorovich and Weiss [56] show a bound exhibiting a logarithmic dependency on the number of classes. However, their bound holds only for the specific nearest-neighbor-based algorithm that they propose, so their analysis does not cover the commonly used multi-class learning machines mentioned in the introduction (such as multinomial logistic regression and classic MC-SVMs). Furthermore, their bound is of the order $\min\Big\{O\big(\gamma^{-1}\big(\frac{\log c}{n}\big)^{\frac{1}{1+D}}\big), O\big(\gamma^{-\frac{D}{2}}\big(\frac{\log c}{n}\big)^{\frac{1}{2}}\big)\Big\}$, which admits an *exponential* dependence on the doubling dimension D of the metric space in which the learning occurs. For instance, for linear learning methods with dimension d, the doubling dimension D grows linearly in d, so the bound in [56] grows exponentially in d. For kernel-based learning using an infinite doubling dimension (e.g., Gaussian kernels) the bound is void.

B. Contributions of this Paper

This paper aims to contribute a solid theoretical foundation for learning with many class labels by presenting data-dependent generalization error bounds with relaxed dependencies on the number of classes. We develop two approaches to establish data-dependent error bounds: one based on multi-class GCs and one based on empirical ℓ_∞-norm CNs. We give specific examples to show that each of these two approaches has its distinct advantages and may yield error bounds tighter than the other. We also develop novel algorithms to train the ℓ_p-norm MC-SVM [28] and report the experimental results. Below we summarize the main results of this paper.

1) Tighter Generalization Bounds by Gaussian Complexities: As an extension of our NIPS 2015 conference paper, our GC-based analysis depends on a novel structural result on GCs (Lemma 1 below) that is able to preserve the correlation among class-wise components. Similar to Maurer [43] and Cortes et al. [44], our structural result applies to function classes induced by operators satisfying a Lipschitz continuity. However, here we measure the Lipschitz continuity with respect to a specially crafted variant of the ℓ_2-norm involving a Lipschitz constant pair (L_1, L_2) (cf. Definition 2 below), motivated by the observation that some multi-class loss functions satisfy this Lipschitz continuity with a relatively small L_1 in a dominant term and a relatively large L_2 in a non-dominant term. This process allows us to improve the error bounds based on the structural result (3) for MC-SVMs with a relatively large L_2.

Based on this new structural result, we present an error bound for multi-class empirical risk minimization algorithms using an arbitrary norm as the regularizer. As instantiations of our general bound, we compute specific bounds for the $\ell_{2,p}$-norm and Schatten p-norm regularizers. We apply this general GC-based bound to some popular MC-SVMs [29, 31–33, 45].

Our GC-based analysis yields the first error bound for top-k MC-SVM [30] as a decreasing function in k. When setting k proportional to c, the bound does not depend on the number of classes. By contrast, error bounds based on the structural result (3) fail to provide insight into the influence of k on the generalization performance because the involved Lipschitz constant is dominated by a constant. For the MC-SVM of Weston and Watkins [32], our analysis yields a bound exhibiting a linear dependency on the number of classes, which improves the dependency $O(c^{\frac{3}{2}})$ based on the structural result (3). For the MC-SVM by Jenssen et al. [45], our analysis yields a bound with no dependencies on c, whereas the error bound based on the structural result (3) has a square-root dependency. This demonstrates the effectiveness of our new structural result in capturing the Lipschitz continuity w.r.t. a variant of the ℓ_2-norm.

2) Tighter Generalization Bounds by Covering Numbers: While the GC-based analysis uses the Lipschitz continuity measured by the ℓ_2-norm or a variant thereof, some multi-class loss functions are Lipschitz continuous w.r.t. the ℓ_∞-norm with a moderate Lipschitz constant. To apply the GC-based error bounds, we need to transform this ℓ_∞-norm Lipschitz continuity into the ℓ_2-norm Lipschitz continuity at the cost of a multiplicative factor of $\sqrt{c}$. Motivated by this observation, we present another data-dependent analysis based on empirical ℓ_∞-norm CNs to fully exploit the Lipschitz continuity measured by the ℓ_∞-norm. We show that this process leads to bounds with a weaker dependency on the number of classes.

The core idea is to introduce a linear and scalar-valued function class induced by training examples to extract all the components of the hypothesis functions on the training examples, which allows us to relate the empirical ℓ_∞-norm CNs of the loss function classes to that of this linear function class. Our main result is a data-dependent error bound for general MC-SVMs expressed in terms of the *worst-case* RC of a linear function class, for which we establish lower and upper bounds that match up to a constant factor. The analysis in this direction is unrelated to the conference version [28] and provides an alternative to GC-based arguments.

As direct applications, we derive other data-dependent generalization error bounds that scale sublinearly for ℓ_p-norm MC-SVM and Schatten-p norm MC-SVM, and *logarithmically* for top-k MC-SVM [30], trace-norm regularized MC-SVM [57], multinomial logistic regression [29] and the MC-SVM by Crammer and Singer [31]. Note that the previously best results for the MC-SVM in [31] and multinomial logistic regression scale as the square root of the number of classes [48].

3) Novel Algorithms with Empirical Verifications: We propose a novel algorithm to train ℓ_p-norm MC-SVM [28] using the Frank-Wolfe algorithm [58], for which we show that the involved linear optimization problem has a closed-form solution, making the implementation of the Frank-Wolfe algorithm simple and efficient. This method avoids the introduction of class weights used in our previous optimization algorithm [28], which moreover applies to only the case $1 \leq p \leq 2$. The effectiveness of ℓ_p-norm MC-SVM is demonstrated by empirical comparisons with several baseline methods on benchmark datasets. We also empirically show that our generalization bounds really capture models' generalization performance on the number of classes, which in turn suggest a structural risk that is able to guide the selection of model parameters.

III. Main Results

A. Problem Setting

In multi-class classification with c classes, we are given training examples $S = \{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the input space, and $\mathcal{Y} = \{1, \ldots, c\}$ is the output space. We assume that $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independently drawn from a probability measure P defined on $\mathcal{Z}$.

Our aim is to learn, from a hypothesis space H, a hypothesis $h = (h_1, \ldots, h_c) : \mathcal{X} \mapsto \mathbb{R}^c$ used for prediction via the rule $\mathbf{x} \mapsto \arg\max_{y \in \mathcal{Y}} h_y(\mathbf{x})$. We consider prediction functions of the form $h_j^{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}_j, \phi(\mathbf{x}) \rangle$, where ϕ is a feature map associated with a Mercer kernel K defined over $\mathcal{X} \times \mathcal{X}$, and $\mathbf{w}_j$ belongs to the reproducing kernel Hilbert space H_K induced from K with the inner product $\langle \cdot, \cdot \rangle$ satisfying $K(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}}) \rangle$.

We consider hypothesis spaces of the form

$$H_\tau = \Big\{ h^{\mathbf{w}} = \big(\langle \mathbf{w}_1, \phi(\mathbf{x}) \rangle, \ldots, \langle \mathbf{w}_c, \phi(\mathbf{x}) \rangle \big) : \mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_c) \in H_K^c, \tau(\mathbf{w}) \leq \Lambda \Big\}, \quad (4)$$

where τ is a functional defined on $H_K^c := \underbrace{H_K \times \cdots \times H_K}_{c \text{ times}}$ and $\Lambda > 0$. Here we omit the dependency on Λ for brevity.

We consider a general problem setting with $\Psi_y(h_1(\mathbf{x}), \ldots, h_c(\mathbf{x}))$ used to measure the prediction quality of model h at $(\mathbf{x}, y)$ [48, 59], where $\Psi_y : \mathbb{R}^c \mapsto \mathbb{R}_+$ is a real-valued function taking a c-component vector as its argument. The general loss function Ψ_y is widely used in many MC-SVMs, including the models of Crammer and Singer [31], Weston and Watkins [32], Lee et al. [33], Zhang [48], and Lapin et al. [30].

TABLE I
NOTATION USED IN THIS PAPER AND THE PAGE NUMBER WHERE IT FIRST OCCURS.

notation	meaning	page
$\mathcal{X}, \mathcal{Y}$	the input space and output space, respectively	4
S	the set of training examples $\{\mathbf{z}_i = (\mathbf{x}_i, y_i)\} \in \mathcal{X} \times \mathcal{Y}$	4
c	number of classes	4
K	Mercer kernel	4
ϕ	feature map associated to a kernel K	4
H_K	reproducing kernel Hilbert space induced by a Mercer kernel K	4
H_K^c	c-fold Cartesian product of the reproducing kernel Hilbert space H_K	4
$\mathbf{w}$	$(\mathbf{w}_1, \ldots, \mathbf{w}_c) \in H_K^c$	4
$h^{\mathbf{w}}$	prediction function $(\langle \mathbf{w}_1, \phi(\mathbf{x}) \rangle, \ldots, \langle \mathbf{w}_c, \phi(\mathbf{x}) \rangle)$	4
H_τ	hypothesis space for MC-SVM constrained by a regularizer τ	4
Ψ_y	multi-class loss function for class label y	4
$\|\cdot\|_p$	ℓ_p-norm defined on $\mathbb{R}^c$	5
$\|\cdot\|_{2,p}$	$\ell_{2,p}$ norm defined on H_K^c	5
$\langle \mathbf{w}, \mathbf{v} \rangle$	inner product on H_K^c as $\sum_{j=1}^c \langle \mathbf{w}_j, \mathbf{v}_j \rangle$	5
$\|\cdot\|_*$	dual norm of $\|\cdot\|$	5
$\mathbb{N}_n$	the set $\{1, \ldots, n\}$	5
p^*	dual exponent of p satisfying $1/p + 1/p^* = 1$	5
$\mathbb{E}_{\mathbf{u}}$	the expectation w.r.t. random $\mathbf{u}$	5
B_Ψ	the constant $\sup_{(\mathbf{x},y)\in\mathcal{Z}, h\in H_\tau} \Psi_y(h(\mathbf{x}))$	5
$\hat{B}_\Psi$	the constant $n^{-\frac{1}{2}} \sup_{h\in H_\tau} \big\|\big(\Psi_{y_i}(h(\mathbf{x}_i))\big)_{i=1}^n\big\|_2$	5
$\hat{B}$	the constant $\max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2 \sup_{\mathbf{w}:\tau(\mathbf{w})\leq\Lambda} \|\mathbf{w}\|_{2,\infty}$	5
A_τ	the term defined in (5)	5
I_y	indices of examples with class label y	5
$\|\cdot\|_{S_p}$	Schatten-p norm of a matrix	5
$\mathfrak{R}_S(H)$	empirical Rademacher complexity of H w.r.t. sample S	5
$\mathfrak{G}_S(H)$	empirical Gaussian complexity of H w.r.t. sample S	5
$\mathfrak{R}_n(H)$	worst-case Rademacher complexity of H w.r.t. n examples	5
$\widetilde{H}_\tau$	class of scalar-valued linear functions defined on H_K^c	6
$\widetilde{S}$	an enlarged set of cardinality nc defined in (9)	6
$\widetilde{S}'$	a set of cardinality n defined in (11)	6
$F_{\tau,\Lambda}$	loss function class for MC-SVM	7
$\rho_h(\mathbf{x}, y)$	margin of h at $(\mathbf{x}, y)$	8
$\mathcal{N}_\infty(\epsilon, F, S)$	empirical covering number of F w.r.t. sample S	19
$\mathrm{fat}_\epsilon(F)$	fat-shattering dimension of F	19

B. Notations

We now present some notation used throughout this paper (see also Table I). We say that a function $f : \mathbb{R}^c \mapsto \mathbb{R}$ is L-Lipschitz continuous w.r.t. a norm $\|\cdot\|$ in $\mathbb{R}^c$ if

$$|f(\mathbf{t}) - f(\mathbf{t}')| \leq L\|(t_1 - t_1', \ldots, t_c - t_c')\|, \quad \forall \mathbf{t}, \mathbf{t}' \in \mathbb{R}^c.$$

The ℓ_p-norm of a vector $\mathbf{t} = (t_1, \ldots, t_c)$ is defined as $\|\mathbf{t}\|_p = \big[\sum_{j=1}^c |t_j|^p\big]^{\frac{1}{p}}$. For any $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_c) \in H_K^c$ and $p \geq 1$, we define the structure norm $\|\mathbf{v}\|_{2,p} = \big[\sum_{j=1}^c \|\mathbf{v}_j\|_2^p\big]^{\frac{1}{p}}$. Here, for brevity, we denote by $\|\mathbf{v}_j\|_2$ the norm of $\mathbf{v}_j$ in H_K. For any $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_c), \mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_c) \in H_K^c$, we denote $\langle \mathbf{w}, \mathbf{v} \rangle = \sum_{j=1}^c \langle \mathbf{w}_j, \mathbf{v}_j \rangle$. For any $n \in \mathbb{N}$, we introduce the notation $\mathbb{N}_n := \{1, \ldots, n\}$. For any $p \geq 1$, we denote by p^* the dual exponent of p satisfying $1/p + 1/p^* = 1$. For any norm $\|\cdot\|$ we use $\|\cdot\|_*$ to represent its dual norm. Furthermore, we define $B_\Psi = \sup_{(\mathbf{x},y)\in\mathcal{Z}} \sup_{h^{\mathbf{w}}\in H_\tau} \Psi_y(h^{\mathbf{w}}(\mathbf{x}))$, $\hat{B}_\Psi = n^{-\frac{1}{2}} \sup_{h^{\mathbf{w}}\in H_\tau} \big\|\big(\Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\big)_{i=1}^n\big\|_2$, and $\hat{B} = \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2 \sup_{\mathbf{w}:\tau(\mathbf{w})\leq\Lambda} \|\mathbf{w}\|_{2,\infty}$. For any functional τ over H_K^c, we introduce the following notation to write our bounds compactly

$$A_\tau := \sup_{h^{\mathbf{w}}\in H_\tau} \Big[\mathbb{E}_{\mathbf{x},y}\Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big] - 3B_\Psi\Big[\frac{\log\frac{2}{\delta}}{2n}\Big]^{\frac{1}{2}}, \quad (5)$$

where we omit the dependency on n and loss function for brevity. Note that, for any random $\mathbf{u}$, the notation $\mathbb{E}_{\mathbf{u}}$ denotes the expectation w.r.t. $\mathbf{u}$. For any $y \in \mathcal{Y}$, we use $I_y = \{i \in \mathbb{N}_n : y_i = y\}$ to represent the indices of the examples with label y.

If ϕ is the identity map, then the hypothesis $h^{\mathbf{w}}$ can be compactly represented by a matrix $W = (\mathbf{w}_1, \ldots, \mathbf{w}_c) \in \mathbb{R}^{d\times c}$. For any $p \geq 1$, the Schatten-p norm of a matrix $W \in \mathbb{R}^{d\times c}$ is defined as the ℓ_p-norm of the vector of singular values $\sigma(W) := (\sigma_1(W), \ldots, \sigma_{\min\{c,d\}}(W))^\top$ (the singular values are assumed to be sorted in non-increasing order), i.e., $\|W\|_{S_p} := \|\sigma(W)\|_p$.

C. Data-dependent Bounds by Gaussian Complexities

We first present data-dependent analysis based on the established methodology of RCs and GCs [37].

Definition 1 (Empirical Rademacher and Gaussian complexities)**.** Let H be a class of real-valued functions defined over a space $\tilde{\mathcal{Z}}$ and $S' = \{\tilde{\mathbf{z}}_i\}_{i=1}^n \in \tilde{\mathcal{Z}}^n$. The *empirical Rademacher and Gaussian complexities* of H with respect to S' are, respectively, defined as

$$\mathfrak{R}_{S'}(H) = \mathbb{E}_{\boldsymbol{\epsilon}}\big[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^n \epsilon_i h(\tilde{\mathbf{z}}_i)\big],$$

$$\mathfrak{G}_{S'}(H) = \mathbb{E}_{\boldsymbol{g}}\big[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^n g_i h(\tilde{\mathbf{z}}_i)\big],$$

where $\epsilon_1,\ldots,\epsilon_n$ are independent Rademacher variables, and $g_1,\ldots,g_n$ are independent $N(0,1)$ random variables. We define the *worst-case* Rademacher complexity as $\mathfrak{R}_n(H) = \sup_{S'\in\tilde{\mathcal{Z}}^n}\mathfrak{R}_{S'}(H)$.

Existing data-dependent analyses build on either the structural result (1) or (3), which either ignore the correlation among predictors associated with individual class labels or require f_i to be Lipschitz continuous w.r.t. the ℓ_2-norm. Below we introduce a new structural complexity result based on the following Lipschitz property w.r.t. a variant of the ℓ_2-norm. The motivation of this Lipschitz continuity is that some multi-class loss functions satisfy (6) with a relatively small L_1 and a relatively large L_2, the latter of which is not strongly influential since it is involved in a single component.

Definition 2 (Lipschitz continuity w.r.t. a variant of the ℓ_2-norm)**.** We say a function $f:\mathbb{R}^c\mapsto\mathbb{R}$ is *Lipschitz continuous w.r.t. a variant of the ℓ_2-norm* involving a Lipschitz constant pair (L_1,L_2) and index $r\in\{1,\ldots,c\}$ if

$$|f(\mathbf{t})-f(\mathbf{t}')| \le L_1\|(t_1-t_1',\ldots,t_c-t_c')\|_2 + L_2|t_r-t_r'| \quad (6)$$

for all $\mathbf{t},\mathbf{t}'\in\mathbb{R}^c$.

We now present our first core result of this paper, the following structural lemma. Proofs of results in this section are given in Section VI-A.

Lemma 1 (Structural Lemma)**.** *Let H be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}^c$. Let $L_1, L_2\ge 0$ be two constants and $r:\mathbb{N}\mapsto\mathcal{Y}$. Let $f_1,\ldots,f_n$ be a sequence of functions from $\mathbb{R}^c$ to $\mathbb{R}$. Suppose that for any $i\in\mathbb{N}_n$, f_i is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving a Lipschitz constant pair (L_1,L_2) and index $r(i)$. Let $g_1,\ldots,g_n,g_{11},\ldots,g_{nc}$ be a sequence of independent $N(0,1)$ random variables. Then, for any sample $\{\tilde{\mathbf{x}}_i\}_{i=1}^n\in\mathcal{X}^n$ we have*

$$\mathbb{E}_{\boldsymbol{g}}\sup_{h\in H}\sum_{i=1}^n g_i f_i(h(\tilde{\mathbf{x}}_i)) \le \sqrt{2}L_1\mathbb{E}_{\boldsymbol{g}}\sup_{h\in H}\sum_{i=1}^n\sum_{j=1}^c g_{ij}h_j(\tilde{\mathbf{x}}_i) + \sqrt{2}L_2\mathbb{E}_{\boldsymbol{g}}\sup_{h\in H}\sum_{i=1}^n g_i h_{r(i)}(\tilde{\mathbf{x}}_i). \quad (7)$$

Lemma 1 controls the GC of the multi-class loss function class by that of the original hypothesis class, thereby removing the dependency on the potentially cumbersome operator f_i in the definition of the loss function class (for instance for Crammer and Singer [31], f_i would be the component-wise maximum). The above lemma is based on a comparison (Slepian's lemma, Lemma 20 below) of the suprema of Gaussian processes.

Equipped with Lemma 1, we can present our main results based on GCs. Eq. (13) is a data-dependent bound in terms of the GC of the following linear scalar-valued function class

$$\widetilde{H}_\tau := \{\mathbf{v}\mapsto\langle\mathbf{w},\mathbf{v}\rangle : \mathbf{w},\mathbf{v}\in H_K^c, \tau(\mathbf{w})\le\Lambda, \mathbf{v}\in\widetilde{S}\}, \quad (8)$$

where $\widetilde{S}$ is defined as follows

$$\widetilde{S} = \Big\{\underbrace{\tilde{\boldsymbol{\phi}}_1(\mathbf{x}_1),\tilde{\boldsymbol{\phi}}_2(\mathbf{x}_1),\ldots,\tilde{\boldsymbol{\phi}}_c(\mathbf{x}_1)}_{\text{induced by }\mathbf{x}_1},\underbrace{\tilde{\boldsymbol{\phi}}_1(\mathbf{x}_2),\tilde{\boldsymbol{\phi}}_2(\mathbf{x}_2),\ldots,\tilde{\boldsymbol{\phi}}_c(\mathbf{x}_2)}_{\text{induced by }\mathbf{x}_2}, \ldots,\underbrace{\tilde{\boldsymbol{\phi}}_1(\mathbf{x}_n),\ldots,\tilde{\boldsymbol{\phi}}_c(\mathbf{x}_n)}_{\text{induced by }\mathbf{x}_n}\Big\} \quad (9)$$

and, for any $\mathbf{x}\in\mathcal{X}$, we use the notation

$$\tilde{\boldsymbol{\phi}}_j(\mathbf{x}) := \big(\underbrace{0,\ldots,0}_{j-1},\phi(\mathbf{x}),\underbrace{0,\ldots,0}_{c-j}\big)\in H_K^c, \quad j\in\mathbb{N}_c. \quad (10)$$

Note that $\widetilde{H}_\tau$ is a class of functions defined on a finite set $\widetilde{S}$. We also introduce

$$\widetilde{S}' = \Big\{\tilde{\boldsymbol{\phi}}_{y_1}(\mathbf{x}_1),\tilde{\boldsymbol{\phi}}_{y_2}(\mathbf{x}_2),\ldots,\tilde{\boldsymbol{\phi}}_{y_n}(\mathbf{x}_n)\Big\}. \quad (11)$$

The terms $\widetilde{S},\widetilde{S}'$ and $\tilde{\boldsymbol{\phi}}_j(\mathbf{x})$ are motivated by the following identity

$$\begin{aligned}\langle\mathbf{w},\tilde{\boldsymbol{\phi}}_k(\mathbf{x})\rangle &= \Big\langle(\mathbf{w}_1,\ldots,\mathbf{w}_c),\big(\underbrace{0,\ldots,0}_{k-1},\phi(\mathbf{x}),\underbrace{0,\ldots,0}_{c-k}\big)\Big\rangle\\ &= \langle\mathbf{w}_k,\phi(\mathbf{x})\rangle, \quad \forall k\in\mathbb{N}_c. \end{aligned}\quad (12)$$

Hence, the right-hand side of (7) can be rewritten as Gaussian complexities of $\widetilde{H}_\tau$ when $H=H_\tau$.

