Multi-Class Learning using Unlabeled Samples: Theory and Algorithm

Jian Li[1,2], **Yong Liu**[1*], **Rong Yin**[1,2] and **Weiping Wang**[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{lijian9026, liuyong, yinrong, wangweiping}@iie.ac.cn

Abstract

In this paper, we investigate the generalization performance of multi-class classification, for which we obtain a shaper error bound by using the notion of local Rademacher complexity and additional unlabeled samples, substantially improving the state-of-the-art bounds in existing multi-class learning methods. The statistical learning motivates us to devise an efficient multi-class learning framework with the local Rademacher complexity and Laplacian regularization. Coinciding with the theoretical analysis, experimental results demonstrate that the stated approach achieves better performance.

1 Introduction

Multi-class classification is an import task in machine learning, with numerous applications such as text categorization, image annotation, etc. Estimating the generalization performance of algorithms is useful for understanding the factors that influence their behavior, as well as suggesting ways to improve them. Many works have studied generalization ability of supervised multi-class classification, but there are still a lot of statistical challenges in the semi-supervised case.

To study the generalization ability of multi-class classification algorithms, researchers have developed useful tools to measure the richness of the hypothesis space, including data-independent measures and data-dependent measures. Data-independent measures, such as VC-dimension [Allwein *et al.*, 2000] and Natarajan dimension [Daniely *et al.*, 2015], typically provide conservative multi-class bounds. As one of the most successful data-dependent complexity measures, the Rademacher complexity was first introduced into the multi-class setting in [Koltchinskii *et al.*, 2001] and further studied in [Cortes *et al.*, 2013; Maximov and Reshetova, 2016]. The convergence rates of multi-class classification bounds using Rademacher complexity are $\mathcal{O}(K/\sqrt{n})$ at best, where K and n are the number of classes and the size of labeled samples, respectively. A delicate error bound which exhibits logarithmic dependence on the class size was proposed in [Lei *et al.*, 2015]. While the global Rademacher complexity captures the complexity of the entire class, the local Rademacher complexity is used to choose a favorable subset from the hypothesis space [Bartlett *et al.*, 2005], always leading better statistical properties. The state-of-the-art error bound of kernel-based multi-class classification was established in [Li *et al.*, 2018] by using the local Rademacher complexity.

In this paper, we derive a novel data-dependent generalization error bound for multi-class classification in linear space by using the notion of the local Rademacher complexity and additional unlabeled samples. The convergence rate of the bound is $\mathcal{O}\big(K/\sqrt{n+u}+1/n\big)$, where u is the number of unlabeled samples, which is much faster than the rate of common bounds $\mathcal{O}(K/\sqrt{n})$. Further, motived by the statistical analysis, we devise an efficient multi-class classification algorithm by combining the local Rademacher complexity and unlabeled samples. Our approach improves both computational efficiency and statistical guarantee. Computational gains come from linear multi-class estimator and stochastic gradient descent algorithm on the primal form. Statistical gains lie in a smaller hypothesis space by using the local Rademacher complexity and unlabeled samples. Experiential learning reveals our approach outperforms other linear multi-class algorithms with or without unlabeled data.

1.1 Related Work

By using the local Rademacher complexity, a sharper multi-class error bound with fast rate $\mathcal{O}\big(\log^2 K/n\big)$ was reached in [Li *et al.*, 2018]. But it focused on supervised settings and kernel-based estimators which limits its applications in real-world datasets. For partially labeled data in multi-class settings, much progress has been accomplished in algorithmic front, but there are still a lot of challenges in theoretical front. Theoretical results for semi-supervised learning mainly consider the binary case, such as generalization analysis based on the global Rademacher complexity [Balcan and Blum, 2010; Oneto *et al.*, 2011] and based on the local Rademacher complexity [Oneto *et al.*, 2015]. In semi-supervised margin-based multi-class learning, the global Rademacher complexity for multi-class classifier trained with a two-step semi-supervised model is exploited in [Maximov *et al.*, 2018], of which the convergence rate is $\mathcal{O}\big(\sqrt{K/n}+K\sqrt{K/u}\big)$. In this work, we derive a much shaper generalization bound for multi-class classification with fast rate $\mathcal{O}\big(K/\sqrt{n+u}+1/n\big)$, by using the local Rademacher complexity and unlabeled samples.