Theorem 2 (Data-dependent bounds for general regularizer and Lipschitz continuous loss w.r.t. Def. 2)**.** *Consider the hypothesis space H_τ in (4) with $\tau(\mathbf{w})=\|\mathbf{w}\|$, where $\|\cdot\|$ is a norm defined on H_K^c. Suppose there exist $L_1,L_2\in\mathbb{R}_+$ such that Ψ_y is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving a Lipschitz constant pair (L_1,L_2) and index y for all $y\in\mathcal{Y}$. Then, for any $0<\delta<1$, with probability of at least $1-\delta$, we have*

$$A_\tau \le 2\sqrt{\pi}\Big[L_1 c\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau) + L_2\mathfrak{G}_{\widetilde{S}'}(\widetilde{H}_\tau)\Big] \quad (13)$$

and

$$A_\tau \le \frac{2\Lambda\sqrt{\pi}}{n}\Big[L_1\mathbb{E}_{\boldsymbol{g}}\Big\|\big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\big)_{j=1}^c\Big\|_* + L_2\mathbb{E}_{\boldsymbol{g}}\Big\|\big(\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\big)_{j=1}^c\Big\|_*\Big], \quad (14)$$

where $g_1,\ldots,g_n,g_{11},\ldots,g_{nc}$ are independent $N(0,1)$ random variables.

Remark 1 (Motivation of Lipschitz continuity w.r.t. Def. 2)**.** The dominant term on the right-hand side of (13) is $L_1 c\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau)$ if $L_2 = O(\sqrt{c}L_1)$. This explains the motivation to introduce the new structural result (7) to exploit the Lipschitz continuity w.r.t. a variant of the ℓ_2-norm involving a large L_2. For comparison, if we apply the previous structural result (3) for loss functions satisfying (6), then the associated ℓ_2-Lipschitz constant is $L_1 + L_2$, resulting in the following bound

$$A_\tau \leq 2\sqrt{\pi}(L_1 + L_2)c\mathfrak{R}_{\widetilde{S}}(\widetilde{H}_\tau),$$

which is worse than (13) when $L_1 = O(L_2)$ since the dominant term becomes $L_2 c\mathfrak{R}_{\widetilde{S}}(\widetilde{H}_\tau)$. Many popular loss functions satisfy (6) with $L_1 = O(L_2)$ [30, 32, 45]. For example, the loss function used in the top-k SVM [30] satisfies (6) with $(L_1, L_2) = (\frac{1}{\sqrt{k}}, 1)$, which, as we will show, allows us to derive data-dependent bounds with no dependencies on the number of classes by setting k proportional to c. By comparison, the $(k^{-\frac{1}{2}} + 1)$-Lipschitz continuity w.r.t. ℓ_2-norm does not capture the special structure of the top-k loss function since $k^{-\frac{1}{2}}$ is dominated by the constant 1. As further examples, the loss function in Weston and Watkins [32] satisfies (6) with $(L_1, L_2) = (\sqrt{c}, c)$, while the loss function in Jenssen et al. [45] satisfies (27) with $(L_1, L_2) = (0, 1)$.

We now consider two applications of Theorem 2 by considering $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}$ defined on H_K^c [28] and $\tau(W) = \|W\|_{S_p}$ defined on $\mathbb{R}^{d\times c}$ [57], respectively.

Corollary 3 (Data-dependent bound for ℓ_p-norm regularizer and Lipschitz continuous loss w.r.t. Def. 2)**.** *Consider the hypothesis space $H_{p,\Lambda} := H_{\tau,\Lambda}$ in* (4) *with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}, p \geq 1$. If there exist $L_1, L_2 \in \mathbb{R}_+$ such that Ψ_y is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving a Lipschitz constant pair (L_1, L_2) and index y for all $y \in \mathcal{Y}$, then for any $0 < \delta < 1$, the following inequality holds with probability of at least $1 - \delta$ (we use the abbreviation $A_p = A_\tau$ with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}$)*

$$A_p \leq \frac{2\Lambda\sqrt{\pi}}{n}\Big[\sum_{i=1}^n K(\mathbf{x}_i, \mathbf{x}_i)\Big]^{\frac{1}{2}} \inf_{q\geq p}\Big[L_1(q^*)^{\frac{1}{2}}c^{\frac{1}{q^*}} + L_2(q^*)^{\frac{1}{2}}\max(c^{\frac{1}{q^*}-\frac{1}{2}}, 1)\Big]. \quad (15)$$

Corollary 4 (Data-dependent bound for Schatten-p norm regularizer and Lipschitz continuous loss w.r.t. Def. 2)**.** *Let ϕ be the identity map and represent $\mathbf{w}$ by a matrix $W \in \mathbb{R}^{d\times c}$. Consider the hypothesis space $H_{S_p,\Lambda} := H_{\tau,\Lambda}$ in* (4) *with $\tau(W) = \|W\|_{S_p}, p \geq 1$. If there exist $L_1, L_2 \in \mathbb{R}_+$ such that Ψ_y is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving a Lipschitz constant pair (L_1, L_2) and index y for all $y \in \mathcal{Y}$, then for any $0 < \delta < 1$ with probability of at least $1 - \delta$, we have (we use the abbreviation $A_{S_p} = A_\tau$ with $\tau(W) = \|W\|_{S_p}$)*

$$A_{S_p} \leq \begin{cases} \frac{2^{\frac{3}{4}}\pi\Lambda}{n\sqrt{e}} \inf\limits_{p\leq q\leq 2}(q^*)^{\frac{1}{2}}\Big\{(L_1 c^{\frac{1}{q^*}} + L_2)\Big[\sum\limits_{i=1}^n \|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}} \\ \quad + L_1 c^{\frac{1}{2}}\Big\|\sum\limits_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_{\frac{q^*}{2}}}^{\frac{1}{2}}\Big\}, & \text{if } p \leq 2, \\ \frac{2^{\frac{5}{4}}\pi\Lambda\big(L_1 c^{\frac{1}{2}} + L_2\big)\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{n\sqrt{e}}\Big[\sum_{i=1}^n \|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}}, & \\ & \text{otherwise}. \end{cases} \quad (16)$$

In comparison to Corollary 3, the error bound of Corollary 4 involves an additional term $O\big(c^{\frac{1}{2}}n^{-1}\big\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\big\|_{S_{\frac{q^*}{2}}}^{\frac{1}{2}}\big)$ for the case $p \leq 2$ due to the need to apply the non-commutative Khintchine-Kahane inequality (71) for Schatten norms. As we will show in Section IV, from Corollaries 3 and 4 we can derive error bounds with sublinear dependencies on the number of classes for ℓ_p-norm and Schatten-p norm MC-SVMs. Furthermore, the dependency is logarithmic for the ℓ_p-norm MC-SVM [28] when p approaches 1.

D. Data-dependent Bounds by Covering Numbers

The data-dependent generalization bounds given in subsection III-C assume the loss function Ψ_y to be Lipschitz continuous w.r.t. a variant of the ℓ_2-norm. However, some typical loss functions used in the multi-class setting are Lipschitz continuous w.r.t. the much milder ℓ_∞-norm with a comparable Lipschitz constant [48]. This mismatch between the norms w.r.t. which Lipschitz continuity is measured requires an additional step of controlling the ℓ_∞-norm of vector-valued predictors by the ℓ_2-norm in the application of Theorem 2, at the cost of a possible multiplicative factor of $\sqrt{c}$. This subsection aims to avoid this loss in the class-size dependency by presenting data-dependent analysis based on empirical ℓ_∞-norm CNs to directly use the Lipschitz continuity measured by the ℓ_∞-norm.

The key step in this approach lies in estimating the empirical CNs of the loss function class

$$F_{\tau,\Lambda} := \{(\mathbf{x}, y) \mapsto \Psi_y(h^{\mathbf{w}}(\mathbf{x})) : h^{\mathbf{w}} \in H_\tau\}. \quad (17)$$

A difficulty towards this aim consists in the non-linearity of $F_{\tau,\Lambda}$ and the fact that $h^{\mathbf{w}} \in H_\tau$ takes vector-valued outputs, whereas standard analyses are limited to scalar-valued and essentially linear (kernel) function classes [60–62]. We bypass this obstacle by considering a related linear scalar-valued function class $\widetilde{H}_\tau$ defined in (8). A key motivation in introducing $\widetilde{H}_\tau$ is that the CNs of $F_{\tau,\Lambda}$ w.r.t. $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (CNs are defined in subsection VI-B) can be related to that of the function class $\{\mathbf{v} \mapsto \langle \mathbf{w}, \mathbf{v}\rangle : \tau(\mathbf{w}) \leq \Lambda\}$, w.r.t. the set $\widetilde{S}$ defined in (9). The latter is easily addressed since it is a linear and scalar-valued function class, to which standard arguments apply. Specifically, to approximate the projection of $F_{\tau,\Lambda}$ onto the examples S with (ϵ, ℓ_∞)-covers (cf. Definition 3 below), the ℓ_∞-Lipschitz continuity of the loss function requires us to approximate the set $\big\{\big(\langle \mathbf{w}_j, \phi(\mathbf{x}_i)\rangle_{i\in\mathbb{N}_n, j\in\mathbb{N}_c}\big) : \tau(\mathbf{w}) \leq \Lambda\big\}$, which, according to (12), is exactly the projection of $\widetilde{H}_\tau$ onto $\widetilde{S}$: $\big\{\big(\langle \mathbf{w}, \tilde{\phi}_j(\mathbf{x}_i)\rangle_{i\in\mathbb{N}_n, j\in\mathbb{N}_c}\big) : \tau(\mathbf{w}) \leq \Lambda\big\}$. This result motivates the definition of $\widetilde{H}_\tau$ in (8) and $\widetilde{S}$ in (9).

Theorem 5 reduces the estimation of $\mathfrak{R}_S(F_{\tau,\Lambda})$ to bounding $\mathfrak{R}_{nc}(\widetilde{H}_\tau)$, based on which the data-dependent error bounds are given in Theorem 6. Note that $\mathfrak{R}_{nc}(\widetilde{H}_\tau)$ is data-dependent since $\widetilde{H}_\tau$ is a class of functions defined on a finite set induced by training examples. The proofs of complexity bounds in Proposition 7 and Proposition 8 are given in subsection VI-C and Appendix B, respectively. The proofs of error bounds in this subsection are given in subsection VI-B.

Theorem 5 (Worst-case RC bound)**.** *Suppose that Ψ_y is L-Lipscthiz continuous w.r.t. the ℓ_∞-norm for any $y \in \mathcal{Y}$ and assume that $\hat{B}_\Psi \leq 2e\hat{B}ncL$. Then the RC of $F_{\tau,\Lambda}$ can be bounded by*

$$\mathfrak{R}_S(F_{\tau,\Lambda}) \leq 16L\sqrt{c\log 2}\mathfrak{R}_{nc}(\widetilde{H}_\tau)\Big(1+\log_2^{\frac{3}{2}}\frac{\hat{B}n\sqrt{c}}{\mathfrak{R}_{nc}(\widetilde{H}_\tau)}\Big).$$

Theorem 6 (Data-dependent bounds for general regularizer and Lipschitz continuous loss function w.r.t. $\|\cdot\|_\infty$)**.** *Under the condition of Theorem 5, for any $0 < \delta < 1$, with probability of at least $1-\delta$, we have*

$$A_\tau \leq 27L\sqrt{c}\mathfrak{R}_{nc}(\widetilde{H}_\tau)\Big(1+\log_2^{\frac{3}{2}}\frac{\hat{B}n\sqrt{c}}{\mathfrak{R}_{nc}(\widetilde{H}_\tau)}\Big).$$

The application of Theorem 6 requires to control the *worst-case* RC of the linear function class $\widetilde{H}_\tau$ from both below and above, to which the following two propositions give tight estimates for $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}$ defined on H_K^c [28] and $\tau(W) = \|W\|_{S_p}$ defined on $\mathbb{R}^{d\times c}$ [57].

Proposition 7 (Lower and upper bound on worst-case RC for ℓ_p-norm regularizer)**.** *For $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}, p \geq 1$ in* (8)*, the function class $\widetilde{H}_\tau$ becomes*

$$\widetilde{H}_p := \big\{\mathbf{v} \mapsto \langle\mathbf{w},\mathbf{v}\rangle : \mathbf{w},\mathbf{v} \in H_K^c, \|\mathbf{w}\|_{2,p} \leq \Lambda, \mathbf{v} \in \widetilde{S}\big\}.$$

The RC of $\widetilde{H}_p$ can be upper and lower bounded by

$$\Lambda \max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2(2n)^{-\frac{1}{2}}c^{-\frac{1}{\max(2,p)}} \leq \mathfrak{R}_{nc}(\widetilde{H}_p) \leq \Lambda \max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2 n^{-\frac{1}{2}}c^{-\frac{1}{\max(2,p)}}. \quad (18)$$

Remark 2 (Phase transition for p-norm regularized space)**.** We see an interesting phase transition at $p = 2$. The *worst-case* RC of $\widetilde{H}_p$ decays as $O((nc)^{-\frac{1}{2}})$ for the case $p \leq 2$, and decays as $O(n^{-\frac{1}{2}}c^{-\frac{1}{p}})$ for the case $p > 2$. Indeed, the definition of $\widetilde{S}$ by (9) implies $\|\mathbf{v}\|_{2,\infty} = \|\mathbf{v}\|_{2,p}$ for all $\mathbf{v} \in \widetilde{S}$ and $p \geq 1$ (sparsity of elements in $\widetilde{S}$), from which we derive the following identity

$$\max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\sum_{j=1}^{c}\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2 = \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\sum_{i=1}^{nc}\|\mathbf{v}^i\|_{2,\infty}^2 = nc\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2^2, \quad (19)$$

where $\mathbf{v}_j^i$ is the j-th component of $\mathbf{v}^i \in \widetilde{S}$. That is, we have an automatic constraint on $\big\|\big(\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2\big)_{j=1}^c\big\|_1$ for all $\mathbf{v}^i \in \widetilde{S}, i \in \mathbb{N}_{nc}$. Furthermore, according to (65), we know $nc\mathfrak{R}_{nc}(\widetilde{H}_p)$ can be controlled in terms of $\max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_n}\big\|\big(\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2\big)_{j=1}^c\big\|_{\frac{p^*}{2}}$, for which an appropriate p to fully use the identity (19) is $p = 2$. This explains the phase transition phenomenon.

Proposition 8 (Lower and upper bound on worst-case RC for Schatten-p norm regularizer)**.** *Let ϕ be the identity map and represent $\mathbf{w}$ by a matrix $W \in \mathbb{R}^{d\times c}$. For $\tau(W) = \|W\|_{S_p}, p \geq 1$ in* (8)*, the function class $\widetilde{H}_\tau$ becomes*

$$\widetilde{H}_{S_p} := \big\{V \mapsto \langle W, V\rangle : W \in \mathbb{R}^{d\times c}, \|W\|_{S_p} \leq \Lambda, V \in \widetilde{S} \subset \mathbb{R}^{d\times c}\big\}. \quad (20)$$

The RC of $\widetilde{H}_{S_p}$ can be upper and lower bounded by

$$\begin{cases} \Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2(2nc)^{-\frac{1}{2}} \leq \mathfrak{R}_{nc}(\widetilde{H}_{S_p}) \leq \Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2(nc)^{-\frac{1}{2}}, & \text{if } p \leq 2, \\ \Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2(2nc)^{-\frac{1}{2}} \leq \mathfrak{R}_{nc}(\widetilde{H}_{S_p}) \leq \frac{\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{nc}}, & \text{otherwise}. \end{cases} \quad (21)$$

The associated data-dependent error bounds given in Corollary 9 and Corollary 10 are then immediate.

Corollary 9 (Data-dependent bound for ℓ_p-norm regularizer and Lipschitz continuous loss w.r.t. $\|\cdot\|_\infty$)**.** *Consider the hypothesis space $H_{p,\Lambda} := H_{\tau,\Lambda}$ in* (4) *with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}, p \geq 1$. Assume that Ψ_y is L-Lipschitz continuous w.r.t. ℓ_∞-norm for any $y \in \mathcal{Y}$ and $\hat{B}_\Psi \leq 2e\hat{B}ncL$. Then, for any $0 < \delta < 1$ with probability of $1-\delta$, we have*

$$A_p \leq \frac{27L\Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2 c^{\frac{1}{2}-\frac{1}{\max(2,p)}}}{\sqrt{n}}\Big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\Big).$$

Corollary 10 (Data-dependent bound for Schatten-p norm regularizer and Lipschitz continuous loss w.r.t. ℓ_∞-norm)**.** *Let ϕ be the identity map and represent $\mathbf{w}$ by a matrix $W \in \mathbb{R}^{d\times c}$. Consider the hypothesis space $H_{S_p,\Lambda} := H_{\tau,\Lambda}$ in* (4) *with $\tau(W) = \|W\|_{S_p}, p \geq 1$. Assume that Ψ_y is L-Lipschitz continuous w.r.t. ℓ_∞-norm for any $y \in \mathcal{Y}$ and $\hat{B}_\Psi \leq 2e\hat{B}ncL$. Then, for any $0 < \delta < 1$ with probability of $1-\delta$, we have*

$$A_{S_p} \leq \begin{cases} \frac{27L\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{n}}\Big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\Big), & \text{if } p \leq 2, \\ \frac{27L\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\Big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\Big), & \text{otherwise}. \end{cases}$$

IV. Applications

In this section, we apply the general results in subsections III-C and III-D to study data-dependent error bounds for some prominent multi-class learning methods. We also compare our data-dependent bounds with the state of the art. In subsection IV-E, we present an in-depth discussion to compare error bounds based on GCs with those based on CNs.

A. Classic MC-SVMs

We first apply the results from the previous section to several classic MC-SVMs. For this purpose, we need to show that the associated loss functions satisfy Lipschitz conditions.

To this end, for any $h : \mathcal{X} \mapsto \mathbb{R}^c$, we denote by

$$\rho_h(\mathbf{x}, y) := h_y(\mathbf{x}) - \max_{y':y'\neq y} h_{y'}(\mathbf{x}) \quad (22)$$

the margin of the model h at $(\mathbf{x}, y)$. It is clear that the prediction rule h makes an error at $(\mathbf{x}, y)$ if $\rho_h(\mathbf{x}, y) < 0$. In Examples 1, 3, and 4 below, we assume that $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ is a decreasing and L_ℓ-Lipschitz function.

Example 1 (Multi-class margin-based loss [31])**.** The loss function defined as

$$\Psi_y^\ell(\mathbf{t}) := \max_{y':y'\neq y}\ell(t_y - t_{y'}), \quad \forall \mathbf{t} \in \mathbb{R}^c \quad (23)$$

is $(2L_\ell)$-Lipschitz continuous w.r.t. the ℓ_∞-norm and the ℓ_2-norm. Furthermore, we have $\ell(\rho_h(\mathbf{x},y)) = \Psi_y^\ell(h(\mathbf{x}))$.

The loss function Ψ_y^ℓ defined above in Eq. (23) is a margin-based loss function widely used in multi-class classification [31] and structured prediction [9].

Next, we study the multinomial logistic loss Ψ_y^m defined below, which is used in multinomial logistic regression [29, Chapter 4.3.4].

Example 2 (Multinomial logistic loss)**.** The multinomial logistic loss $\Psi_y^m(\mathbf{t})$ defined as

$$\Psi_y^m(\mathbf{t}) := \log\big(\sum_{j=1}^c \exp(t_j - t_y)\big), \quad \forall \mathbf{t} \in \mathbb{R}^c \tag{24}$$

is 2-Lipschitz continuous w.r.t. the ℓ_∞-norm and the ℓ_2-norm.

The loss $\tilde{\Psi}_y^\ell$ defined in Eq. (25) below is used in [32] to make pairwise comparisons among components of the predictor.

Example 3 (Loss function used in [32])**.** The loss function defined as

$$\tilde{\Psi}_y^\ell(\mathbf{t}) = \sum_{j=1}^c \ell(t_y - t_j), \quad \forall \mathbf{t} \in \mathbb{R}^c \tag{25}$$

is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving the Lipschitz constant pair $(L_\ell\sqrt{c}, L_\ell c)$ and index y. Furthermore, it is also $(2L_\ell c)$-Lipschitz continuous w.r.t. the ℓ_∞-norm.

Finally, the loss $\bar{\Psi}_y^\ell$ defined in Eq. (26) and the loss $\hat{\Psi}_y^\ell$ defined in Eq. (27) are used separately in [33] based on constrained comparisons.

Example 4 (Loss function used in [33])**.** The loss function defined as

$$\bar{\Psi}_y^\ell(\mathbf{t}) = \sum_{j=1, j\neq y}^c \ell(-t_j), \quad \forall \mathbf{t} \in \Omega = \{\tilde{\mathbf{t}} \in \mathbb{R}^c : \sum_{j=1}^c \tilde{t}_j = 0\} \tag{26}$$

is $(L_\ell\sqrt{c})$-Lipschitz continuous w.r.t. the ℓ_2-norm and $(L_\ell c)$-Lipschitz continuous w.r.t. the ℓ_∞-norm.

Example 5 (Loss function used in [45])**.** The loss function defined as

$$\hat{\Psi}_y^\ell(\mathbf{t}) = \ell(t_y), \quad \forall \mathbf{t} \in \Omega = \{\tilde{\mathbf{t}} \in \mathbb{R}^c : \sum_{j=1}^c \tilde{t}_j = 0\} \tag{27}$$

is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving the Lipschitz constant pair $(0, L_\ell)$ and index y, and L_ℓ-Lipschitz continuous w.r.t. the ℓ_∞-norm.

The following data-dependent error bounds are immediate by plugging the Lipschitz conditions established in Examples 1, 2, 3, 4 and 5 into Corollaries 3, 4, 9 and 10, separately. In the following, we always assume that the condition $\hat{B}_\Psi \leq 2e\hat{B}ncL$ holds, where L is the Lipschitz constant in Theorem 5.

Corollary 11 (Generalization bounds for Crammer and Singer MC-SVM)**.** *Consider the MC-SVM in [31] with the loss function Ψ_y^ℓ (23) and the hypothesis space H_τ with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,2}$. Let $0 < \delta < 1$. Then,*

(a) with probability of at least $1-\delta$, we have (by GCs)

$$A_2 \leq \frac{4L_\ell\Lambda\sqrt{2\pi c}}{n}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}};$$

(b) with probability of at least $1-\delta$, we have (by CNs)

$$A_2 \leq \frac{54L_\ell\Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

Analogous to Corollary 11, we have the following corollary on error bounds for the multinomial logistic regression in [29].

Corollary 12 (Generalization bounds for multinomial logistic regression)**.** *Consider the multinomial logistic regression with the loss function Ψ_y^ℓ (24) and the hypothesis space H_τ with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,2}$. Let $0 < \delta < 1$. Then,*

(a) with probability of at least $1-\delta$, we have (by GCs)

$$A_2 \leq \frac{4\Lambda\sqrt{2\pi c}}{n}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}};$$

(b) with probability of at least $1-\delta$, we have (by CNs)

$$A_2 \leq \frac{54\Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

The following three corollaries give error bounds for MC-SVMs in [32, 33, 45]. The MC-SVM in Corollary 15 is a minor variant of that in [45] with a fixed functional margin.

Corollary 13 (Generalization bounds for Weston and Watkins MC-SVM)**.** *Consider the MC-SVM in Weston and Watkins [32] with the loss function $\tilde{\Psi}_y^\ell$ (25) and the hypothesis space H_τ with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,2}$. Let $0 < \delta < 1$. Then,*

(a) with probability of at least $1-\delta$, we have (by GCs)

$$A_2 \leq \frac{4L_\ell\Lambda c\sqrt{2\pi}}{n}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}};$$

(b) with probability of at least $1-\delta$, we have (by CNs)

$$A_2 \leq \frac{54L_\ell\Lambda c\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

Corollary 14 (Generalization bounds for Lee et al. MC-SVM)**.** *Consider the MC-SVM in Lee et al. [33] with the loss function $\bar{\Psi}_y^\ell$ (26) and the hypothesis space H_τ with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,2}$. Let $0 < \delta < 1$. Then,*

(a) with probability of at least $1-\delta$, we have (by GCs)

$$A_2 \leq \frac{2L_\ell\Lambda c\sqrt{2\pi}}{n}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}};$$

(b) with probability of at least $1-\delta$, we have (by CNs)

$$A_2 \leq \frac{27L_\ell\Lambda c\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

Corollary 15 (Generalization bounds for Jenssen et al. MC-SVM)**.** *Consider the MC-SVM in Jenssen et al. [45] with the loss function $\hat{\Psi}_y^\ell$ (26) and the hypothesis space H_τ with $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,2}$. Let $0 < \delta < 1$. Then,* *(a) with probability of at least* $1-\delta$, *we have (by GCs)*

$$A_2 \leq \frac{2L_\ell \Lambda\sqrt{2\pi}}{n}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}};$$

(b) with probability of at least $1-\delta$, *we have (by CNs)*

$$A_2 \leq \frac{27L_\ell\Lambda \max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

Remark 3 (Comparison with the state of the art)**.** It is interesting to compare the above error bounds with the best known results in the literature. To start with, the data-dependent error bound of Corollary 11 (a) exhibits a square-root dependency on the number of classes, matching the state of the art from the conference version of this paper [28], which is significantly improved to a *logarithmic* dependency in Corollary 11 (b).

The error bound in Corollary 13 (a) for the MC-SVM by Weston and Watkins [32] scales linearly in c. On the other hand, according to Example 3, it is evident that $\tilde{\Psi}_y^\ell$ is $(c+\sqrt{c})L_\ell$-Lipschitz continuous w.r.t. the ℓ_2-norm, for any $y\in\mathcal{Y}$. Therefore, one can apply the structural result (3) from [43, 44] to derive the bound $O(c^{\frac{3}{2}}n^{-1}[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)]^{\frac{1}{2}})$. Furthermore, according to Example 5, $\hat{\Psi}_y^\ell$ is L_ℓ-Lipschitz continuity w.r.t. $\|\cdot\|_2$. Hence, one can apply the structural result (3) to derive the bound $O(c^{\frac{1}{2}}n^{-1}[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)]^{\frac{1}{2}})$, which is worse than the error bound $O(n^{-1}[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)]^{\frac{1}{2}})$ based on Lemma 1 and stated in Corollary 15 (a), which has no dependency on the number of classes. This justifies the effectiveness of our new structural result (Lemma 1) in capturing the Lipschitz continuity of loss functions w.r.t. a variant of the ℓ_2-norm to allow for a relatively large L_2, which is exactly the case for some popular MC-SVMs [30, 32, 45].

Note that for the MC-SVMs by Weston and Watkins [32], Lee et al. [33], Jenssen et al. [45], the GC-based error bounds are tighter than the corresponding error bounds based on CNs, up to logarithmic factors.

B. Top-k MC-SVM

Motivated by the ambiguity in class labels caused by the rapid increase in number of classes in modern computer vision benchmarks, Lapin et al. [30, 63] introduce the top-k MC-SVM by using the top-k hinge loss to allow k predictions for each object $\mathbf{x}$. For any $\mathbf{t}\in\mathbb{R}^c$, let the brackets $[\cdot]$ denote a permutation such that $[j]$ is the index of the j-th largest score, i.e., $t_{[1]}\geq t_{[2]}\geq\cdots\geq t_{[c]}$.

Example 6 (Top-k hinge loss [30])**.** The top-k hinge loss defined for any $\mathbf{t}\in\mathbb{R}^c$

$$\Psi_y^k(\mathbf{t}) = \max\Big\{0, \frac{1}{k}\sum_{j=1}^k(1_{y\neq 1}+t_1-t_y,\ldots,1_{y\neq c}+t_c-t_y)_{[j]}\Big\} \tag{28}$$

is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving a Lipschitz constant pair $\big(\frac{1}{\sqrt{k}},1\big)$ and index y. Furthermore, it is also 2-Lipschitz continuous w.r.t. the ℓ_∞-norm.

With the Lipschitz conditions established in Example 6, we can now give the generalization error bounds for the top-k MC-SVM [30].

Corollary 16 (Generalization bounds for top-k MC-SVM)**.** *Consider the top-k MC-SVM with the loss functions* (28) *and the hypothesis space* H_τ *with* $\tau(\mathbf{w})=\|\mathbf{w}\|_{2,2}$. *Let* $0<\delta<1$. *Then,*

(a) with probability of at least $1-\delta$, *we have (by GCs)*

$$A_2 \leq \frac{2\Lambda\sqrt{2\pi}}{n}(c^{\frac{1}{2}}k^{-\frac{1}{2}}+1)\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}};$$

(b) with probability of at least $1-\delta$, *we have (by CNs)*

$$A_2 \leq \frac{54\Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

Remark 4 (Comparison with the state of the art)**.** An appealing property of Corollary 16 (a) is the involvement of the factor $k^{-\frac{1}{2}}$. Note that we even can get error bounds with no dependencies on c if we choose $k>\tilde{C}c$ for a universal constant $\tilde{C}$.

Comparing our result to the state of the art, it follows again from Example 6 that Ψ_y^k is $(1+k^{-\frac{1}{2}})$-Lipschitz continuous w.r.t. the ℓ_2-norm for all $y\in\mathcal{Y}$. Using the structural result (3) [28, 43, 44], one can derive an error bound decaying as $O\big(n^{-1}c^{\frac{1}{2}}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}}\big)$, which is suboptimal to Corollary 16 (a) since it does not shed insight on how the parameter k would affect the generalization performance. Furthermore, the error bound in Corollary 16 (b) enjoys a logarithmic dependency on the number of classes.

C. ℓ_p-norm MC-SVM

In our previous work [28], we introduce the ℓ_p-norm MC-SVM as an extension of the Crammer & Singer MC-SVM by replacing the associated ℓ_2-norm regularizer with a general block $\ell_{2,p}$-norm regularizer [28]. We establish data-dependent error bounds in [28], showing a logarithmic dependency on the number of classes as p decreases to 1. The present analysis yields the following bounds, which also hold for the MC-SVM with the multinomial logistic loss and the block $\ell_{2,p}$-norm regularizer.

Corollary 17 (Generalization bounds for ℓ_p-norm MC-SVM)**.** *Consider the ℓ_p-norm MC-SVM with loss function* (23) *and the hypothesis space* H_τ *with* $\tau(\mathbf{w})=\|\mathbf{w}\|_{2,p}, p\geq 1$. *Let* $0<\delta<1$. *Then,*

(a) with probability of at least $1-\delta$, *we have (by GCs):*

$$A_p \leq \frac{4L_\ell\Lambda\sqrt{\pi}}{n}\big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big]^{\frac{1}{2}}\inf_{q\geq p}[(q^*)^{\frac{1}{2}}c^{\frac{1}{q^*}}];$$

(b) with probability of at least $1-\delta$, *we have (by CNs):*

$$A_p \leq \frac{54L_\ell\Lambda\max\limits_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2 c^{\frac{1}{2}-\frac{1}{\max(2,p)}}}{\sqrt{n}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big).$$

Remark 5 (Comparison with the state of the art)**.** Corollary 17 (a) is an extension of error bounds in the conference version [28] from $1\leq p\leq 2$ to the case $p\geq 1$. We can see how p affects the generalization performance of ℓ_p-norm MC-SVM. The function $f:\mathbb{R}_+\mapsto\mathbb{R}_+$ defined by $f(t)=t^{\frac{1}{2}}c^{\frac{1}{t}}$ is monotonically decreasing on the interval $(0,2\log c)$ and increasing on the interval $(2\log c,\infty)$. Therefore, the data-dependent error bounds in Corollary 17 (a) transfer to

$$A_p\leq\begin{cases}4\Lambda L_\ell\sqrt{\pi p^*}n^{-1}c^{1-\frac{1}{p}}\Big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{1}{2}}, & \text{if } p>\frac{2\log c}{2\log c-1},\\ 4\Lambda L_\ell(2\pi e\log c)^{\frac{1}{2}}n^{-1}\Big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{1}{2}}, & \text{otherwise.}\end{cases}$$

That is, the dependency on the number of classes would be polynomial with exponent $1/p^*$ if $p>\frac{2\log c}{2\log c-1}$ and logarithmic otherwise. On the other hand, the error bounds in Corollary 17 (b) significantly improve those in Corollary 17 (a). Indeed, the error bounds in Corollary 17 (b) enjoy a logarithmic dependency on the number of classes if $p\leq 2$ and a polynomial dependency with exponent $\frac{1}{2}-\frac{1}{p}$ otherwise (up to logarithmic factors). This phase transition phenomenon at $p=2$ is explained in Remark 2. It is also clear that error bounds based on CNs outperform those based on GCs by a factor of $\sqrt{c}$ for $p\geq 2$ (up to logarithmic factors), which, as we will explain in subsection IV-E, is due to the use of the Lipschitz continuity measured by a norm suitable to the loss function.

D. Schatten-p Norm MC-SVM

Amit et al. [57] propose to use trace-norm regularization in multi-class classification to uncover shared structures that always exist in the learning regime with many classes. Here we consider error bounds for the more general Schatten-p norm MC-SVM.

Corollary 18 (Generalization bounds for Schatten-p norm MC-SVM)**.** *Let ϕ be the identity map and represent $\mathbf{w}$ by a matrix $W\in\mathbb{R}^{d\times c}$. Consider Schatten-$p$ norm MC-SVM with loss functions* (23) *and the hypothesis space H_τ with $\tau(W)=\|W\|_{S_p},p\geq 1$. Let $0<\delta<1$. Then,*

(a) with probability of at least $1-\delta$, we have (by GCs):

$$A_{S_p}\leq\begin{cases}\frac{2^{\frac{7}{4}}\pi\Lambda L_\ell}{n\sqrt{e}}\inf_{p\leq q\leq 2}(q^*)^{\frac{1}{2}}\Big[c^{\frac{1}{q^*}}\big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\big]^{\frac{1}{2}}\\ \quad +c^{\frac{1}{2}}\|\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\|_{S_{\frac{q^*}{2}}}^{\frac{1}{2}}\Big], & \text{if } p\leq 2,\\ \frac{2^{\frac{9}{4}}\pi\Lambda L_\ell c^{\frac{1}{2}}\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{n\sqrt{e}}\big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\big]^{\frac{1}{2}}, & \text{otherwise.}\end{cases}$$

(b) with probability of at least $1-\delta$, we have (by CNs):

$$A_{S_p}\leq\begin{cases}\frac{54L_\ell\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{n}}\Big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\Big), & \text{if } p\leq 2,\\ \frac{54L_\ell\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\Big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\Big), & \text{otherwise.}\end{cases}$$

Remark 6 (Analysis of Schatten-p norm MC-SVM)**.** Analogous to Remark 5, error bounds of Corollary 18 (a) transfer to

$$\begin{cases}O\Big(n^{-1}(p^*)^{\frac{1}{2}}\big(c^{\frac{1}{p^*}}\big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\big]^{\frac{1}{2}}+c^{\frac{1}{2}}\|\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\|_{S_{\frac{p^*}{2}}}^{\frac{1}{2}}\big)\Big), & \text{if } 2\leq p^*\leq 2\log c,\\ O\Big(n^{-1}\sqrt{\log c}\big(\big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\big]^{\frac{1}{2}}+c^{\frac{1}{2}}\|\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\|_{S_{\log c}}^{\frac{1}{2}}\big)\Big), & \text{if } 2<2\log c<p^*,\\ O\Big(n^{-1}c^{1-\frac{1}{p}}\big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\big]^{\frac{1}{2}}\Big), & \text{if } p>2.\end{cases}$$

As a comparison, error bounds in Corollary 18 (b) would decay as $O(n^{-\frac{1}{2}}\log^{\frac{3}{2}}(n^{\frac{3}{2}}c))$ if $p\leq 2$ and $O\big(n^{-\frac{1}{2}}c^{\frac{1}{2}-\frac{1}{p}}\log^{\frac{3}{2}}(n^{\frac{3}{2}}c)\big)$ otherwise, which significantly outperform those in Corollary 18 (a).

E. Comparison of the GC and the CN Approach

In this paper, we develop two methods to derive data-dependent error bounds that are applicable to learning with many classes. We summarize these two types of error bounds for some specific MC-SVMs in the third and fourth columns of Table II, from which it is clear that each approach can yield better bounds than the other for some MC-SVMs. For example, for multinomial logistic regression and the Crammer & Singer MC-SVM, the GC-based error bound has a square-root dependency on the number of classes, whereas the CN-based bound has a logarithmic dependency. CN-based error bounds also have significant advantages for ℓ_p-norm MC-SVM and Schatten-p norm MC-SVM. On the other hand, GC-based analyses have their own advantages. First, for the MC-SVMs in Weston and Watkins [32], Lee et al. [33], the GC-based error bounds decay as $O(n^{-\frac{1}{2}}c)$, while the CN-based bounds decay as $O(n^{-\frac{1}{2}}c\log^{\frac{3}{2}}(nc))$. Second, the GC-based error bounds involve a summation of $K(\mathbf{x}_i,\mathbf{x}_i)$ over training examples, while the CN-based error bounds involve a maximum of $\|\phi(\mathbf{x}_i)\|_i$ over the training examples. In this sense, the GC-based error bounds better capture the properties of the distribution from which the training examples are drawn.

An in-depth discussion can explain the mismatch between these two types of generalization error bounds. Our GC-based bounds are based on a structural result (Lemma 1) of empirical GCs to exploit the Lipschitz continuity of loss functions w.r.t. a variant of the ℓ_2-norm, while our CN-based analysis is based on a structural result of empirical ℓ_∞-norm CNs to directly use the Lipschitz continuity of loss functions w.r.t. the ℓ_∞-norm. Which approach is better depends on the Lipschitz continuity of the associated loss functions. Specifically, if Ψ_y is Lipschitz continuous w.r.t. a variant of the ℓ_2-norm involving the Lipschitz constant pair (L_1,L_2) and is L-Lipschitz continuous w.r.t. the ℓ_∞-norm, then one can show the following inequality with probability of at least $1-\delta$ for $\delta\in(0,1)$ (Theorem 2 and Theorem 6, respectively)

$$A_\tau\leq\begin{cases}2\sqrt{\pi}\Big[L_1c\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau)+L_2\mathfrak{G}_{\widetilde{S}'}(\widetilde{H}_\tau)\Big]\ \text{(by GCs)}, & \text{(29a)}\\ 27L\sqrt{c}\mathfrak{R}_{nc}(\widetilde{H}_\tau)\Big(1+\log_2^{\frac{3}{2}}\frac{\hat{B}n\sqrt{c}}{\mathfrak{R}_{nc}(\widetilde{H}_\tau)}\Big)\text{(by CNs)}. & \text{(29b)}\end{cases}$$

It is reasonable to assume that $\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau)$ and $\mathfrak{R}_{nc}(\widetilde{H}_\tau)$ decay at the same order. For example, if $\tau(\mathbf{w})=\|\mathbf{w}\|_{2,p},p\geq 2$, then one can show (the first inequality follows from (39), (40) and (41), and the second inequality follows from Proposition 7)

$$\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau)=O\Big(n^{-1}c^{-\frac{1}{p}}\big(\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big)^{\frac{1}{2}}\Big),$$
$$\mathfrak{R}_{nc}(\widetilde{H}_\tau)=O\Big(n^{-\frac{1}{2}}c^{-\frac{1}{p}}\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2\Big).$$

We further assume that the dominant term in (29a) is $L_1c\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau)$ to clearly illustrate the relative behavior of these

TABLE II
COMPARISON OF DATA-DEPENDENT GENERALIZATION ERROR BOUNDS DERIVED IN THIS PAPER. We use the notation $B_1 = \big(\frac{1}{n}\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\big)^{\frac{1}{2}}$ and $B_\infty = \max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2$. The best bound for each MC-SVM is followed by a bullet.

MC-SVM	by structural result (3)	by GCs	by CNs
Crammer & Singer	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_\infty n^{-\frac{1}{2}}\log^{\frac{3}{2}}(nc))$ •
Multinomial Logistic	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_\infty n^{-\frac{1}{2}}\log^{\frac{3}{2}}(nc))$ •
Weston and Watkins	$O(B_1 n^{-\frac{1}{2}}c^{\frac{3}{2}})$	$O(B_1 n^{-\frac{1}{2}}c)$ •	$O(B_\infty n^{-\frac{1}{2}}c\log^{\frac{3}{2}}(nc))$
Lee et al.	$O(B_1 n^{-\frac{1}{2}}c)$ •	$O(B_1 n^{-\frac{1}{2}}c)$ •	$O(B_\infty n^{-\frac{1}{2}}c\log^{\frac{3}{2}}(nc))$
Jenssen et al.	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_1 n^{-\frac{1}{2}})$ •	$O(B_\infty n^{-\frac{1}{2}}\log^{\frac{3}{2}}(nc))$
top-k	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_1 n^{-\frac{1}{2}}(ck^{-1})^{\frac{1}{2}})$	$O(B_\infty n^{-\frac{1}{2}}\log^{\frac{3}{2}}(nc))$ •
ℓ_p-norm $p\in(1,\infty)$	$O(B_1 n^{-\frac{1}{2}}c^{1-\frac{1}{p}})$	$O(B_1 n^{-\frac{1}{2}}c^{1-\frac{1}{p}})$	$O(B_\infty n^{-\frac{1}{2}}c^{\frac{1}{2}-\frac{1}{\max(2,p)}}\log^{\frac{3}{2}}(nc))$ •
Schatten-p $p\in[1,2)$	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_1 n^{-\frac{1}{2}}c^{\frac{1}{2}})$	$O(B_\infty n^{-\frac{1}{2}}\log^{\frac{3}{2}}(nc))$ •
Schatten-p $p\in[2,\infty)$	$O(B_1 n^{-\frac{1}{2}}c^{1-\frac{1}{p}})$	$O(B_1 n^{-\frac{1}{2}}c^{1-\frac{1}{p}})$	$O(B_\infty n^{-\frac{1}{2}}c^{\frac{1}{2}-\frac{1}{p}}\log^{\frac{3}{2}}(nc))$ •

two types of error bounds. If L_1 and L are of the same order, as exemplified by Example 1 and Example 2, then the error bounds based on CNs outperform those based on GCs by a factor of $\sqrt{c}$ (up to logarithmic factors). If $L_1 = O(c^{-\frac{1}{2}}L)$, as exemplified by Example 3, Example 4 and Example 5, then the error bounds based on GCs outperform those based on CNs by a factor of $\log^{\frac{3}{2}}(nc)$. The underlying reason is that the Lipschitz continuity w.r.t. $\|\cdot\|_2$ is a stronger assumption than that w.r.t. $\|\cdot\|_\infty$ in the magnitude of Lipschitz constants. Indeed, if Ψ_y is L_1-Lipschitz continuous w.r.t. $\|\cdot\|_2$, then one may expect that Ψ_y is $(L_1\sqrt{c})$-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ due to the inequality $\|\mathbf{t}\|_2\le\sqrt{c}\|\mathbf{t}\|_\infty$ for any $\mathbf{t}\in\mathbb{R}^c$. This explains why (29b) outperforms (29a) by a factor of $\sqrt{c}$ if we ignore the Lipschitz constants. To summarize, if $L_1 = O(c^{-\frac{1}{2}}L)$, then (29a) outperforms (29b). Otherwise, (29b) is better. Therefore, one should choose an appropriate approach according to the associated loss function to exploit the inherent Lipschitz continuity.

We also include the error bounds based on the structural result (3) in the second column to demonstrate the advantages of the structural result based on the variant of the ℓ_2-norm over (3).

V. EXPERIMENTS

In this section, we report experimental results to show the effectiveness of our theory. We consider the ℓ_p-norm MC-SVM with multinomial logistic loss $\Psi_y(\mathbf{t}) = \Psi_y^m(\mathbf{t})$ defined in Example 2 and hypothesis space H_τ, where $\tau(\mathbf{w}) = \|\mathbf{w}\|_{2,p}, p\ge 1$ and $\phi(\mathbf{x}) = \mathbf{x}$. In subsection V-A, we aim to show that our error bounds capture well the effects of the number of classes on the generalization performance. In subsection V-B, we aim to show that our error analysis is able to imply a structural risk that works well in model selection, as well as the efficiency of ℓ_p-norm MC-SVM. We use several benchmark datasets in our experiments: MNIST [64], NEWS20 [65], LETTER [3], RCV1 [66], SECTOR [67] and ALOI [68]. For ALOI, we include the first 67% of the instances of each class in the training dataset and use the remaining instances as the test dataset. Table III gives some information on these datasets, which can be downloaded from the LIBSVM website [69].

TABLE III
DESCRIPTION OF THE DATASETS USED IN THE EXPERIMENTS.

Dataset	c	n	# Test Examples	d
MNIST	10	60,000	10,000	778
NEWS20	20	15,935	3,993	62,060
LETTER	26	10,500	5,000	16
RCV1	53	15,564	518,571	47,236
SECTOR	105	6,412	3,207	55,197
ALOI	1,000	72,000	36,000	128

A. Empirical verification of generalization bounds

According to the proof of Corollary 17 (b), we know

$$
\begin{aligned}
&\mathrm{GAP}(\mathbf{w}_{p,\Lambda}) := \mathbb{E}_{\mathbf{x},y}\Psi_y(h^{\mathbf{w}_{p,\Lambda}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n\Psi_{y_i}(h^{\mathbf{w}_{p,\Lambda}}(\mathbf{x}_i))\\
&\le \sup_{h^{\mathbf{w}}\in H_\tau}\Big[\mathbb{E}_{\mathbf{x},y}\Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n\Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big]\\
&= O(1)\mathfrak{R}_S(F_{\tau,\Lambda}) = O(\Lambda n^{-\frac{1}{2}}c^{\frac{1}{2}-\frac{1}{\max(2,p)}}\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2\log^{\frac{3}{2}}(nc)),
\end{aligned}
$$

where the trained model $\mathbf{w}_{p,\Lambda}$ associated with a pair (p,Λ) is defined by

$$
\mathbf{w}_{p,\Lambda} := \arg\min_{\substack{\mathbf{w}\in\mathbb{R}^{d\times c}\\ \|\mathbf{w}\|_{2,p}\le\Lambda}}\frac{1}{n}\sum_{i=1}^n\Psi_{y_i}^m\big(\langle\mathbf{w}_1,\mathbf{x}_i\rangle,\ldots,\langle\mathbf{w}_c,\mathbf{x}_i\rangle\big). \tag{30}
$$

Note that GAP measures the difference between the generalization error and the empirical error for the *particular* learned model, which is the quantity we are interested in. For comparison, $\mathfrak{R}_S(F_{\tau,\Lambda})$ controls the *uniform* deviation between generalization errors and empirical errors over the hypothesis space and is a standard tool used to control GAPs [37, 47]. Our purpose here is to validate whether our bounds capture the dependency of $\mathfrak{R}_S(F_{\tau,\Lambda})$ and GAPs on the number of classes in practice. To this aim, we first discuss how to approximate $\mathfrak{R}_S(F_{\tau,\Lambda})$ and GAPs.

Approximation of $\mathfrak{R}_S(F_{\tau,\Lambda})$. We approximate $\mathfrak{R}_S(F_{\tau,\Lambda})$ by an *Approximation of the Empirical Rademacher Complexity* (AERC) defined by $\mathrm{AERC}(F_{\tau,\Lambda}) := \frac{1}{50}\sum_{t=1}^{50}\widetilde{\mathfrak{R}}_S(\boldsymbol{\epsilon}^{(t)},F_{\tau,\Lambda})$, where $\boldsymbol{\epsilon}^{(t)} = \{\epsilon_i^{(t)}\}_{i\in\mathbb{N}_n}, t = 1,\ldots,50$, are independent sequences of independent Rademacher random variables and

$$\widetilde{\mathfrak{R}}_S(\boldsymbol{\epsilon}, F_{\tau,\Lambda}) := \frac{1}{n}\sup_{\substack{\mathbf{w}\in\mathbb{R}^{d\times c}\\ \|\mathbf{w}\|_{2,p}\leq\Lambda}}\sum_{i=1}^{n}\epsilon_i\Psi_{y_i}^m\big(\langle\mathbf{w}_1,\mathbf{x}_i\rangle,\ldots,\langle\mathbf{w}_c,\mathbf{x}_i\rangle\big). \tag{31}$$

It can be checked that $\widetilde{\mathfrak{R}}_S(\boldsymbol{\epsilon}, F_{\tau,\Lambda})$ (as a function of $\boldsymbol{\epsilon}$) satisfies the increment condition (36) in McDiarmid's inequality below and concentrates sharply around its expectation $\mathfrak{R}_S(F_{\tau,\Lambda})$. Therefore, AERC is a good approximation of $\mathfrak{R}_S(F_{\tau,\Lambda})$. The calculation of AERC involves the constrained non-convex optimization problem (31), which we solve by the classic Frank-Wolfe algorithm [58, 70]. We describe the Frank-Wolfe algorithm to solve $\min_{\mathbf{w}\in\triangle_p} f(\mathbf{w})$ for a general function f defined on the feasible set $\triangle_p = \{\mathbf{w}\in\mathbb{R}^{d\times c} : \|\mathbf{w}\|_{2,p}\leq\Lambda\}$ with $p\geq 1$ and $\Lambda > 0$ in Algorithm 1. This is a projection-free method that involves a constrained linear optimization problem at each iteration, which, as shown in the following proposition, has a closed-form solution. In line 4 of Algorithm 1, we use a backtracking line search to find the step size γ satisfying the *Armijo condition* (e.g., page 33 in [71]). Proposition 19 can be proved by checking $\|\mathbf{w}^*\|_{2,p}\leq 1$ and $\langle\mathbf{w}^*,\mathbf{v}\rangle = -\|\mathbf{v}\|_{2,p^*}$, which is deferred to Appendix C.

Algorithm 1: Frank-Wolfe Algorithm

```
1 Let k = 0 and w^(0) = 0 ∈ R^{d×c}
2 while Optimality conditions are not satisfied do
3     Compute w̃ = arg min_{w:‖w‖_{2,p}≤Λ} ⟨w, ∇f(w^(k))⟩
4     Calculate the direction v = w̃ − w^(k) and step size
        γ ∈ [0, 1]
5     Update w^(k+1) = w^(k) + γv
6     Set k = k + 1
7 end
```

Proposition 19. *Let $\mathbf{v} = (\mathbf{v}_1,\ldots,\mathbf{v}_c)\in\mathbb{R}^{d\times c}$ have nonzero column vectors and $p\geq 1$. Then the problem*

$$\arg\min_{\mathbf{w}\in\mathbb{R}^{d\times c}}\langle\mathbf{w},\mathbf{v}\rangle \quad \textit{s.t. } \|\mathbf{w}\|_{2,p}\leq 1 \tag{32}$$

has a closed-form solution $\mathbf{w}^* = (\mathbf{w}_1^*,\ldots,\mathbf{w}_c^*)$ as follows

$$\mathbf{w}_j^* = \begin{cases} -\mathbf{v}_{\bar{j}}\|\mathbf{v}_{\bar{j}}\|_2^{-1}, & \textit{if } p=1 \textit{ and } j=\bar{j},\\ 0, & \textit{if } p=1 \textit{ and } j\neq\bar{j},\\ -\frac{\|\mathbf{v}_j\|_2^{p^*-2}\mathbf{v}_j}{\big(\sum_{\tilde{j}=1}^{c}\|\mathbf{v}_{\tilde{j}}\|_2^{p^*}\big)^{\frac{1}{p}}}, & \textit{if } 1<p<\infty,\\ -\|\mathbf{v}_j\|_2^{-1}\mathbf{v}_j, & \textit{if } p=\infty, \end{cases} \tag{33}$$

where $\bar{j}$ is the smallest index satisfying $\|\mathbf{v}_{\bar{j}}\|_2 = \max_{\tilde{j}\in\mathbb{N}_c}\|\mathbf{v}_{\tilde{j}}\|_2$ and $p^* = p/(p-1)$.

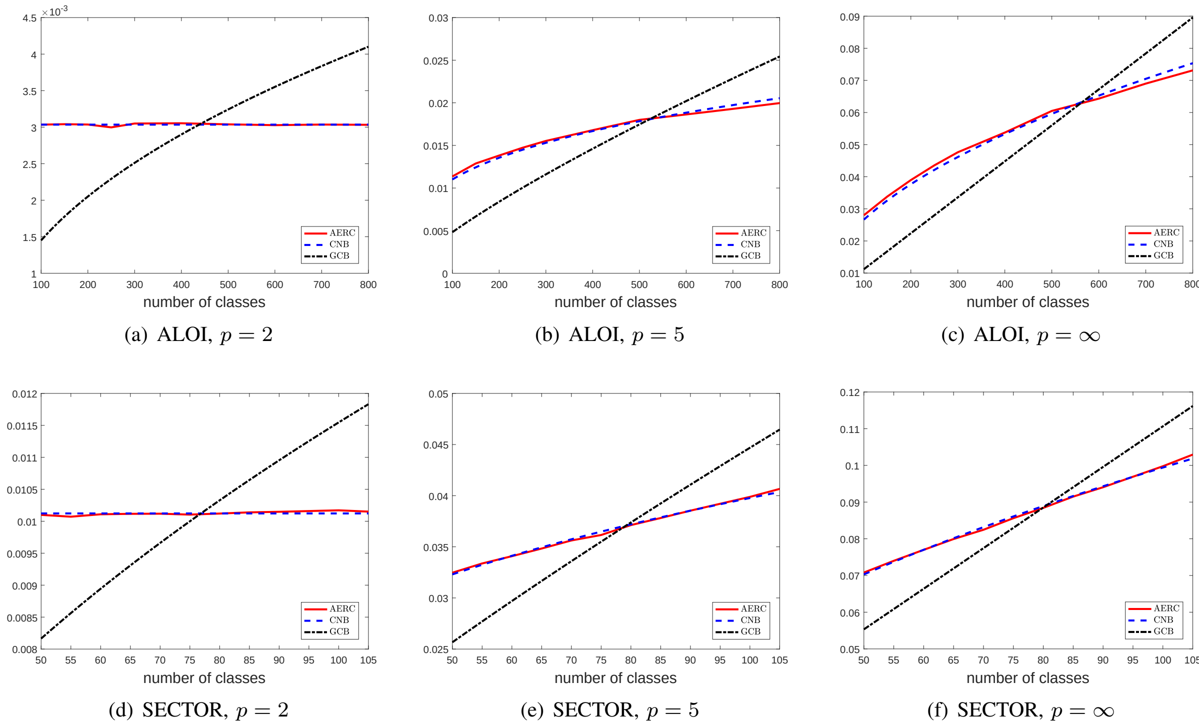
Estimation of GAPs. To calculate GAPs, we need to solve the convex optimization problem (30), which is solved by introducing class weights and alternating the update w.r.t. class weights and the update w.r.t. the model $\mathbf{w}$ in [28]. In this paper, we propose to solve this optimization problem with the Frank-Wolfe algorithm (Algorithm 1), which avoids the introduction of additional class weights and extends the algorithm in [28] to the case of $p > 2$. The closed-form solution established in Proposition 19 makes the implementation of this algorithm simple and efficient for training ℓ_p-norm MC-SVM.

Behavior with respect to the number of classes. We now show that our generalization bounds capture the dependency of AERCs and GAPs on the number of classes. To this aim, we need to construct several datasets with different numbers of classes. We fix the input $\{\mathbf{x}_i\}_{i=1}^n$ of either ALOI or SECTOR, the parameter p and $\Lambda = 1$, and vary the number of classes $\tilde{c}$ over the set $\{100,150,200,250,300,350,400,500,600,700,800\}$ (ALOI) or $\{50,55,60,65,70,75,80,85,90,100,105\}$ (SECTOR). For each $\tilde{c}$ and dataset, we create a dataset with $\tilde{c}$ classes as $S^{(\tilde{c})} = \{(\mathbf{x}_i, y_i^{(\tilde{c})})\}_{i=1}^n$, where $y_i^{(\tilde{c})} = \lceil y_i\tilde{c}/c\rceil$, y_i is the i-th output and $\lceil a\rceil$ denotes the least integer not smaller than a. Note that this strategy of grouping class labels may affect the meaning of labels and further influence the classification quality. However, it is reasonable here since we are interested in the behavior of AERCs and GAPs w.r.t. the number of classes. For each $\tilde{c}$, we can calculate the corresponding AERCs and GAPs. We repeat the experiment 50 times and report the average of the experimental results. We plot AERCs and GAPs as functions of $\tilde{c}$ in Fig. 2 and Fig. 3, respectively, for $p = 2, 5, \infty$. In each of these panels, we also include plots of the function $\mathrm{CNB}_\tau(\tilde{c}) = \tau\tilde{c}^{\frac{1}{2}-\frac{1}{\max(2,p)}}$ and $\mathrm{GCB}_{\tilde{\tau}}(\tilde{c}) = \tilde{\tau}\tilde{c}^{1-\frac{1}{p}}$, where the corresponding parameters τ and $\tilde{\tau}$ are computed by fitting the AERCs/GAPs with models $\{\tilde{c}\mapsto\mathrm{CNB}_\tau(\tilde{c}) : \tau\in\mathbb{R}_+\}$ and $\{\tilde{c}\mapsto\mathrm{GCB}_{\tilde{\tau}}(\tilde{c}) : \tilde{\tau}\in\mathbb{R}_+\}$, respectively. Note that the CNBs and GCBs are constructed based on CN analysis and GC analysis, as listed in Table II (we ignore logarithmic factors here).

According to Fig. 2, we see clearly that AERCs match very well with the CNB plot, which indicates that our CN-based analysis captures the dependency of the generalization performance on the number of classes. By comparison, there is a clear discrepancy between the AERC and GCB plots, indicating a crudeness of the GC-based analysis. Furthermore, AERCs behave nearly as constants in the case of $p = 2$, which is consistent with the almost class-size independent bounds based on CN analysis for $p = 2$ (up to a logarithmic factor). One can see a similar phenomenon in Fig. 3: CNBs behave much better than GCBs in fitting the GAPs. It should be mentioned that the fitting of GAPs by CNBs is not as perfect as the fitting of AERCs by CNBs. The underlying reason is as follows. Our generalization bounds directly apply to $\mathfrak{R}_S(F_{\tau,\Lambda})$ which controls the *uniform* deviation between generalization errors and empirical errors over all $\mathbf{w}\in H_\tau$, whereas GAPs correspond to the deviation for the *particular* trained model $\mathbf{w}_{p,\Lambda}$. Nevertheless, as shown in Fig. 3, CNBs already capture well the behavior of GAPs as a function of the class size, which justifies the usefulness of our theoretical analysis since it is the trained model $\mathbf{w}_{p,\Lambda}$ that we are most interested in for practical learning processes.

B. Behavior of the ℓ_p-norm MC-SVM and model selection

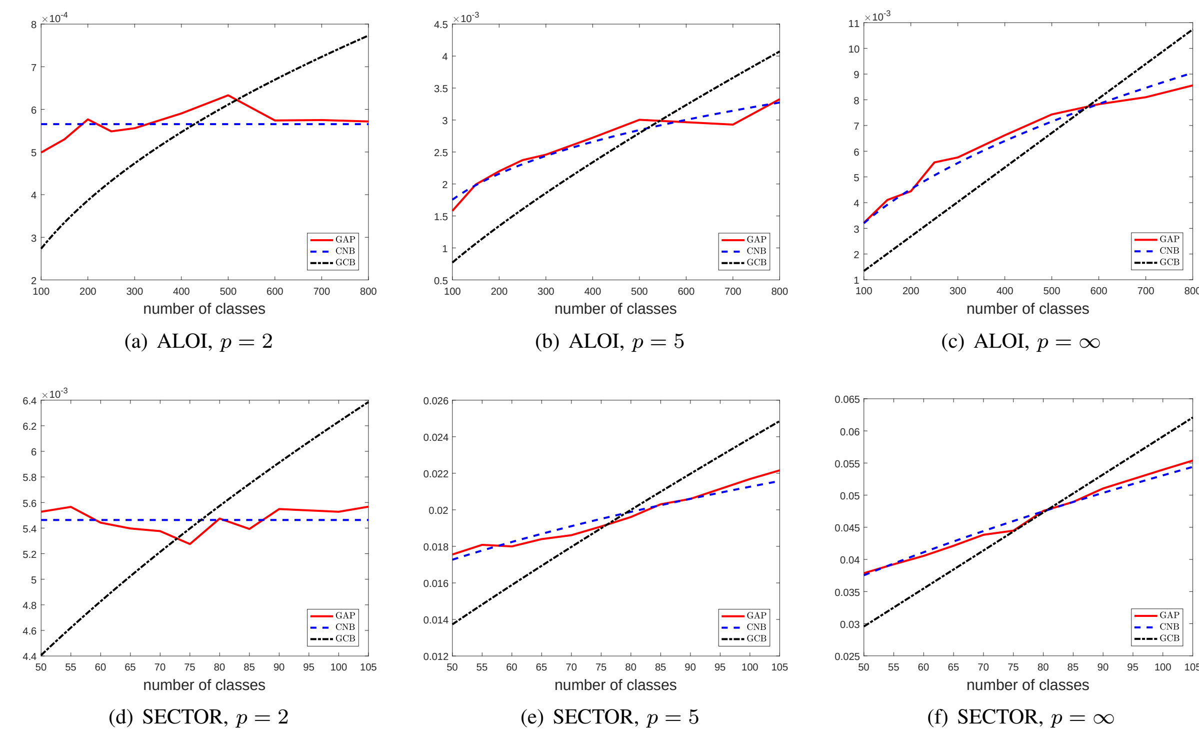
In this section, we describe the application of our error bounds in model selection, as well as the effectiveness of the



(a) ALOI, $p=2$ (b) ALOI, $p=5$ (c) ALOI, $p=\infty$

(d) SECTOR, $p=2$ (e) SECTOR, $p=5$ (f) SECTOR, $p=\infty$

Fig. 2. AERCs as a function of the number of classes. Based on ALOI or SECTOR, we construct datasets with a varying number of classes $\tilde{c}$, for each of which we compute the associated AERC. We also include plots of $\mathrm{CNB}_\tau(\tilde{c})$ and $\mathrm{GCB}_{\bar{\tau}}(\tilde{c})$ in this figure, where both τ and $\bar{\tau}$ are calculated by applying the least-squares method to fit these AERCs with $\mathrm{CNB}_\tau(\tilde{c})$ and $\mathrm{GCB}_{\bar{\tau}}(\tilde{c})$, respectively. Each panel corresponds to a specific dataset and a parameter p.



(a) ALOI, $p=2$ (b) ALOI, $p=5$ (c) ALOI, $p=\infty$

(d) SECTOR, $p=2$ (e) SECTOR, $p=5$ (f) SECTOR, $p=\infty$

Fig. 3. GAPs as a function of the number of classes. Based on ALOI or SECTOR, we construct datasets with a varying number of classes $\tilde{c}$, for each of which we compute the associated GAP. We also include plots of $\mathrm{CNB}_\tau(\tilde{c})$ and $\mathrm{GCB}_{\bar{\tau}}(\tilde{c})$ in this figure, where both τ and $\bar{\tau}$ are calculated by applying the least squares method to fit these GAPs with $\mathrm{CNB}_\tau(\tilde{c})$ and $\mathrm{GCB}_{\bar{\tau}}(\tilde{c})$, respectively. Each panel corresponds to a specific dataset and a parameter p.

ℓ_p-norm MC-SVM as compared to multinomial logistic regression (MLR) [29] and the Weston & Watkins (WW) MC-SVM in Corollary 13 with $\ell(t)=\log(1+\exp(-t))$. We traverse p over the set $\{1,1.2,1.5,1.8,2,2.33,2.5,2.67,3,4,8,\infty\}$ and Λ over the set $\{10^{0.5},10,10^{1.5},\ldots,10^{3.5}\}$. For each pair (p,Λ), we train the model $\mathbf{w}_{p,\Lambda}$ defined in (30) by Algorithm 1 as candidate models, and compute the accuracy (the percent of instances labeled correctly) on the test examples. We also train a model by MLR and a model by WW MC-SVM for each candidate Λ. Our aim is to identify an appropriate model from these candidate models based on our generalization analysis, and to compare the behavior of MLR, ℓ_p-norm MC-SVM and WW MC-SVM on several datasets.

Model selection strategy. Since $\mathbf{w}_{p,\Lambda}\in H_{\tilde{p},\|\mathbf{w}_{p,\Lambda}\|_{2,\tilde{p}}}$ for any $\tilde{p}\geq 1$, one can derive from Corollary 17 the following inequality with probability of $1-\delta$ (here we omit the randomness of $\|\mathbf{w}_{p,\Lambda}\|_{2,\tilde{p}}$ for brevity)

$$\mathbb{E}_{\mathbf{x},y}\Psi_y(h^{\mathbf{w}_{p,\Lambda}}(\mathbf{x}))\leq\frac{1}{n}\sum_{i=1}^n\Psi_{y_i}(h^{\mathbf{w}_{p,\Lambda}}(\mathbf{x}_i))+3B_\Psi\Big[\frac{\log\frac{4}{\delta}}{2n}\Big]^{\frac{1}{2}}+$$
$$54\|\mathbf{w}_{p,\Lambda}\|_{2,\tilde{p}}\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2 c^{\frac{1}{2}-\frac{1}{\max(2,\tilde{p})}}\big(1+\log_2^{\frac{3}{2}}\big(\sqrt{2}n^{\frac{3}{2}}c\big)\big)/\sqrt{n}.$$

According to the inequality $\|\mathbf{w}\|_{2,2}\leq\|\mathbf{w}\|_{2,\tilde{p}}c^{\frac{1}{2}-\frac{1}{\tilde{p}}}$ for any $\tilde{p}\geq 2$, the term $\|\mathbf{w}\|_{2,\tilde{p}}c^{\frac{1}{2}-\frac{1}{\max(2,\tilde{p})}}$ attains its minimum at $\tilde{p}=2$. Hence, we construct the following structural risk (ignoring logarithmic factors here)

$$\mathrm{Err}_{\mathrm{str},\lambda}(\mathbf{w}):=\frac{1}{n}\sum_{i=1}^n\Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))+\lambda\|\mathbf{w}\|_{2,2}\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2/\sqrt{n} \tag{34}$$

and use it to select a model with the minimal structural risk among all candidates $\mathbf{w}_{p,\Lambda}$. According to Table II, we construct a different structural risk for WW MC-SVM with the penalty being $\lambda c\|\mathbf{w}\|_{2,2}\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2/\sqrt{n}$. We use $\lambda=1/25$ in this paper.

In Table IV, we report the accuracies of MLR, ℓ_p-norm MC-SVM and WW MC-SVM on several benchmark datasets. For each method, we report the best accuracy achieved by the candidate model and the accuracy of the model selected from these candidate models with the minimal structural risk, as shown in the columns termed "Oracle" and "Model selection", respectively. For ℓ_p-norm MC-SVM, we also report the parameter p at which the corresponding accuracy is achieved.

According to Table IV, our structural risk based on generalization analysis behaves well in guiding the selection of a model with comparable prediction accuracy to the best candidate model. For ℓ_p-norm MC-SVM, the accuracies for the model selected according to (34) and the best candidate model differ by less than 0.17% on all datasets. ℓ_p-norm MC-SVM consistently outperforms both MLR and WW MC-SVM. For example, for ALOI and the model selection strategy, ℓ_p-norm MC-SVM achieves an accuracy of 88.48%, while MLR and WW MC-SVM achieve accuracies of 85.70% and 78.53%, respectively.

VI. PROOFS

In this section, we present the proofs of the results presented in the previous sections.

A. Proof of Bounds by Gaussian Complexities

In this subsection, we present the proofs for data-dependent bounds in subsection III-C. The proof of Lemma 1 requires to use a comparison result (Lemma 20) on Gaussian processes attributed to Slepian [42], while the proof of Theorem 2 is based on a concentration inequality in [72].

Lemma 20. *Let $\{\mathfrak{X}_\theta:\theta\in\Theta\}$ and $\{\mathfrak{Y}_\theta:\theta\in\Theta\}$ be two mean-zero separable Gaussian processes indexed by the same set Θ and suppose that*

$$\mathbb{E}[(\mathfrak{X}_\theta-\mathfrak{X}_{\bar\theta})^2]\leq\mathbb{E}[(\mathfrak{Y}_\theta-\mathfrak{Y}_{\bar\theta})^2],\quad\forall\theta,\bar\theta\in\Theta. \tag{35}$$

Then $\mathbb{E}[\sup_{\theta\in\Theta}\mathfrak{X}_\theta]\leq\mathbb{E}[\sup_{\theta\in\Theta}\mathfrak{Y}_\theta]$.

Lemma 21 (McDiarmid's inequality [72])**.** *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*

$$\sup_{\mathbf{z}_1,\ldots,\mathbf{z}_n,\bar{\mathbf{z}}_i\in\mathcal{Z}}|f(\mathbf{z}_1,\cdots,\mathbf{z}_n)-f(\mathbf{z}_1,\cdots,\mathbf{z}_{i-1},\bar{\mathbf{z}}_i,\mathbf{z}_{i+1},\cdots,\mathbf{z}_n)|\leq c_i \tag{36}$$

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 Lemma 1. Define two mean-zero separable Gaussian processes indexed by the finite dimensional Euclidean space $\{(h(\mathbf{x}_1),\ldots,h(\mathbf{x}_n)):h\in H\}$

$$\mathfrak{X}_h:=\sum_{i=1}^n g_i f_i(h(\mathbf{x}_i)),$$
$$\mathfrak{Y}_h:=\sqrt{2}L_1\sum_{i=1}^n\sum_{j=1}^c g_{ij}h_j(\mathbf{x}_i)+\sqrt{2}L_2\sum_{i=1}^n g_i h_{r(i)}(\mathbf{x}_i).$$

For any $h,h'\in H$, the independence among g_i,g_{ij} and $\mathbb{E}g_i^2=1,\mathbb{E}g_{ij}^2=1,\forall i\in\mathbb{N}_n,j\in\mathbb{N}_c$ imply that

$$\begin{aligned}\mathbb{E}[(\mathfrak{X}_h-\mathfrak{X}_{h'})^2]&=\mathbb{E}\Big[\Big(\sum_{i=1}^n g_i\big(f_i(h(\mathbf{x}_i))-f_i(h'(\mathbf{x}_i))\big)\Big)^2\Big]\\&=\sum_{i=1}^n\big[f_i(h(\mathbf{x}_i))-f_i(h'(\mathbf{x}_i))\big]^2\\&\leq\sum_{i=1}^n\Big[L_1\Big[\sum_{j=1}^c|h_j(\mathbf{x}_i)-h_j'(\mathbf{x}_i)|^2\Big]^{\frac{1}{2}}+L_2|h_{r(i)}(\mathbf{x}_i)-h_{r(i)}'(\mathbf{x}_i)|\Big]^2\\&\leq 2L_1^2\sum_{i=1}^n\sum_{j=1}^c|h_j(\mathbf{x}_i)-h_j'(\mathbf{x}_i)|^2+2L_2^2\sum_{i=1}^n|h_{r(i)}(\mathbf{x}_i)-h_{r(i)}'(\mathbf{x}_i)|^2\\&=\mathbb{E}[(\mathfrak{Y}_h-\mathfrak{Y}_{h'})^2],\end{aligned}$$

where we have used the Lipschitz continuity of f_i w.r.t. a variant of the ℓ_2-norm in the first inequality, and the elementary inequality $(a+b)^2\leq 2(a^2+b^2)$ in the second

TABLE IV
PERFORMANCE OF MC-SVMS ON SEVERAL BENCHMARK DATASETS.
We consider MLR, ℓ_p-norm MC-SVM and WW MC-SVM in Corollary 13 with $\ell(t) = \log(1+\exp(-t))$. We traverse p over $\{1, 1.2, 1.5, 1.8, 2, 2.33, 2.5, 2.67, 3, 4, 8, \infty\}$ and Λ over $\{10^{0.5}, 10, \dots, 10^{3.5}\}$ to obtain the candidate models. We report the accuracy of the best candidate model and the selected model with a minimal structural risk in the columns "Oracle" and "Model selection", respectively. For ℓ_p-norm MC-SVM, we also report the parameter p at which the corresponding accuracy is achieved.

Dataset	MLR		ℓ_p-norm MC-SVM				WW MC-SVM	
	Oracle	Model Selection	Oracle		Model Selection		Oracle	Model Selection
	Accuracy	Accuracy	p	Accuracy	p	Accuracy	Accuracy	Accuracy
MNIST	91.43	91.39	3	91.99	8	91.82	91.00	90.98
NEWS20	84.07	83.25	4	84.45	4	84.45	84.10	84.10
LETTER	73.52	73.52	∞	73.74	∞	73.68	69.28	68.92
RCV1	88.67	88.62	1.8	88.71	2.33	88.65	88.68	86.96
SECTOR	93.08	93.08	4	93.30	2.33	93.20	92.83	91.21
ALOI	85.70	85.70	∞	88.48	∞	88.48	78.56	78.53

inequality. Therefore, the condition (35) holds and Lemma 20 can be applied here to give

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}} \sup_{h\in H} \sum_{i=1}^n g_i f_i(h(\mathbf{x}_i)) \\
&\leq \mathbb{E}_{\boldsymbol{g}} \sup_{h\in H} \Big[\sqrt{2}L_1 \sum_{i=1}^n \sum_{j=1}^c g_{ij} h_j(\mathbf{x}_i) + \sqrt{2}L_2 \sum_{i=1}^n g_i h_{r(i)}(\mathbf{x}_i)\Big] \\
&\leq \sqrt{2}L_1 \mathbb{E}_{\boldsymbol{g}} \sup_{h\in H} \sum_{i=1}^n \sum_{j=1}^c g_{ij} h_j(\mathbf{x}_i) + \sqrt{2}L_2 \mathbb{E}_{\boldsymbol{g}} \sup_{h\in H} \sum_{i=1}^n g_i h_{r(i)}(\mathbf{x}_i).
\end{aligned}$$

The proof of Lemma 1 is complete. □

Proof of Theorem 2. It can be checked that

$$f(\mathbf{z}_1, \dots, \mathbf{z}_n) = \sup_{h^{\mathbf{w}}\in H_\tau} \Big[\mathbb{E}_{\mathbf{z}} \Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big]$$

satisfies the increment condition (36) with $c_i = B_\Psi/n$. An application of McDiarmid's inequality (Lemma 21) then shows the following inequality with probability of $1-\delta/2$

$$\begin{aligned}
&\sup_{h^{\mathbf{w}}\in H_\tau} \Big[\mathbb{E}_{\mathbf{z}} \Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big] \leq \\
&\mathbb{E}_{\mathbf{z}} \sup_{h^{\mathbf{w}}\in H_\tau} \Big[\mathbb{E}_{\mathbf{z}} \Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big] + B_\Psi \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
\end{aligned}$$

It follows from the standard symmetrization technique (see, e.g., proof of Theorem 3.1 in [9]) that

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}} \sup_{h^{\mathbf{w}}\in H_\tau} &\Big[\mathbb{E}_{\mathbf{x},y} \Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big] \\
&\leq 2\mathbb{E}_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h^{\mathbf{w}}\in H_\tau} \Big[\frac{1}{n}\sum_{i=1}^n \epsilon_i \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big].
\end{aligned}$$

It can also be checked that the function

$$f(\mathbf{z}_1, \dots, \mathbf{z}_n) = \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h^{\mathbf{w}}\in H_\tau} \big[\frac{1}{n}\sum_{i=1}^n \epsilon_i \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\big]$$

satisfies the increment condition (36) with $c_i = B_\Psi/n$. Another application of McDiarmid's inequality shows the inequality

$$\mathbb{E}_{\mathbf{z}} \mathfrak{R}_S(F_{\tau,\Lambda}) \leq \mathfrak{R}_S(F_{\tau,\Lambda}) + B_\Psi \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

with probability of $1-\delta/2$, which together with the above two inequalities then imply the following inequality with probability of at least $1-\delta$

$$\begin{aligned}
&\sup_{h^{\mathbf{w}}\in H_\tau} \Big[\mathbb{E}_{\mathbf{z}} \Psi_y(h^{\mathbf{w}}(\mathbf{x})) - \frac{1}{n}\sum_{i=1}^n \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))\Big] \leq \\
&\qquad 2\mathfrak{R}_S(F_{\tau,\Lambda}) + 3B_\Psi \sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \qquad (37)
\end{aligned}$$

Furthermore, according to the following relationship between Gaussian and Rademacher processes for any function class $\tilde{H}$ [37] ($|S|$ is the cardinality of S)

$$\mathfrak{R}_S(\tilde{H}) \leq \sqrt{\frac{\pi}{2}} \mathfrak{G}_S(\tilde{H}) \leq 3\sqrt{\frac{\pi \log |S|}{2}} \mathfrak{R}_S(\tilde{H}),$$

we derive

$$\begin{aligned}
&\mathfrak{R}_S\big(\big\{\Psi_y(h^{\mathbf{w}}(\mathbf{x})) : h^{\mathbf{w}} \in H_\tau\big\}\big) \\
&\leq \sqrt{\frac{\pi}{2}} \mathfrak{G}_S\big(\big\{\Psi_y(h^{\mathbf{w}}(\mathbf{x})) : h^{\mathbf{w}} \in H_\tau\big\}\big) \\
&= \sqrt{\frac{\pi}{2}} \frac{1}{n} \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau} \sum_{i=1}^n g_i \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i)) \\
&\leq \frac{L_1\sqrt{\pi}}{n} \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau} \sum_{i=1}^n \sum_{j=1}^c g_{ij} h_j^{\mathbf{w}}(\mathbf{x}_i) \\
&\qquad + \frac{L_2\sqrt{\pi}}{n} \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau} \sum_{i=1}^n g_i h_{y_i}^{\mathbf{w}}(\mathbf{x}_i),
\end{aligned}$$

where the last step follows from Lemma 1 with $f_i = \Psi_{y_i}$ and $r(i) = y_i, \forall i \in \mathbb{N}_n$. Plugging the above RC bound into (37) gives the following inequality with probability of at least $1-\delta$

$$A_\tau \leq \frac{2L_1\sqrt{\pi}}{n}\mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{i=1}^n\sum_{j=1}^c g_{ij}\langle\mathbf{w}_j,\phi(\mathbf{x}_i)\rangle + \frac{2L_2\sqrt{\pi}}{n}\mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{i=1}^n g_i\langle\mathbf{w}_{y_i},\phi(\mathbf{x}_i)\rangle. \quad (38)$$

It remains to estimate the two terms on the right-hand side of (38). By (12), the definition of $\widetilde{H}_\tau, \widetilde{S}$ and $\widetilde{S}'$, we know

$$\mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{i=1}^n\sum_{j=1}^c g_{ij}\langle\mathbf{w}_j,\phi(\mathbf{x}_i)\rangle = \mathbb{E}_{\boldsymbol{g}} \sup_{\mathbf{w}:\tau(\mathbf{w})\leq\Lambda}\sum_{i=1}^n\sum_{j=1}^c g_{ij}\langle\mathbf{w},\tilde{\boldsymbol{\phi}}_j(\mathbf{x}_i)\rangle = nc\mathfrak{G}_{\widetilde{S}}(\widetilde{H}_\tau) \quad (39)$$

and

$$\mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{i=1}^n g_i\langle\mathbf{w}_{y_i},\phi(\mathbf{x}_i)\rangle = \mathbb{E}_{\boldsymbol{g}} \sup_{\mathbf{w}:\tau(\mathbf{w})\leq\Lambda}\sum_{i=1}^n g_i\langle\mathbf{w},\tilde{\boldsymbol{\phi}}_{y_i}(\mathbf{x}_i)\rangle = n\mathfrak{G}_{\widetilde{S}'}(\widetilde{H}_\tau).$$

Plugging the above two identities back into (38) gives (13).

We now show (14). According to the definition of dual norm, we derive

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{i=1}^n\sum_{j=1}^c g_{ij}\langle\mathbf{w}_j,\phi(\mathbf{x}_i)\rangle \\
&= \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{j=1}^c \Big\langle\mathbf{w}_j,\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big\rangle \\
&= \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\langle\mathbf{w},\big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\big)_{j=1}^c\rangle \\
&\leq \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\|\mathbf{w}\|\big\|\big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\big)_{j=1}^c\big\|_* \\
&= \Lambda\mathbb{E}_{\boldsymbol{g}}\Big\|\big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\big)_{j=1}^c\Big\|_*. \quad (40)
\end{aligned}$$

Analogously, we also have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{i=1}^n g_i\langle\mathbf{w}_{y_i},\phi(\mathbf{x}_i)\rangle \\
&= \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\sum_{j=1}^c\langle\mathbf{w}_j,\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\rangle \\
&= \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}}\in H_\tau}\langle\mathbf{w},\big(\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\big)_{j=1}^c\rangle \\
&\leq \Lambda\mathbb{E}_{\boldsymbol{g}}\big\|\big(\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\big)_{j=1}^c\big\|_*.
\end{aligned}$$

Plugging the above two inequalities back into (38) gives (14). □

Proof of Corollary 3. Let $q\geq p$ be any real number. It follows from Jensen's inequality and Khintchine-Kahane inequality (69) that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}}\Big\|\Big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*} = \mathbb{E}_{\boldsymbol{g}}\Big[\sum_{j=1}^c\Big\|\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big\|_2^{q^*}\Big]^{\frac{1}{q^*}} \\
&\leq \Big[\sum_{j=1}^c\mathbb{E}_{\boldsymbol{g}}\Big\|\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big\|_2^{q^*}\Big]^{\frac{1}{q^*}} \leq \Big[\sum_{j=1}^c\Big[q^*\sum_{i=1}^n\|\phi(\mathbf{x}_i)\|_2^2\Big]^{\frac{q^*}{2}}\Big]^{\frac{1}{q^*}} \\
&= c^{\frac{1}{q^*}}\Big[q^*\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{1}{2}}. \quad (41)
\end{aligned}$$

Applying again Jensen's inequality and Khintchine-Kahane inequality (69), we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}}\Big\|\Big(\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*} &\leq \Big[\mathbb{E}_{\boldsymbol{g}}\sum_{j=1}^c\Big\|\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\Big\|_2^{q^*}\Big]^{\frac{1}{q^*}} \\
&\leq \sqrt{q^*}\Big[\sum_{j=1}^c\Big[\sum_{i\in I_j}\|\phi(\mathbf{x}_i)\|_2^2\Big]^{\frac{q^*}{2}}\Big]^{\frac{1}{q^*}}. \quad (42)
\end{aligned}$$

We now control the last term in the above inequality by distinguishing whether $q\geq 2$ or not. If $q\leq 2$, we have $2^{-1}q^*\geq 1$ and it follows from the elementary inequality $a^s+b^s\leq(a+b)^s, \forall a,b\geq 0, s\geq 1$ that

$$\begin{aligned}
\sum_{j=1}^c\Big[\sum_{i\in I_j}K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{q^*}{2}} &\leq \Big[\sum_{j=1}^c\sum_{i\in I_j}K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{q^*}{2}} \\
&= \Big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{q^*}{2}}. \quad (43)
\end{aligned}$$

Otherwise we have $2^{-1}q^*\leq 1$ and Jensen's inequality implies

$$\begin{aligned}
\sum_{j=1}^c\Big[\sum_{i\in I_j}K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{q^*}{2}} &\leq c\Big[\sum_{j=1}^c\frac{1}{c}\sum_{i\in I_j}K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{q^*}{2}} \\
&= c^{1-\frac{q^*}{2}}\Big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{q^*}{2}}. \quad (44)
\end{aligned}$$

Combining (42), (43) and (44) together implies

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}}\Big\|\Big(\sum_{i\in I_j} g_i\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*} \\
&\leq \max(c^{\frac{1}{q^*}-\frac{1}{2}},1)\Big[q^*\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{1}{2}}. \quad (45)
\end{aligned}$$

According to the monotonicity of $\|\cdot\|_{2,p}$ w.r.t. p, we have $H_{p,\Lambda}\subset H_{q,\Lambda}$ if $p\leq q$. Plugging the complexity bound established in Eqs. (41), (45) into the generalization bound given in Theorem 2, we get the following inequality with probability of at least $1-\delta$

$$\begin{aligned}
A_\tau \leq \frac{2\Lambda\sqrt{\pi}}{n}\Big[&L_1 c^{\frac{1}{q^*}}\Big[q^*\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{1}{2}} \\
&+ L_2\max(c^{\frac{1}{q^*}-\frac{1}{2}},1)\Big[q^*\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac{1}{2}}\Big], \ \forall q\geq p.
\end{aligned}$$

The proof is complete. □

Remark 7 (Tightness of the Rademacher Complexity Bound)**.** Eq. (41) gives an upper bound on $\mathbb{E}_{\boldsymbol{g}}\Big\|\Big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*}$. We now show that this bound is tight up to a constant factor. Indeed, according to the elementary inequality for $a_1,\ldots,a_c\geq 0$

$$\big(a_1+\cdots+a_c\big)^{\frac{1}{q^*}}\geq c^{\frac{1}{q^*}-1}\big(a_1^{\frac{1}{q^*}}+\cdots+a_c^{\frac{1}{q^*}}\big),$$

we derive

$$\begin{aligned}\Big\|\Big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*}&=\Big[\sum_{j=1}^c\Big\|\sum_{i=1}^n g_{ij}\phi(x_i)\Big\|_2^{q^*}\Big]^{\frac{1}{q^*}}\\&\geq c^{\frac{1}{q^*}-1}\sum_{j=1}^c\Big\|\sum_{i=1}^n g_{ij}\phi(x_i)\Big\|_2.\end{aligned}$$

Taking expectations on both sides, we get that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}}\Big\|\Big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*}&\geq c^{\frac{1}{q^*}-1}\sum_{j=1}^c\mathbb{E}_{\boldsymbol{g}}\Big\|\sum_{i=1}^n g_{ij}\phi(x_i)\Big\|_2\\&\geq 2^{-\frac12}c^{\frac{1}{q^*}}\Big[\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)\Big]^{\frac12},\end{aligned}$$

where the second inequality is due to (70). The above lower bound coincides with the upper bound (41) up to a constant factor. Specifically, the above upper and lower bounds show that $\mathbb{E}_{\boldsymbol{g}}\Big\|\Big(\sum_{i=1}^n g_{ij}\phi(\mathbf{x}_i)\Big)_{j=1}^c\Big\|_{2,q^*}$ enjoys exactly a square-root dependency on the number of classes if $q=2$.

Proof of Corollary 4. We first consider the case $1\leq p\leq 2$. Let $q\in\mathbb{R}$ satisfy $p\leq q\leq 2$. Denote $\tilde{X}_i^j=(0,\ldots,0,\mathbf{x}_i,0,\ldots,0)$ with the j-th column being $\mathbf{x}_i$. Then, we have

$$\begin{aligned}\big(\sum_{i=1}^n g_{ij}\mathbf{x}_i\big)_{j=1}^c&=\sum_{i=1}^n\sum_{j=1}^c g_{ij}\tilde{X}_i^j\\\big(\sum_{i\in I_1}g_i\mathbf{x}_i,\ldots,\sum_{i\in I_c}g_i\mathbf{x}_i\big)&=\sum_{j=1}^c\sum_{i\in I_j}g_i\tilde{X}_i^j.\end{aligned}\tag{46}$$

Since $q^*\geq 2$, we can apply Jensen's inequality and Khintchine-Kahane inequality (71) to derive (recall $\sigma_r(X)$ denotes the r-th singular value of X)

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}}\Big\|\sum_{i=1}^n\sum_{j=1}^c g_{ij}\tilde{X}_i^j\Big\|_{S_{q^*}}&\leq\Big[\mathbb{E}_{\boldsymbol{g}}\sum_{r=1}^{\min\{c,d\}}\sigma_r^{q^*}\big(\sum_{i=1}^n\sum_{j=1}^c g_{ij}\tilde{X}_i^j\big)\Big]^{\frac{1}{q^*}}\\&\leq 2^{-\frac14}\sqrt{\frac{\pi q^*}{e}}\max\Big\{\Big\|\Big[\sum_{i=1}^n\sum_{j=1}^c(\tilde{X}_i^j)^\top\tilde{X}_i^j\Big]^{\frac12}\Big\|_{S_{q^*}},\\&\qquad\Big\|\Big[\sum_{i=1}^n\sum_{j=1}^c\tilde{X}_i^j(\tilde{X}_i^j)^\top\Big]^{\frac12}\Big\|_{S_{q^*}}\Big\}.\end{aligned}\tag{47}$$

For any $\mathbf{u}=(u_1,\ldots,u_c)\in\mathbb{R}^c$, we denote by $\mathrm{diag}(\mathbf{u})$ the diagonal matrix in $\mathbb{R}^{c\times c}$ with the j-th diagonal element being u_j. The following identities can be directly checked

$$\begin{aligned}\sum_{i=1}^n\sum_{j=1}^c(\tilde{X}_i^j)^\top(\tilde{X}_i^j)&=\sum_{i=1}^n\sum_{j=1}^c\|\mathbf{x}_i\|_2^2\mathrm{diag}(\mathbf{e}_j)=\sum_{i=1}^n\|\mathbf{x}_i\|_2^2 I_{c\times c},\\\sum_{i=1}^n\sum_{j=1}^c(\tilde{X}_i^j)(\tilde{X}_i^j)^\top&=\sum_{i=1}^n\sum_{j=1}^c\mathbf{x}_i\mathbf{x}_i^\top=c\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top,\end{aligned}$$

where $(\mathbf{e}_1,\ldots,\mathbf{e}_c)$ forms the identity matrix $I_{c\times c}\in\mathbb{R}^{c\times c}$ Therefore,

$$\begin{aligned}\Big\|\Big[\sum_{i=1}^n\sum_{j=1}^c(\tilde{X}_i^j)^\top(\tilde{X}_i^j)\Big]^{\frac12}\Big\|_{S_{q^*}}&=\Big\|\Big(\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big)^{\frac12}I_{c\times c}\Big\|_{S_{q^*}}\\&=c^{\frac{1}{q^*}}\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac12},\end{aligned}\tag{48}$$

and

$$\begin{aligned}&\Big\|\Big[\sum_{i=1}^n\sum_{j=1}^c(\tilde{X}_i^j)(\tilde{X}_i^j)^\top\Big]^{\frac12}\Big\|_{S_{q^*}}\\&=\sqrt{c}\Big\|\Big(\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\Big)^{\frac12}\Big\|_{S_{q^*}}=\sqrt{c}\Big[\sum_{r=1}^{\min\{c,d\}}\sigma_r^{q^*}\Big(\big(\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\big)^{\frac12}\Big)\Big]^{\frac{1}{q^*}}\\&=\sqrt{c}\Big[\sum_{r=1}^{\min\{c,d\}}\sigma_r^{\frac{q^*}{2}}\big(\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\big)\Big]^{\frac{1}{q^*}}=\Big[c\Big\|\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_{\frac{q^*}{2}}}\Big]^{\frac12}.\end{aligned}\tag{49}$$

Plugging (48) and (49) into (47) gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}}\Big\|\sum_{i=1}^n\sum_{j=1}^c g_{ij}\tilde{X}_i^j\Big\|_{S_{q^*}}&\leq 2^{-\frac14}\sqrt{\frac{\pi q^*}{e}}\max\Big\{c^{\frac{1}{q^*}}\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac12},\\&\qquad c^{\frac12}\Big\|\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_{\frac{q^*}{2}}}^{\frac12}\Big\}.\end{aligned}\tag{50}$$

Applying again Jensen's inequality and Khintchine-Kahane inequality (71) gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}}\Big\|\sum_{j=1}^c\sum_{i\in I_j}g_i\tilde{X}_i^j\Big\|_{S_{q^*}}&\leq\Big[\mathbb{E}_{\boldsymbol{g}}\sum_{r=1}^{\min\{c,d\}}\sigma_r^{q^*}\big(\sum_{j=1}^c\sum_{i\in I_j}g_i\tilde{X}_i^j\big)\Big]^{\frac{1}{q^*}}\\&\leq 2^{-\frac14}\sqrt{\frac{\pi q^*}{e}}\max\Big\{\Big\|\Big[\sum_{j=1}^c\sum_{i\in I_j}(\tilde{X}_i^j)^\top\tilde{X}_i^j\Big]^{\frac12}\Big\|_{S_{q^*}},\\&\qquad\Big\|\Big[\sum_{j=1}^c\sum_{i\in I_j}\tilde{X}_i^j(\tilde{X}_i^j)^\top\Big]^{\frac12}\Big\|_{S_{q^*}}\Big\}.\end{aligned}\tag{51}$$

It can be directly checked that

$$\begin{aligned}\sum_{j=1}^c\sum_{i\in I_j}(\tilde{X}_i^j)^\top(\tilde{X}_i^j)&=\sum_{j=1}^c\sum_{i\in I_j}\|\mathbf{x}_i\|_2^2\mathrm{diag}(\mathbf{e}_j)\\&=\mathrm{diag}\big(\sum_{i\in I_1}\|\mathbf{x}_i\|_2^2,\ldots,\sum_{i\in I_c}\|\mathbf{x}_i\|_2^2\big)\end{aligned}$$

and

$$\sum_{j=1}^c\sum_{i\in I_j}(\tilde{X}_i^j)(\tilde{X}_i^j)^\top=\sum_{j=1}^c\sum_{i\in I_j}\mathbf{x}_i\mathbf{x}_i^\top=\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top,$$

from which and $q^*\geq 2$ we derive

$$\Big\|\Big[\sum_{j=1}^c\sum_{i\in I_j}(\tilde{X}_i^j)^\top\tilde{X}_i^j\Big]^{\frac{1}{2}}\Big\|_{S_{q^*}}=\Big[\sum_{j=1}^c\Big(\sum_{i\in I_j}\|\mathbf{x}_i\|_2^2\Big)^{\frac{q^*}{2}}\Big]^{\frac{1}{q^*}}$$
$$\leq\Big[\sum_{j=1}^c\sum_{i\in I_j}\|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}}=\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}}$$

and

$$\Big\|\Big[\sum_{j=1}^c\sum_{i\in I_j}\tilde{X}_i^j(\tilde{X}_i^j)^\top\Big]^{\frac{1}{2}}\Big\|_{S_{q^*}}=\Big\|\Big[\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\Big]^{\frac{1}{2}}\Big\|_{S_{q^*}}$$
$$\leq\Big\|\Big[\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\Big]^{\frac{1}{2}}\Big\|_{S_2}=\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}},$$

where we have used deduction similar to (49) in the last identity. Plugging the above two inequalities back into (51) implies

$$\mathbb{E}_{\boldsymbol{g}}\Big\|\sum_{j=1}^c\sum_{i\in I_j}g_i\tilde{X}_i^j\Big\|_{S_{q^*}}\leq 2^{-\frac{1}{4}}\sqrt{\frac{\pi q^*}{e}}\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}}.\quad(52)$$

Plugging (50) and (52) into Theorem 2 and noting that $H_{S_p}\subset H_{S_q}$ we get the following inequality with probability of at least $1-\delta$

$$A_{S_p}\leq\frac{2^{\frac{3}{4}}\pi\Lambda}{n\sqrt{e}}\inf_{p\leq q\leq 2}(q^*)^{\frac{1}{2}}\Big\{L_1\max\Big\{c^{\frac{1}{q^*}}\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}},$$
$$c^{\frac{1}{2}}\Big\|\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_{\frac{q^*}{2}}}^{\frac{1}{2}}\Big\}+L_2\Big[\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\Big]^{\frac{1}{2}}\Big\}.\quad(53)$$

This finishes the proof for the case $p\leq 2$.

We now consider the case $p>2$. For any W with $\|W\|_{S_p}\leq\Lambda$, we have $\|W\|_{S_2}\leq\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}\Lambda$. The stated bound (16) for the case $p>2$ then follows by recalling the established generalization bound (53) for $p=2$. □

B. Proof of Bounds by Covering Numbers

We use the tool of empirical ℓ_∞-norm CNs to prove data-dependent bounds given in subsection III-D. The key observation to proceed with the proof is that the empirical ℓ_∞-norm CNs of $F_{\tau,\Lambda}$ w.r.t. the training examples can be controlled by that of $\widetilde{H}_\tau$ w.r.t. an enlarged data set of cardinality nc, due to the Lipschitz continuity of loss functions w.r.t. the ℓ_∞-norm [48, 73]. The remaining problem is to estimate the empirical CNs of $\widetilde{H}_\tau$, which, by the universal relationship between fat-shattering dimension and CNs (Part (a) of Lemma 22), can be further transferred to the estimation of fat-shattering dimension. Finally, the problem of estimating fat-shattering dimension reduces to the estimation of *worst case* RC (Part (b) of Lemma 22). We summarize this deduction process in the proof of Theorem 23.

Definition 3 (Covering number)**.** Let F be a class of real-valued functions defined over a space $\widetilde{\mathcal{Z}}$ and $S':=\{\tilde{\mathbf{z}}_1,\ldots,\tilde{\mathbf{z}}_n\}\in\tilde{\mathcal{Z}}^n$ of cardinality n. For any $\epsilon>0$, the empirical ℓ_∞-norm CN $\mathcal{N}_\infty(\epsilon,F,S')$ w.r.t. S' is defined as the minimal number m of a collection of vectors $\mathbf{v}^1,\ldots,\mathbf{v}^m\in\mathbb{R}^n$ such that ($\mathbf{v}_i^j$ is the i-th component of the vector $\mathbf{v}^j$)

$$\sup_{f\in F}\min_{j=1,\ldots,m}\max_{i=1,\ldots,n}|f(\tilde{\mathbf{z}}_i)-\mathbf{v}_i^j|\leq\epsilon.$$

In this case, we call $\{\mathbf{v}^1,\ldots,\mathbf{v}^m\}$ an (ϵ,ℓ_∞)-cover of F w.r.t. S'.

Definition 4 (Fat-Shattering Dimension)**.** Let F be a class of real-valued functions defined over a space $\widetilde{\mathcal{Z}}$. We define the fat-shattering dimension $\mathrm{fat}_\epsilon(F)$ at scale $\epsilon>0$ as the largest $D\in\mathbb{N}$ such that there exist D points $\tilde{\mathbf{z}}_1,\ldots,\tilde{\mathbf{z}}_D\in\tilde{\mathcal{Z}}$ and witnesses $s_1,\ldots,s_D\in\mathbb{R}$ satisfying: for any $\delta_1,\ldots,\delta_D\in\{\pm1\}$ there exists $f\in F$ with

$$\delta_i(f(\tilde{\mathbf{z}}_i)-s_i)\geq\epsilon/2,\qquad\forall i=1,\ldots,D.$$

Lemma 22 ([74, 75])**.** *Let F be a class of real-valued functions defined over a space $\widetilde{\mathcal{Z}}$ and $S':=\{\tilde{\mathbf{z}}_1,\ldots,\tilde{\mathbf{z}}_n\}\in\tilde{\mathcal{Z}}^n$ of cardinality n.*

(a) If functions in F take values in $[-B,B]$, then for any $\epsilon>0$ with $\mathrm{fat}_\epsilon(F)<n$ we have

$$\log\mathcal{N}_\infty(\epsilon,F,S')\leq\mathrm{fat}_\epsilon(F)\log\frac{2eBn}{\epsilon}.$$

(b) For any $\epsilon>2\mathfrak{R}_n(F)$, we have $\mathrm{fat}_\epsilon(F)<\frac{16n}{\epsilon^2}\mathfrak{R}_n^2(F)$.

(c) For any monotone sequence $(\epsilon_k)_{k=0}^\infty$ decreasing to 0 such that $\epsilon_0\geq\sqrt{n^{-1}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f^2(\tilde{\mathbf{z}}_i)}$, the following inequality holds for every non-negative integer N:

$$\mathfrak{R}_{S'}(F)\leq 2\sum_{k=1}^N\big(\epsilon_k+\epsilon_{k-1}\big)\sqrt{\frac{\log\mathcal{N}_\infty(\epsilon_k,F,S')}{n}}+\epsilon_N.\quad(54)$$

Theorem 23 (Covering number bounds)**.** *Assume that, for any $y\in\mathcal{Y}$, the function Ψ_y is L-Lipschitz continuous w.r.t. the ℓ_∞-norm. Then, for any $\epsilon>4L\mathfrak{R}_{nc}(\widetilde{H}_\tau)$, the CN of $F_{\tau,\Lambda}$ w.r.t. $S=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$ can be bounded by*

$$\log\mathcal{N}_\infty(\epsilon,F_{\tau,\Lambda},S)\leq\frac{16ncL^2\mathfrak{R}_{nc}^2(\widetilde{H}_\tau)}{\epsilon^2}\log\frac{2e\hat{B}ncL}{\epsilon}.$$

Proof. We proceed with the proof in three steps. Note that $\widetilde{H}_\tau$ is a class of functions defined on a finite set $\widetilde{S}=\big\{\tilde{\boldsymbol{\phi}}_j(\mathbf{x}_i):i\in\mathbb{N}_n,j\in\mathbb{N}_c\big\}$.

Step 1. We first estimate the CN of $\widetilde{H}_\tau$ w.r.t. $\widetilde{S}$. For any $\epsilon>4\mathfrak{R}_{nc}(\widetilde{H}_\tau)$, Part (b) of Lemma 22 implies that

$$\mathrm{fat}_\epsilon(\widetilde{H}_\tau)<\frac{16nc}{\epsilon^2}\mathfrak{R}_{nc}^2(\widetilde{H}_\tau)\leq nc.\quad(55)$$

According to (12) and the definition of $\hat{B}$, we derive the following inequality for any $\mathbf{w}$ with $\tau(\mathbf{w})\leq\Lambda$ and $i\in\mathbb{N}_n,j\in\mathbb{N}_c$

$$|\langle\mathbf{w},\tilde{\boldsymbol{\phi}}_j(\mathbf{x}_i)\rangle|=|\langle\mathbf{w}_j,\phi(\mathbf{x}_i)\rangle|\leq\|\mathbf{w}_j\|_2\|\phi(\mathbf{x}_i)\|_2$$
$$\leq\sup_{\mathbf{w}:\tau(\mathbf{w})\leq\Lambda}\|\mathbf{w}\|_{2,\infty}\|\phi(\mathbf{x}_i)\|_2\leq\hat{B}.$$

Then, the conditions of Part (a) in Lemma 22 are satisfied with $F = \widetilde{H}_\tau, B = \hat{B}$ and $S' = \widetilde{S}$, and we can apply it to control the CNs for any $\epsilon > 4\mathfrak{R}_{nc}(\widetilde{H}_\tau)$ (note $\mathrm{fat}_\epsilon(\widetilde{H}_\tau) < nc$ in (55))

$$\begin{aligned}\log \mathcal{N}_\infty(\epsilon, \widetilde{H}_\tau, \widetilde{S}) &\leq \mathrm{fat}_\epsilon(\widetilde{H}_\tau) \log \frac{2e\hat{B}nc}{\epsilon} \\ &\leq \frac{16nc\mathfrak{R}^2_{nc}(\widetilde{H}_\tau)}{\epsilon^2} \log \frac{2e\hat{B}nc}{\epsilon}, \qquad (56)\end{aligned}$$

where the second inequality is due to (55).

Step 2. We now relate the empirical ℓ_∞-norm CNs of $\widetilde{H}_\tau$ w.r.t. $\widetilde{S}$ to that of $F_{\tau,\Lambda}$ w.r.t. S. Let

$$\Big\{\mathbf{r}^j = \big(r^j_{1,1}, r^j_{1,2}, \ldots, r^j_{1,c}, \ldots, r^j_{n,1}, r^j_{n,2}, \ldots, r^j_{n,c}\big) : j = 1, \ldots, N\Big\} \subset \mathbb{R}^{nc} \quad (57)$$

be an (ϵ, ℓ_∞)-cover of

$$\Big\{\big(\underbrace{\langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_1(\mathbf{x}_1)\rangle, \ldots, \langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_c(\mathbf{x}_1)\rangle}_{\text{related to } \mathbf{x}_1}, \underbrace{\langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_1(\mathbf{x}_2)\rangle, \ldots, \langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_c(\mathbf{x}_2)\rangle}_{\text{related to } \mathbf{x}_2}, \ldots \underbrace{\langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_1(\mathbf{x}_n)\rangle, \ldots, \langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_c(\mathbf{x}_n)\rangle}_{\text{related to } \mathbf{x}_n}\big) : \tau(\mathbf{w}) \leq \Lambda\Big\} \subset \mathbb{R}^{nc}$$

with N not larger than the right-hand side of (56). Define $\mathbf{r}^j_i = \big(r^j_{i,1}, \ldots, r^j_{i,c}\big)$ for all $i \in \mathbb{N}_n, j \in \mathbb{N}_N$. Now, we show that

$$\Big\{\big(\Psi_{y_1}(\mathbf{r}^j_1), \Psi_{y_2}(\mathbf{r}^j_2) \ldots, \Psi_{y_n}(\mathbf{r}^j_n)\big) : j = 1, \ldots, N\Big\} \subset \mathbb{R}^n \quad (58)$$

would be an $(L\epsilon, \ell_\infty)$-cover of the set (note $h^{\mathbf{w}}(\mathbf{x}) = \big(\langle \mathbf{w}_1, \phi(\mathbf{x})\rangle, \ldots, \langle \mathbf{w}_c, \phi(\mathbf{x})\rangle\big)$)

$$\Big\{\big(\Psi_{y_1}(h^{\mathbf{w}}(\mathbf{x}_1)), \ldots, \Psi_{y_n}(h^{\mathbf{w}}(\mathbf{x}_n))\big) : \tau(\mathbf{w}) \leq \Lambda\Big\} \subset \mathbb{R}^n.$$

Indeed, for any $\mathbf{w} \in H^c_K$ with $\tau(\mathbf{w}) \leq \Lambda$, the construction of the cover in Eq. (57) guarantees the existence of $j_{(\mathbf{w})} \in \{1, \ldots, N\}$ such that

$$\max_{1\leq i\leq n} \max_{1\leq k\leq c} \big|r^{j_{(\mathbf{w})}}_{i,k} - \langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_k(\mathbf{x}_i)\rangle\big| \leq \epsilon. \quad (59)$$

Then, the Lipschitz continuity of Ψ_y w.r.t. the ℓ_∞-norm implies that

$$\begin{aligned}&\max_{1\leq i\leq n} |\Psi_{y_i}(\mathbf{r}^{j(\mathbf{w})}_i) - \Psi_{y_i}(h^{\mathbf{w}}(\mathbf{x}_i))| \\ &\leq L \max_{1\leq i\leq n} \|\mathbf{r}^{j(\mathbf{w})}_i - h^{\mathbf{w}}(\mathbf{x}_i)\|_\infty \\ &= L \max_{1\leq i\leq n} \max_{1\leq k\leq c} \big|r^{j_{(\mathbf{w})}}_{i,k} - \langle \mathbf{w}_k, \phi(\mathbf{x}_i)\rangle\big| \\ &= L \max_{1\leq i\leq n} \max_{1\leq k\leq c} \big|r^{j_{(\mathbf{w})}}_{i,k} - \langle \mathbf{w}, \tilde{\boldsymbol{\phi}}_k(\mathbf{x}_i)\rangle\big| \\ &\leq L\epsilon,\end{aligned}$$

where we have used (12) in the third step and (59) in the last step. That is, the set defined in (58) is also an $(L\epsilon, \ell_\infty)$-cover of $F_{\tau,\Lambda}$ w.r.t. $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Therefore,

$$\log \mathcal{N}_\infty(\epsilon, F_{\tau,\Lambda}, S) \leq \log \mathcal{N}_\infty(\epsilon/L, \widetilde{H}_\tau, \widetilde{S}), \quad \forall \epsilon > 0. \quad (60)$$

Step 3. The stated result follows directly if we plug the complexity bound of $\widetilde{H}_\tau$ established in (56) into (60). The proof is complete. □

We can now apply the entropy integral (54) to control $\mathfrak{R}_S(F_{\tau,\Lambda})$ in terms of $\mathfrak{R}_{nc}(\widetilde{H}_\tau)$.

Proof of Theorem 5. Let

$$N = \left\lceil \log_2 \frac{n^{-\frac{1}{2}} \sup_{h\in H_\tau} \big\|\big(\Psi_{y_i}(h(\mathbf{x}_i))\big)^n_{i=1}\big\|_2}{16L\sqrt{c\log 2}\mathfrak{R}_{nc}(\widetilde{H}_\tau)} \right\rceil,$$

$\epsilon_N = 16L\sqrt{c\log 2}\mathfrak{R}_{nc}(\widetilde{H}_\tau)$ and $\epsilon_k = 2^{N-k}\epsilon_N, k = 0, \ldots, N-1$. It is clear that

$$\epsilon_0 \geq n^{-\frac{1}{2}} \sup_{h\in H_\tau} \big\|\big(\Psi_{y_i}(h(\mathbf{x}_i))\big)^n_{i=1}\big\|_2 \geq \epsilon_0/2$$

and $\epsilon_N \geq 4L\mathfrak{R}_{nc}(\widetilde{H}_\tau)$. Plugging the CN bounds established in Theorem 23 into the entropy integral (54), we derive the following inequality

$$\mathfrak{R}_S(F_{\tau,\Lambda}) \leq 8L\sqrt{c}\mathfrak{R}_{nc}(\widetilde{H}_\tau) \sum_{k=1}^N \frac{\epsilon_k + \epsilon_{k-1}}{\epsilon_k} \sqrt{\log \frac{2e\hat{B}ncL}{\epsilon_k}} + \epsilon_N. \quad (61)$$

We know

$$\begin{aligned}&\sum_{k=1}^N \sqrt{\log \frac{2e\hat{B}ncL}{\epsilon_k}} = \sum_{k=1}^N \sqrt{k\log 2 + \log(2e\hat{B}ncL\epsilon_0^{-1})} \\ &\leq \sqrt{\log 2} \int_1^{N+1} \sqrt{x + \log_2(2e\hat{B}ncL\epsilon_0^{-1})}dx \\ &= \frac{2\sqrt{\log 2}}{3} \int_1^{N+1} d\big(x + \log_2(2e\hat{B}ncL\epsilon_0^{-1})\big)^{\frac{3}{2}} \\ &\leq \frac{2\sqrt{\log 2}}{3} \log_2^{\frac{3}{2}}(4e\hat{B}ncL\epsilon_N^{-1}),\end{aligned}$$

where the last inequality follows from

$$4e\hat{B}ncL \geq 2n^{-\frac{1}{2}} \sup_{h\in H_\tau} \big\|\big(\Psi_{y_i}(h(\mathbf{x}_i))\big)^n_{i=1}\big\|_2 \geq \epsilon_0.$$

Plugging the above inequality back into (61) gives

$$\begin{aligned}\mathfrak{R}_S(F_{\tau,\Lambda}) &\leq 16L\sqrt{c\log 2}\mathfrak{R}_{nc}(\widetilde{H}_\tau) \log_2^{\frac{3}{2}}\big(4e\hat{B}ncL\epsilon_N^{-1}\big) + \epsilon_N \\ &= 16L\sqrt{c\log 2}\mathfrak{R}_{nc}(\widetilde{H}_\tau)\Big(1 + \log_2^{\frac{3}{2}} \frac{\sqrt{c}e\hat{B}n}{4\sqrt{\log 2}\mathfrak{R}_{nc}(\widetilde{H}_\tau)}\Big).\end{aligned}$$

The proof is complete by noting $e \leq 4\sqrt{\log 2}$. □

The proof of Theorem 6 is now immediate.

Proof of Theorem 6. The proof is complete if we plug the RC bounds established in Theorem 5 back into (37) and noting $32\sqrt{\log 2} \leq 27$. □

Proof of Corollary 9. Plugging the complexity bounds of $\widetilde{H}_p$ given in (18) into Theorem 6 gives the following inequality with probability of at least $1-\delta$

$$\begin{aligned}A_p &\leq \frac{27\sqrt{c}L\Lambda \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2}{n^{\frac{1}{2}} c^{\frac{1}{\max(2,p)}}}\Big(1 + \log_2^{\frac{3}{2}} \frac{\sqrt{2}\hat{B}n^{\frac{3}{2}} c^{\frac{1}{2} + \frac{1}{\max(2,p)}}}{\Lambda \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2}\Big) \\ &\leq \frac{27L\Lambda \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2 c^{\frac{1}{2} - \frac{1}{\max(2,p)}}}{\sqrt{n}}\Big(1 + \log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)\Big),\end{aligned}$$

where we have used the following inequality in the last step

$$\hat{B} = \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2 \sup_{\mathbf{w}:\|\mathbf{w}\|_{2,p}\leq\Lambda} \|\mathbf{w}\|_{2,\infty} \leq \Lambda \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2.$$

The proof of Corollary 9 is complete. □

Proof of Corollary 10. Consider any $W = (\mathbf{w}_1, \ldots, \mathbf{w}_c) \in \mathbb{R}^{d\times c}$. If $1 < p \leq 2$, then

$$\|W\|_{S_p} \geq \|W\|_{S_2} = \|W\|_{2,2} \geq \|W\|_{2,\infty}.$$

Otherwise, according to the following inequality for any semi-definite positive matrix $A = (a_{j\tilde{j}})_{j,\tilde{j}=1}^c$ (e.g., (1.67) in [76])

$$\|A\|_{S_{\tilde{p}}} \geq \big[\sum_{j=1}^c |a_{jj}|^{\tilde{p}}\big]^{\frac{1}{\tilde{p}}}, \quad \forall \tilde{p} \geq 1,$$

we derive

$$\begin{aligned}\|W\|_{S_p} &= \|(W^\top W)^{\frac12}\|_{S_p} = \Big\|\Big[\big(\mathbf{w}_j^\top \mathbf{w}_{\tilde{j}}\big)_{j,\tilde{j}=1}^c\Big]^{\frac12}\Big\|_{S_p} \\ &= \Big\|\big(\mathbf{w}_j^\top \mathbf{w}_{\tilde{j}}\big)_{j,\tilde{j}=1}^c\Big\|_{S_{\frac{p}{2}}}^{\frac12} \geq \Big[\sum_{j=1}^c (\mathbf{w}_j^\top \mathbf{w}_j)^{\frac{p}{2}}\Big]^{\frac1p} \\ &\geq \max_{j=1,\ldots,c} \|\mathbf{w}_j\|_2 = \|W\|_{2,\infty}.\end{aligned}$$

Thereby, for the specific choice $\tau(W) = \|W\|_{S_p}, p \geq 1$, we have

$$\hat{B} = \max_{i\in\mathbb{N}_n} \|\mathbf{x}_i\|_2 \sup_{W:\|W\|_{S_p}\leq\Lambda} \|W\|_{2,\infty} \leq \Lambda \max_{i\in\mathbb{N}_n} \|\mathbf{x}_i\|_2. \quad (62)$$

We now consider two cases. If $1 < p \leq 2$, plugging the RC bounds of $\widetilde{H}_{S_p}$ given in (21) into Theorem 6 gives the following inequality with probability of at least $1-\delta$

$$\begin{aligned}A_{S_p} &\leq \frac{27L\Lambda \max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{n}}\Big(1 + \log_2^{\frac32}\frac{\sqrt{2}\hat{B}n^{\frac32}c}{\Lambda \max_{i\in\mathbb{N}_n}\|\mathbf{x}\|_i}\Big) \\ &\leq \frac{27L\Lambda \max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{n}}\Big(1 + \log_2^{\frac32}\big(\sqrt{2}n^{\frac32}c\big)\Big),\end{aligned}$$

where the last step follows from (62). If $p > 2$, analyzing analogously yields the following inequality with probability of at least $1-\delta$

$$A_{S_p} \leq \frac{27L\Lambda \max\limits_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2 \min\{c,d\}^{\frac12-\frac1p}}{\sqrt{n}}\Big(1 + \log_2^{\frac32}(\sqrt{2}n^{\frac32}c)\Big).$$

The stated error bounds follow by combining the above two cases together. □

C. Proofs on worst-case Rademacher Complexities

Proof of Proposition 7. We proceed with the proof by distinguishing two cases according to the value of p.

We first consider the case $1 \leq p \leq 2$, for which the RC can be lower bounded by

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_p) &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\mathbb{E}_\epsilon \sup_{\|\mathbf{w}\|_{2,p}\leq\Lambda} \sum_{i=1}^{nc} \epsilon_i \langle \mathbf{w}, \mathbf{v}^i\rangle \\ &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\mathbb{E}_\epsilon \sup_{\|\mathbf{w}\|_{2,p}\leq\Lambda} \langle \mathbf{w}, \sum_{i=1}^{nc}\epsilon_i \mathbf{v}^i\rangle \\ &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{\Lambda}{nc}\mathbb{E}_\epsilon \big\|\sum_{i=1}^{nc}\epsilon_i\mathbf{v}^i\big\|_{2,p^*} \qquad (63) \\ &\geq \max_{\mathbf{v}^1\in\widetilde{S}} \frac{\Lambda}{nc}\mathbb{E}_\epsilon\big|\sum_{i=1}^{nc}\epsilon_i\big|\|\mathbf{v}^1\|_{2,p^*},\end{aligned}$$

where the equality (63) follows from the definition of dual norm and the inequality follows by taking $\mathbf{v}^1 = \cdots = \mathbf{v}^{nc}$. Applying the Khitchine-Kahane inequality (70) and using the definition of $\widetilde{S}$ in (9), we then derive ($\|\mathbf{v}\|_{2,p} = \|\mathbf{v}\|_{2,\infty}$ for $\mathbf{v}\in\widetilde{S}$)

$$\mathfrak{R}_{nc}(\widetilde{H}_p) \geq \frac{\Lambda}{\sqrt{2nc}}\max_{\mathbf{v}^1\in\widetilde{S}}\|\mathbf{v}^1\|_{2,p^*} = \frac{\Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{2nc}}.$$

Furthermore, according to the subset relationship $\widetilde{H}_p \subset \widetilde{H}_2, 1 \leq p \leq 2$ due to the monotonicity of $\|\cdot\|_{2,p}$, the term $\mathfrak{R}_{nc}(\widetilde{H}_p)$ can also be upper bounded by ($\mathbf{v}_j^i$ denotes the j-th component of $\mathbf{v}^i$)

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_p) \leq \mathfrak{R}_{nc}(\widetilde{H}_2) &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{\Lambda}{nc}\mathbb{E}_\epsilon\big\|\sum_{i=1}^{nc}\epsilon_i\mathbf{v}^i\big\|_{2,2} \\ &\leq \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{\Lambda}{nc}\sqrt{\sum_{j=1}^c \mathbb{E}_\epsilon\|\sum_{i=1}^{nc}\epsilon_i\mathbf{v}_j^i\|_2^2} \\ &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{\Lambda}{nc}\sqrt{\sum_{j=1}^c\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2} \\ &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{\Lambda}{nc}\sqrt{\sum_{i=1}^{nc}\|\mathbf{v}^i\|_{2,\infty}^2} \\ &= \frac{\Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2}{\sqrt{nc}},\end{aligned}$$

where the first identity is due to (63), the second inequality is due to Jensen's inequality and the last second identity is due to $\sum_{j=1}^c\|\mathbf{v}_j\|_2^2 = \|\mathbf{v}\|_{2,\infty}^2$ for all $\mathbf{v}\in\widetilde{S}$.

We now turn to the case $p > 2$. In this case, we have

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_p) &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\mathbb{E}_\epsilon \sup_{\|\mathbf{w}\|_{2,p}\leq\Lambda}\sum_{i=1}^{nc}\epsilon_i\sum_{j=1}^c\langle\mathbf{w}_j,\mathbf{v}_j^i\rangle \\ &\geq \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\mathbb{E}_\epsilon \sup_{\|\mathbf{w}_j\|_2^p\leq\frac{\Lambda^p}{c}:j\in\mathbb{N}_c}\sum_{i=1}^{nc}\epsilon_i\sum_{j=1}^c\langle\mathbf{w}_j,\mathbf{v}_j^i\rangle \\ &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\sum_{j=1}^c\mathbb{E}_\epsilon \sup_{\|\mathbf{w}_j\|_2^p\leq\frac{\Lambda^p}{c}}\sum_{i=1}^{nc}\epsilon_i\langle\mathbf{w}_j,\mathbf{v}_j^i\rangle \\ &= \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\sum_{j=1}^c\mathbb{E}_\epsilon \sup_{\|\mathbf{w}_j\|_2^p\leq\frac{\Lambda^p}{c}}\langle\mathbf{w}_j,\sum_{i=1}^{nc}\epsilon_i\mathbf{v}_j^i\rangle,\end{aligned}$$

where we can exchange the summation over j with the supremum in the second identity since the constraint $\|\mathbf{w}_j\|_2^p \leq \frac{\Lambda^p}{c}, j\in\mathbb{N}_c$ are decoupled. According to the definition of dual norm and the Khitchine-Kahane inequality (70), $\mathfrak{R}_{nc}(\widetilde{H}_p)$ can be further controlled by

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_p) &\geq \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\sum_{j=1}^c\frac{\Lambda}{c^{\frac1p}}\mathbb{E}_\epsilon\|\sum_{i=1}^{nc}\epsilon_i\mathbf{v}_j^i\|_2 \\ &\geq \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}} \frac{1}{nc}\sum_{j=1}^c\frac{\Lambda}{\sqrt{2}c^{\frac1p}}\big[\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2\big]^{\frac12}. \qquad (64)\end{aligned}$$

We can find $\bar{\mathbf{v}}^1,\ldots,\bar{\mathbf{v}}^{nc}\in\widetilde{S}$ such that for each $j\in\mathbb{N}_c$, there are exactly n $\bar{\mathbf{v}}^k$ with $\|\bar{\mathbf{v}}_j^k\|_2 = \max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2$.

Then, $\sum_{i=1}^{nc}\|\bar{\mathbf{v}}_j^i\|_2^2 = n\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2^2, \forall j\in\mathbb{N}_c$, which, coupled with (64), implies that

$$\mathfrak{R}_{nc}(\widetilde{H}_p) \ge \frac{1}{nc}\sum_{j=1}^{c}\frac{\Lambda}{\sqrt{2}c^{\frac{1}{p}}}\big[n\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2^2\big]^{\frac{1}{2}} \ge \Lambda\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2(2n)^{-\frac{1}{2}}c^{-\frac{1}{p}}.$$

On the other hand, according to (63) and Jensen's inequality, we derive

$$\frac{nc\mathfrak{R}_{nc}(\widetilde{H}_p)}{\Lambda} = \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\mathbb{E}_\epsilon\big\|\sum_{i=1}^{nc}\epsilon_i\mathbf{v}^i\big\|_{2,p^*} \le \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\Big[\mathbb{E}_\epsilon\sum_{j=1}^{c}\|\sum_{i=1}^{nc}\epsilon_i\mathbf{v}_j^i\|_2^{p^*}\Big]^{\frac{1}{p^*}}.$$

By the Khitchine-Kahane inequality (69) with $p^*\le 2$ and the following elementary inequality

$$\sum_{j=1}^{c}|t_j|^{\tilde{p}} \le c^{1-\tilde{p}}\big(\sum_{j=1}^{c}|t_j|\big)^{\tilde{p}}, \forall 0<\tilde{p}\le 1,$$

we get

$$\frac{nc\mathfrak{R}_{nc}(\widetilde{H}_p)}{\Lambda} \le \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\Big[\sum_{j=1}^{c}\big(\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2\big)^{\frac{p^*}{2}}\Big]^{\frac{1}{p^*}} \tag{65}$$

$$\le \max_{\mathbf{v}^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\Big[c^{1-\frac{p^*}{2}}\Big(\sum_{j=1}^{c}\sum_{i=1}^{nc}\|\mathbf{v}_j^i\|_2^2\Big)^{\frac{p^*}{2}}\Big]^{\frac{1}{p^*}}$$

$$\le \sqrt{nc}c^{\frac{1}{p^*}-\frac{1}{2}}\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2 = \sqrt{n}c^{1-\frac{1}{p}}\max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2,$$

where we have used the inequality $\sum_{j=1}^{c}\|\mathbf{v}_j\|_2^2 \le \max_{i\in\mathbb{N}_n}\|\phi(\mathbf{x}_i)\|_2^2$ for all $\mathbf{v}\in\widetilde{S}$ in the last inequality.

The above upper and lower bounds in the two cases can be written compactly as (18). The proof is complete. □

D. Proofs on Applications

Proof of Example 1. According to the monotonicity of ℓ, there holds

$$\ell(\rho_h(\mathbf{x},y)) = \ell\big(\min_{y':y'\neq y}(h_y(\mathbf{x})-h_{y'}(\mathbf{x}))\big) = \max_{y':y'\neq y}\ell(h_y(\mathbf{x})-h_{y'}(\mathbf{x})) = \Psi_y^\ell(h(\mathbf{x})).$$

It remains to show the Lipschitz continuity of Ψ_y^ℓ. Indeed, for any $\mathbf{t},\mathbf{t}'\in\mathbb{R}^c$, we have

$$\begin{aligned}|\Psi_y^\ell(\mathbf{t})-\Psi_y^\ell(\mathbf{t}')| &= \big|\max_{y':y'\neq y}\ell(t_y-t_{y'})-\max_{y':y'\neq y}\ell(t'_y-t'_{y'})\big|\\ &\le \max_{y':y'\neq y}\big|\ell(t_y-t_{y'})-\ell(t'_y-t'_{y'})\big|\\ &\le \max_{y':y'\neq y}L_\ell|(t_y-t_{y'})-(t'_y-t'_{y'})|\\ &\le 2L_\ell\max_{y'\in\mathcal{Y}}|t_{y'}-t'_{y'}|\\ &\le 2L_\ell\|\mathbf{t}-\mathbf{t}'\|_2,\end{aligned}$$

where in the first inequality we have used the elementary inequality

$$|\max\{a_1,\dots,a_c\}-\max\{b_1,\dots,b_c\}| \le \max\{|a_1-b_1|,\dots,|a_c-b_c|\}, \quad \forall \boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^c \tag{66}$$

and the second inequality is due to the Lipschitz continuity of ℓ. □

Proof of Example 2. Define the function $f^m:\mathbb{R}^c\mapsto\mathbb{R}$ by $f^m(\mathbf{t}) = \log\big(\sum_{j=1}^{c}\exp(t_j)\big)$. For any $\mathbf{t}\in\mathbb{R}^c$, the partial gradient of f^m with respect to t_k is

$$\frac{\partial f^m(\mathbf{t})}{\partial t_k} = \frac{\exp(t_k)}{\sum_{j=1}^{c}\exp(t_j)}, \quad \forall k=1,\dots,c,$$

from which we derive that $\|\nabla f^m(\mathbf{t})\|_1 = 1, \forall\mathbf{t}\in\mathbb{R}^c$. Here ∇ denotes the gradient operator. For any $\mathbf{t},\mathbf{t}'\in\mathbb{R}^c$, according to the mean-value theorem we know the existence of $\alpha\in[0,1]$ such that

$$\begin{aligned}|f^m(\mathbf{t})-f^m(\mathbf{t}')| &= \big|\big\langle\nabla f^m(\alpha\mathbf{t}+(1-\alpha)\mathbf{t}'),\mathbf{t}-\mathbf{t}'\big\rangle\big|\\ &\le \|\nabla f^m(\alpha\mathbf{t}+(1-\alpha)\mathbf{t}')\|_1\|\mathbf{t}-\mathbf{t}'\|_\infty = \|\mathbf{t}-\mathbf{t}'\|_\infty.\end{aligned}$$

It then follows that

$$\begin{aligned}|\Psi_y^m(\mathbf{t})-\Psi_y^m(\mathbf{t}')| &= |f^m\big((t_j-t_y)_{j=1}^c\big)-f^m\big((t'_j-t'_y)_{j=1}^c\big)|\\ &\le \big\|(t_j-t_y)_{j=1}^c-(t'_j-t'_y)_{j=1}^c\big\|_\infty\\ &\le 2\|\mathbf{t}-\mathbf{t}'\|_\infty.\end{aligned}$$

That is, Ψ_y^m is 2-Lipschitz continuous w.r.t. the ℓ_∞-norm. □

Proof of Example 3. For any $\mathbf{t},\mathbf{t}'\in\mathbb{R}^c$, we have

$$\begin{aligned}|\tilde{\Psi}_y^\ell(\mathbf{t})-\tilde{\Psi}_y(\mathbf{t}')| &= \big|\sum_{j=1}^{c}\ell(t_y-t_j)-\sum_{j=1}^{c}\ell(t'_y-t'_j)\big|\\ &\le \sum_{j=1}^{c}\big|\ell(t_y-t_j)-\ell(t'_y-t'_j)\big|\\ &\le L_\ell c|t_y-t'_y|+L_\ell\sum_{j=1}^{c}|t_j-t'_j|\\ &\le L_\ell c|t_y-t'_y|+L_\ell\sqrt{c}\|\mathbf{t}-\mathbf{t}'\|_2.\end{aligned}$$

The Lipschitz continuity of $\tilde{\Psi}_y^\ell(\mathbf{t})$ w.r.t. ℓ_∞-norm is also clear. □

Proof of Example 4. For any $\mathbf{t},\mathbf{t}'\in\Omega$, we have

$$|\bar{\Psi}_y^\ell(\mathbf{t})-\bar{\Psi}_y^\ell(\mathbf{t}')| = \Big|\sum_{j=1,j\neq y}^{c}\big[\ell(-t_j)-\ell(-t'_j)\big]\Big|$$

$$\le L_\ell\sum_{j=1,j\neq y}^{c}|t_j-t'_j| \le L_\ell\sqrt{c}\|\mathbf{t}-\mathbf{t}'\|_2 \le L_\ell c\|\mathbf{t}-\mathbf{t}'\|_\infty.$$

This establishes the Lipschitz continuity of $\bar{\Psi}_y^\ell$. □

Proof of Example 5. For any $\mathbf{t},\mathbf{t}'\in\Omega$, we have

$$\big|\hat{\Psi}_y^\ell(\mathbf{t})-\hat{\Psi}_y^\ell(\mathbf{t}')\big| = |\ell(t_y)-\ell(t'_y)| \le L_\ell|t_y-t'_y| \le L_\ell\|\mathbf{t}-\mathbf{t}'\|_\infty.$$

This establishes the Lipschitz continuity of $\hat{\Psi}_y^\ell$. □

Proof of Example 6. It is clear that

$$\sum_{j=1}^{k}t_{[j]} = \max_{1\le i_1<i_2<\cdots<i_k\le c}[t_{i_1}+\cdots+t_{i_k}], \quad \forall\mathbf{t}\in\mathbb{R}^c. \tag{67}$$

For any $\mathbf{t}, \mathbf{t}' \in \mathbb{R}^c$, we have

$$
\begin{aligned}
&|\Psi_y^k(\mathbf{t}) - \Psi_y^k(\mathbf{t}')| \\
&\leq \frac{1}{k}\Big|\sum_{j=1}^{k}(1_{y\neq 1} + t_1 - t_y, \ldots, 1_{y\neq c} + t_c - t_y)_{[j]} \\
&\qquad - \sum_{j=1}^{k}(1_{y\neq 1} + t'_1 - t'_y, \ldots, 1_{y\neq c} + t'_c - t'_y)_{[j]}\Big| \\
&= \frac{1}{k}\Big|\max_{1\leq i_1 < i_2 < \cdots < i_k \leq c} \sum_{r=1}^{k}(1_{y\neq i_r} + t_{i_r} - t_y) \\
&\qquad - \max_{1\leq i_1 < i_2 < \cdots < i_k \leq c} \sum_{r=1}^{k}(1_{y\neq i_r} + t'_{i_r} - t'_y)\Big| \\
&\leq \frac{1}{k}\max_{1\leq i_1 < i_2 < \cdots < i_k \leq c}\Big|\sum_{r=1}^{k}(1_{y\neq i_r} + t_{i_r} - t_y) \\
&\qquad - \sum_{r=1}^{k}(1_{y\neq i_r} + t'_{i_r} - t'_y)\Big| \\
&\leq \frac{1}{k}\max_{1\leq i_1 < i_2 < \cdots < i_k \leq c}\Big|\sum_{r=1}^{k}(t_{i_r} - t'_{i_r})\Big| + |t_y - t'_y| \\
&\leq \frac{1}{\sqrt{k}}\max_{1\leq i_1 < i_2 < \cdots < i_k \leq c}\Big[\sum_{r=1}^{k}(t_{i_r} - t'_{i_r})^2\Big]^{\frac{1}{2}} + |t_y - t'_y| \qquad (68) \\
&\leq \frac{1}{\sqrt{k}}\Big[\sum_{j=1}^{c}(t_j - t'_j)^2\Big]^{\frac{1}{2}} + |t_y - t'_y|,
\end{aligned}
$$

where the first and the second inequality are due to (66) and the first identity is due to (67). This establishes the Lipschitz continuity w.r.t. a variant of the ℓ_2-norm. The 2-Lipschitz continuity of Ψ_y^k w.r.t. ℓ_∞-norm is clear from (68). The proof is complete. □

VII. Conclusion

Motivated by the ever-growing number of label classes in classification problems, we develop two approaches to derive data-dependent error bounds that scale favorably with the number of labels. The two approaches are based on the Gaussian and Rademacher complexities, respectively, of a related linear function class defined over a finite set induced from the training examples, for which we establish tight upper and lower bounds that match within a constant factor. Due to the ability to preserve the correlation among class-wise components, both of these data-dependent bounds admit an improved dependency on the number of classes over the state-of-the-art methods.

Our first approach is based on a novel structural result on the Gaussian complexities of function classes composed by Lipschitz operators measured by a variant of the ℓ_2-norm. We show the advantage of our structural result over the previous one (3) in [28, 43, 44] by better capturing the Lipschitz property of loss functions and yielding tighter bounds, which is the case for some popular MC-SVMs [30, 32, 45].

Our second approach is based on a novel structural result controlling the worst-case Rademacher complexity of the loss function class by the ℓ_∞-norm covering numbers of an associated linear function class. Our approach addresses the fact that several loss functions are Lipschitz continuous w.r.t. the ℓ_∞ norm with a moderate Lipschitz constant [48]. This allows us to obtain error bounds exhibiting a logarithmic dependency on the number of classes for the MC-SVM in Crammer and Singer [31] and MLR, significantly improving the existing square-root dependency [28, 48].

We show that each of these two approaches has its own advantages and can outperform the other for some applications depending on the Lipschitz continuity of the associated loss function. We report experimental results to show that our theoretical bounds capture the influence of class size on models' generalization performance, which in turn imply a structural risk that works well in model selection. Furthermore, we propose an efficient algorithm to train ℓ_p-norm MC-SVM based on the Frank-Wolfe algorithm.

We now present here some possible directions for future study. First, our generalization analysis gives generalization bounds with a logarithmic dependency for MLR and Crammer & Singer MC-SVM. It would be interesting to investigate whether this logarithmic dependency can be further relaxed to a class-size independency. Second, research in classification with many classes increasingly focuses on *multi-label* classification with each output $\mathbf{y}_i$ taking values in $\{0,1\}^c$ [18, 22, 77]. It would be interesting to transfer the results obtained in the present analysis to the multi-label case. To this aim, it is helpful to check the Lipschitz continuity of loss functions in multi-label learning, which, as in the present work, are typically of the form $\Psi_{\mathbf{y}}(h(\mathbf{x}))$ [77, 78], (e.g., Hamming loss, subset zero-one loss, and ranking loss [78]). Third, we study examples with the functional τ depending on the components of $\mathbf{w}$ in the RKHS. It would be interesting to consider examples with τ defined in other forms, such as those in [79, 80]. Fourth, our error bounds are derived for convex surrogates of the 0-1 loss. It would be interesting to relate these error bounds to excess generalization errors measured by the 0-1 loss [48, 59, 81, 82].

Acknowledgment

We are grateful to the associate editor and anonymous referees for their constructive comments. We thank Rohit Babbar, Alexander Binder, Moustapha Cisse, Vitaly Kuznetsov, Stephan Mandt, Mehryar Mohri and Robert Vandermeulen for interesting discussions.

YL acknowledges support from the National Key Research and Development Program of China (Grant No. 2017YFC0804003), the National Natural Science Foundation of China (Grant No. 61806091), the Shenzhen Peacock Plan (Grant No. KQTD2016112514355531), the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284) and the Alexander von Humboldt Foundation for a Humboldt Research Fellowship. DZ acknowledges support from the NSFC/RGC Joint Research Scheme [RGC Project No. N_C CityU120/14 and NSFC Project No. 11461161006]. MK acknowledges funding by the German Research Foundation (DFG) awards KL 2698/2-1 and GRK1589/2 and by the Federal Ministry of Science and Education (BMBF) awards 031L0023A, 01IS18051A.

APPENDIX A
KHINTCHINE-KAHANE INEQUALITY

The following Khintchine-Kahane inequality [83, 84] provides a powerful tool to control the p-th norm of the summation of Rademacher (Gaussian) series.

Lemma 24. *(a) Let $\mathbf{v}_1,\ldots,\mathbf{v}_n\in\mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated norm. Let $\epsilon_1,\ldots,\epsilon_n$ be a sequence of independent Rademacher variables. Then, for any $p\geq 1$ there holds*

$$\min(\sqrt{p-1},1)\big[\sum_{i=1}^n\|\mathbf{v}_i\|^2\big]^{\frac{1}{2}}\leq\big[\mathbb{E}_\epsilon\|\sum_{i=1}^n\epsilon_i\mathbf{v}_i\|^p\big]^{\frac{1}{p}}\leq\max(\sqrt{p-1},1)\big[\sum_{i=1}^n\|\mathbf{v}_i\|^2\big]^{\frac{1}{2}},\quad(69)$$

and

$$\mathbb{E}_\epsilon\|\sum_{i=1}^n\epsilon_i\mathbf{v}_i\|\geq 2^{-\frac{1}{2}}\big[\sum_{i=1}^n\|\mathbf{v}_i\|^2\big]^{\frac{1}{2}}.\quad(70)$$

The above inequalities also hold when the Rademacher variables are replaced by $N(0,1)$ random variables.

(b) Let $X_1,\ldots,X_n$ be a set of matrices of the same dimension and let $g_1,\ldots,g_n$ be a sequence of independent $N(0,1)$ random variables. For all $q\geq 2$,

$$\Big(\mathbb{E}_{\boldsymbol{g}}\|\sum_{i=1}^n g_iX_i\|_{S_q}^q\Big)^{\frac{1}{q}}\leq 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}}\times\max\Big\{\|(\sum_{i=1}^n X_i^\top X_i)^{\frac{1}{2}}\|_{S_q},\|(\sum_{i=1}^n X_iX_i^\top)^{\frac{1}{2}}\|_{S_q}\Big\}.\quad(71)$$

Proof. For Part (b), the original Khintchine-Kahane inequality for matrices is stated for Rademacher random variables, i.e, the Gaussian variables g_i are replaced by Rademacher variables ϵ_i. We now show that it also holds for Gaussian variables. Let $\psi_i^{(k)}=\frac{1}{\sqrt{k}}\sum_{j=1}^k\epsilon_{ik+j}$ with ϵ_{ik+j} being a sequence of independent Rademacher variables, then we have

$$\begin{aligned}\big(\mathbb{E}_\epsilon\|\sum_{i=1}^n\psi_i^{(k)}X_i\|_{S_q}^q\big)^{\frac{1}{q}}&=\big(\mathbb{E}_\epsilon\|\sum_{i=1}^n\sum_{j=1}^k\epsilon_{ik+j}\frac{1}{\sqrt{k}}X_i\|_{S_q}^q\big)^{\frac{1}{q}}\\&\leq\sqrt{\frac{q\pi}{2^{\frac{1}{2}}e}}\max\Big\{\|(\sum_{i=1}^n\sum_{j=1}^k\frac{X_i^\top X_i}{k})^{\frac{1}{2}}\|_{S_q},\|(\sum_{i=1}^n\sum_{j=1}^k\frac{X_iX_i^\top}{k})^{\frac{1}{2}}\|_{S_q}\Big\}\\&\leq\sqrt{\frac{q\pi}{2^{\frac{1}{2}}e}}\max\Big\{\|(\sum_{i=1}^n X_i^\top X_i)^{\frac{1}{2}}\|_{S_q},\|(\sum_{i=1}^n X_iX_i^\top)^{\frac{1}{2}}\|_{S_q}\Big\},\end{aligned}$$

where the first inequality is due to the Khintchine-Kahane inequality for matrices involving Rademacher random variables [84]. The proof is complete if we take k to ∞ and use central limit theorem. $\square$

APPENDIX B
PROOF OF PROPOSITION 8

We present the proof of Proposition 8 in the appendix due to its similarity to the proof of Proposition 7.

We first consider the case $1\leq p\leq 2$. Since the dual norm of $\|\cdot\|_{S_p}$ is $\|\cdot\|_{S_{p^*}}$, we have the following lower bound on RC in this case

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_{S_p})&=\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{1}{nc}\mathbb{E}_\epsilon\sup_{\|W\|_{S_p}\leq\Lambda}\sum_{i=1}^{nc}\epsilon_i\langle W,V^i\rangle\\&=\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{1}{nc}\mathbb{E}_\epsilon\sup_{\|W\|_{S_p}\leq\Lambda}\langle W,\sum_{i=1}^{nc}\epsilon_iV^i\rangle\\&=\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{\Lambda}{nc}\mathbb{E}_\epsilon\big\|\sum_{i=1}^{nc}\epsilon_iV^i\big\|_{S_{p^*}}.\end{aligned}\quad(72)$$

Taking $V^1=\cdots=V^{nc}$ and applying the Khitchine-Kahane inequality (70) further imply

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_{S_p})&\geq\max_{V^1\in\widetilde{S}}\frac{\Lambda}{nc}\mathbb{E}_\epsilon\big|\sum_{i=1}^{nc}\epsilon_i\big|\|V^1\|_{S_{p^*}}\\&\geq\frac{\Lambda}{\sqrt{2nc}}\max_{V^1\in\widetilde{S}}\|V^1\|_{S_{p^*}}=\frac{\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{2nc}},\end{aligned}$$

where the last identity follows from the following identity for any $V\in\widetilde{S}$

$$\|V\|_{S_{p^*}}=\|V\|_{S_2}=\|V\|_{2,2}=\|V\|_{2,\infty}.\quad(73)$$

We now turn to the upper bound. It follows from the relationship $\widetilde{H}_{S_p}\subset\widetilde{H}_{S_2},\forall 1\leq p\leq 2$ and (72) that ($\mathrm{tr}(A)$ denotes the trace of A)

$$\begin{aligned}\mathfrak{R}_{nc}(\widetilde{H}_{S_p})\leq\mathfrak{R}_{nc}(\widetilde{H}_{S_2})&=\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{\Lambda}{nc}\mathbb{E}_\epsilon\big\|\sum_{i=1}^{nc}\epsilon_iV^i\big\|_{S_2}\\&=\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{\Lambda}{nc}\mathbb{E}_\epsilon\sqrt{\mathrm{tr}\big(\sum_{i,\tilde{i}=1}^{nc}\epsilon_i\epsilon_{\tilde{i}}V^i(V^{\tilde{i}})^\top\big)}\\&\leq\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{\Lambda}{nc}\sqrt{\sum_{i=1}^{nc}\mathrm{tr}(V^i(V^i)^\top)}\\&=\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{\Lambda}{nc}\sqrt{\sum_{i=1}^{nc}\|V^i\|_{2,\infty}^2}\leq\frac{\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{nc}},\end{aligned}\quad(74)$$

where the second identity follows from the identity between Frobenius norm and $\|\cdot\|_{S_2}$, the second inequality follows from the Jensen's inequality and the last identity is due to (73).

We now consider the case $p>2$. According to the relationship $\widetilde{H}_{S_2}\subseteq\widetilde{H}_{S_p}$ for all $p>2$ and the discussion for the case $p=2$, we know

$$\mathfrak{R}_{nc}(\widetilde{H}_{S_p})\geq\mathfrak{R}_{nc}(\widetilde{H}_{S_2})\geq\frac{\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2}{\sqrt{2nc}}.$$

Furthermore, for any W with $\|W\|_{S_p}\leq\Lambda$ we have $\|W\|_{S_2}\leq\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}\Lambda$, which, combined with (74), implies that

$$\begin{aligned}&\mathfrak{R}_{nc}(\widetilde{H}_{S_p})\\&\leq\max_{V^i\in\widetilde{S}:i\in\mathbb{N}_{nc}}\frac{1}{nc}\mathbb{E}_\epsilon\sup_{\|W\|_{S_2}\leq\Lambda\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}\sum_{i=1}^{nc}\epsilon_i\langle W,V^i\rangle\\&\leq\frac{\Lambda\max_{i\in\mathbb{N}_n}\|\mathbf{x}_i\|_2\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{nc}}.\end{aligned}$$

The proof is complete.

APPENDIX C
PROOF OF PROPOSITION 19

It suffices to check $\|\mathbf{w}^*\|_{2,p} \leq 1$ and $\langle \mathbf{w}^*, \mathbf{v}\rangle = -\|\mathbf{v}\|_{2,p^*}$. We consider three cases.

If $p = 1$, it is clear that $\|\mathbf{w}^*\|_{2,1} \leq 1$ and $\langle \mathbf{w}^*, \mathbf{v}\rangle = -\|\mathbf{v}\|_{2,\infty}$.

If $p = \infty$, it is clear that $\|\mathbf{w}^*\|_{2,\infty} \leq 1$ and $\langle \mathbf{w}^*, \mathbf{v}\rangle = -\sum_{j=1}^{c} \|\mathbf{v}_j\|_2 = -\|\mathbf{v}\|_{2,1}$.

If $1 < p < \infty$, it is clear that

$$\|\mathbf{w}^*\|_{2,p} = \big(\sum_{\tilde{j}=1}^{c} \|\mathbf{v}_{\tilde{j}}\|_2^{(p^*-1)p}\big)^{\frac{1}{p}} / \big(\sum_{\tilde{j}=1}^{c} \|\mathbf{v}_{\tilde{j}}\|_2^{p^*}\big)^{\frac{1}{p}} = 1$$

and

$$\langle \mathbf{w}^*, \mathbf{v}\rangle = -\big(\sum_{\tilde{j}=1}^{c} \|\mathbf{v}_{\tilde{j}}\|_2^{p^*}\big)^{-\frac{1}{p}} \sum_{\tilde{j}=1}^{c} \|\mathbf{v}_{\tilde{j}}\|_2^{p^*} = -\|\mathbf{v}\|_{2,p^*}.$$

The proof is complete.

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998, vol. 1.

[2] S. Har-Peled, D. Roth, and D. Zimak, "Constraint classification: A new approach to multiclass classification," in *Algorithmic Learning Theory*. Springer, 2002, pp. 365–379.

[3] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[4] U. Dogan, T. Glasmachers, and C. Igel, "A unified view on multi-class support vector classification," *Journal of Machine Learning Research*, vol. 17, no. 45, pp. 1–32, 2016.

[5] N. Kato, M. Suzuki, S. I. Omachi, H. Aso, and Y. Nemoto, "A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 3, pp. 258–262, 1999.

[6] A. Voutilainen, "Part-of-speech tagging," *The Oxford Handbook of Computational Linguistics*, pp. 219–232, 2003.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[8] A. Binder, K.-R. Müller, and M. Kawanabe, "On taxonomies for multi-class image categorization," *International Journal of Computer Vision*, vol. 99, no. 3, pp. 281–301, 2012.

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

[10] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *Annals of Statistics*, pp. 1–50, 2002.

[11] Y. Guermeur, "Combining discriminant models with new multi-class SVMs," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 168–179, 2002.

[12] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[13] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits," 1998.

[14] A. Asuncion and D. Newman, "UCI machine learning repository," 2007.

[15] I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari, "Lshtc: A benchmark for large-scale text classification," *arXiv preprint arXiv:1503.08581*, 2015.

[16] M. Varma and J. Langford, "NIPS Workshop on eXtreme Classification," 2013. [Online]. Available: https://manikvarma.github.io/events/XC13

[17] B. Varadarajan, G. Toderici, S. Vijayanarasimhan, and A. Natsev, "Efficient large scale video classification," *arXiv preprint arXiv:1505.06250*, 2015.

[18] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Advances in Neural Information Processing Systems*, 2015, pp. 730–738.

[19] S. Bengio, J. Weston, and D. Grangier, "Label embedding trees for large multi-class tasks," in *Advances in Neural Information Processing Systems*, 2010, pp. 163–171.

[20] A. Beygelzimer, J. Langford, Y. Lifshits, G. Sorkin, and A. Strehl, "Conditional probability tree estimation analysis and algorithms," in *Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 51–58.

[21] S. Sedhai and A. Sun, "Hspam14: A collection of 14 million tweets for hashtag-oriented spam research," in *ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 223–232.

[22] H. Jain, Y. Prabhu, and M. Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 935–944.

[23] R. Babbar, I. Partalas, E. Gaussier, M.-R. Amini, and C. Amblard, "Learning taxonomy adaptation in large-scale classification," *Journal of Machine Learning Research*, vol. 17, no. 98, pp. 1–37, 2016.

[24] Y. Prabhu and M. Varma, "Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 263–272.

[25] M. Alber, J. Zimmert, U. Dogan, and M. Kloft, "Distributed optimization of multi-class SVMs," *PloS one*, vol. 12, no. 6, p. e0178161, 2017.

[26] R. Babbar, K. Maundet, and B. Schölkopf, "Tersesvm: A scalable approach for learning compact models in large-scale classification," in *SIAM International Conference on Data Mining*. SIAM, 2016, pp. 234–242.

[27] M. Varma and M. Cisse, "NIPS Workshop on eXtreme Classification," 2015. [Online]. Available: https://manikvarma.github.io/events/XC15

[28] Y. Lei, U. Dogan, A. Binder, and M. Kloft, "Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms," in *Advances in Neural Information Processing Systems*, 2015, pp. 2026–2034.

[29] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[30] M. Lapin, M. Hein, and B. Schiele, "Top-k multiclass SVM," in *Advances in Neural Information Processing Systems*, 2015, pp. 325–333.

[31] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2002.

[32] J. Weston and C. Watkins, "Multi-class support vector machines," Citeseer, Tech. Rep., 1998.

[33] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, 2004.

[34] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-iid processes," in *Advances in Neural Information Processing Systems*, 2009, pp. 1097–1104.

[35] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *Journal of Multivariate Analysis*, vol. 100, no. 1, pp. 175–194, 2009.

[36] V. Koltchinskii, "Rademacher penalties and structural risk min-

imization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[37] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[38] S. Mendelson, "Rademacher averages and phase transitions in glivenko-cantelli classes," *IEEE Transactions on Information Theory*, vol. 48, no. 1, pp. 251–263, 2002.

[39] C. Cortes, M. Kloft, and M. Mohri, "Learning kernels using local rademacher complexity," in *Advances in Neural Information Processing Systems*, 2013, pp. 2760–2768.

[40] C. Cortes, M. Mohri, and A. Rostamizadeh, "Multi-class classification with maximum margin multiple kernel," in *International Conference on Machine Learning*, 2013, pp. 46–54.

[41] V. Kuznetsov, M. Mohri, and U. Syed, "Multi-class deep boosting," in *Advances in Neural Information Processing Systems*, 2014, pp. 2501–2509.

[42] D. Slepian, "The one-sided barrier problem for gaussian noise," *Bell System Technical Journal*, vol. 41, no. 2, pp. 463–501, 1962.

[43] A. Maurer, "A vector-contraction inequality for rademacher complexities," in *International Conference on Algorithmic Learning Theory*. Springer, 2016, pp. 3–17.

[44] C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang, "Structured prediction theory based on factor graph complexity," in *Advances in Neural Information Processing Systems*, 2016, pp. 2514–2522.

[45] R. Jenssen, M. Kloft, A. Zien, S. Sonnenburg, and K.-R. Müller, "A scatter-based prototype framework and multi-class extension of support vector machines," *PloS one*, vol. 7, no. 10, p. e42947, 2012.

[46] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Science & Business Media, 2008.

[47] V. N. Vapnik and A. Y. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[48] T. Zhang, "Statistical analysis of some multi-category large margin classification methods," *Journal of Machine Learning Research*, vol. 5, pp. 1225–1251, 2004.

[49] Z.-W. Pan, D.-H. Xiang, Q.-W. Xiao, and D.-X. Zhou, "Parzen windows for multi-class classification," *Journal of complexity*, vol. 24, no. 5, pp. 606–618, 2008.

[50] Y. Guermeur, "Sample complexity of classifiers taking values in $\mathbb{R}^q$, application to multi-class SVMs," *Communications in Statistics Theory and Methods*, vol. 39, no. 3, pp. 543–557, 2010.

[51] ——, "VC theory of large margin multi-category classifiers," *Journal of Machine Learning Research*, vol. 8, no. Nov, pp. 2551–2594, 2007.

[52] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz, "Multiclass learnability and the erm principle," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2377–2404, 2015.

[53] A. Daniely, S. Sabato, and S. S. Shwartz, "Multiclass learning approaches: A theoretical comparison with implications," in *Advances in Neural Information Processing Systems*, 2012, pp. 485–493.

[54] B. K. Natarajan, "On learning sets and functions," *Machine Learning*, vol. 4, no. 1, pp. 67–97, 1989.

[55] Y. Guermeur, "Lp-norm sauer-shelah lemma for margin multi-category classifiers," *Journal of Computer and System Sciences*, vol. 89, pp. 450–473, 2017.

[56] A. Kontorovich and R. Weiss, "Maximum margin multiclass nearest neighbors," in *International Conference on Machine Learning*, 2014, pp. 892–900.

[57] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *International Conference on Machine Learning*. ACM, 2007, pp. 17–24.

[58] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics*, vol. 3, no. 1-2, pp. 95–110, 1956.

[59] A. Tewari and P. L. Bartlett, "On the consistency of multiclass classification methods," *Journal of Machine Learning Research*, vol. 8, pp. 1007–1025, 2007.

[60] T. Zhang, "Covering number bounds of certain regularized linear function classes," *Journal of Machine Learning Research*, vol. 2, pp. 527–550, 2002.

[61] D.-X. Zhou, "The covering number in learning theory," *Journal of Complexity*, vol. 18, no. 3, pp. 739–767, 2002.

[62] ——, "Capacity of reproducing kernel spaces in learning theory," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1743–1752, 2003.

[63] M. Lapin, M. Hein, and B. Schiele, "Loss functions for top-k error: Analysis and insights," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1468–1477.

[64] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[65] K. Lang, "Newsweeder: Learning to filter netnews," in *International Conference on Machine Learning*, 1995, pp. 331–339.

[66] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[67] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Madison, WI, 1998, pp. 41–48.

[68] J.-M. Geusebroek, G. J. Burghouts, and A. W. Smeulders, "The Amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[69] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[70] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *International Conference on Machine Learning*, 2013, pp. 427–435.

[71] J. Nocedal and S. J. Wright, "Numerical optimization," 2006.

[72] C. McDiarmid, "On the method of bounded differences," in *Surveys in combinatorics*, J. Siemous, Ed. Cambridge: Cambridge Univ. Press, 1989, pp. 148–188.

[73] A. Tewari and S. Chaudhuri, "Generalization error bounds for learning to rank: Does the length of document lists matter?" in *International Conference on Machine Learning*, 2015, pp. 315–323.

[74] N. Srebro, K. Sridharan, and A. Tewari, "Smoothness, low noise and fast rates," in *Advances in Neural Information Processing Systems*, 2010, pp. 2199–2207.

[75] A. Rakhlin, K. Sridharan, and A. Tewari, "Sequential complexities and uniform martingale laws of large numbers," *Probability Theory and Related Fields*, vol. 161, no. 1-2, pp. 111–153, 2014.

[76] T. Tao, *Topics in random matrix theory*. American Mathematical Soc., 2012, vol. 132.

[77] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels." in *International Conference on Machine Learning*, 2014, pp. 593–601.

[78] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "On label dependence and loss minimization in multi-label classification," *Machine Learning*, vol. 88, no. 1-2, pp. 5–45, 2012.

[79] L. Shi, Y.-L. Feng, and D.-X. Zhou, "Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces," *Applied and Computational Harmonic Analysis*, vol. 31, no. 2, pp. 286–302, 2011.

[80] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou, "Thresholded spectral algorithms for sparse approximations," *Analysis and Applications*, vol. 15, no. 03, pp. 433–455, 2017.

[81] T. Zhang, "Statistical behavior and consistency of classification

methods based on convex risk minimization," *Annals of Statistics*, pp. 56–85, 2004.

[82] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[83] V. De la Pena and E. Giné, *Decoupling: from Dependence to Independence*. Springer Science & Business Media, 2012.

[84] F. Lust-Piquard and G. Pisier, "Non commutative khintchine and paley inequalities," *Arkiv för Matematik*, vol. 29, no. 1, pp. 241–260, 1991.

Yunwen Lei received his Ph.D. degree in computer science in 2014 from Wuhan University, Wuhan, China. From 2015 to 2017, he was a postdoctoral research fellow at Department of Mathematics, City University of Hong Kong. He is currently a research assistant professor at Department of Computer Science and Engineering, Southern University of Science and Technology. His main research interests include machine learning, statistical learning theory and convex optimization.

Ürün Dogan is a machine learning researcher at Microsoft. Previously he was a postdoctoral researcher at University of Potsdam. He earned his Ph.D. degree from University of Bochum.

Ding-Xuan Zhou received his B.Sc. and Ph.D. degrees in mathematics in 1988 and 1991, respectively, from Zhejiang University, Hangzhou, China. He joined the faculty of City University of Hong Kong in 1996, and is currently a Chair Professor in the School of Data Science and Department of Mathematics. His research interests include deep learning, learning theory, data science, wavelet analysis and approximation theory. He has published over 100 journal papers, is serving on editorial board of more than 10 international journals, and is the Editor-in-Chief of the journal "Analysis and Application". He was rated in 2014–2017 by Thomson Reuters/Clarivate Analytics as a Highly-cited Researcher.

Marius Kloft is a professor of computer science at TU Kaiserslautern and an adjunct faculty member of the University of Southern California. Previously he was a junior professor at HU Berlin and a joint postdoctoral fellow at the Courant Institute of Mathematical Sciences and Memorial Sloan-Kettering Cancer Center, New York. He earned his Ph.D. degree at TU Berlin and UC Berkeley.