*Corresponding author

2 Problem Definition

In a standard semi-supervised learning setting, a set of labeled samples $\mathcal{D}_l = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ are i.i.d. sampled from distribution μ over $\mathcal{X} \times \mathcal{Y}$ and unlabeled ones $\mathcal{D}_u = \{\mathbf{x}_1, \cdots, \mathbf{x}_u\}$ are i.i.d. drawn according to the marginal distribution μ_X of μ over $\mathcal{X}$, typically $n \ll u$. Further, we consider multi-class classification with $K \geq 2$ categories, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{1, 2, \cdots, K\}$.

2.1 Maximum Margin Multi-Class Classification

To evaluate the probability of $\mathbf{x}$ belonging to each category, we wish to learn a scoring rule from the hypothesis space $\mathcal{H}$

$$h(\mathbf{x}) = \mathbf{W}^T\mathbf{x},$$

where $h \in \mathcal{H}, \mathbf{W} \in \mathbb{R}^{d\times K}$ and $\mathbf{x} \in \mathbb{R}^d$, thus h is a vector-valued function with mapping $\mathcal{X} \to \mathbb{R}^K$. The predictor uses the following mapping to predict labels

$$\mathbf{x} \to \arg\max_y h(\mathbf{x}, y),$$

where $h(\mathbf{x}, y) = [\mathbf{W}^T\mathbf{x}]_y$ means the y-th value in vector $\mathbf{W}^T\mathbf{x}$. For any hypothesis $h \in \mathcal{H}$, the margin of a labeled example $(\mathbf{x}, y)$ is defined as

$$\rho_h(\mathbf{x}, y) = h(\mathbf{x}, y) - \max_{y\neq y'} h(\mathbf{x}, y').$$

The h misclassifies example $(\mathbf{x}, y)$ if $\rho_h(\mathbf{x}, y) \leq 0$, thus the expected risk incurred from using h for prediction is $L(h) := \mathbb{E}_\mu[1_{\rho(\mathbf{x},y)\leq 0}]$, where $1_{t\leq 0}$ is the 0-1 loss. Since 0-1 loss is noncontinuous thus hard to deal with, we consider the popular hinge loss $\ell(\rho_h(\mathbf{x}, y)) = |1-\rho_h(\mathbf{x}, y)|_+$ which upper bounds the 0-1 loss and its smooth extension the square hinge loss $\ell(\rho_h(\mathbf{x}, y)) = (1 - \rho_h(\mathbf{x}, y))_+^2$.

Let the expected loss be $L(\ell) = \mathbb{E}_\mu[\ell(\rho_h(\mathbf{x}, y))]$ and empirical loss be $\widehat{L}(\ell) = \frac{1}{n}\sum_{i=1}^n \ell(\rho_h(\mathbf{x}, y))$ with respect to h. The loss space associated with $\mathcal{H}$ is defined as bellow

$$\mathcal{L} = \{\ell(\rho_h(\mathbf{x}, y))|h \in \mathcal{H}\}.$$

In theoretical analysis, we use two standard assumptions:

(1) Any $\ell(\rho_h(\mathbf{x}, y))$ is continuous and bounded in $[0, 1]$, satisfied by normalized $h(\mathbf{x})$.

(2) ℓ is L-Lipschitz continuous, such that

$$|\ell(\rho_h(\mathbf{x}, y)) - \ell(\rho_h(\mathbf{x}', y'))| \leq L|\rho_h(\mathbf{x}, y) - \rho_h(\mathbf{x}', y')|.$$

Note that both hinge loss and square hinge loss satisfy the above two assumptions.

Remark 1. *Although the estimator is defined in linear space, it can be extended into reproducing kernel Hilbert space by feature mapping $\phi(\cdot)$, i.e., $\kappa(\mathbf{x}, \mathbf{x}') = \langle\phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$ where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel with ϕ being the associated feature mapping.*

2.2 Definitions

Before theoretical analysis, we introduce some definitions.

Definition 1. *Consider a local normalized loss space $\mathcal{L}_r$*

$$\mathcal{L}_r = \left\{\alpha\ell|\alpha \in [0, 1], \ell \in \mathcal{L}, L[(\alpha\ell)^2] \leq r\right\},$$

where $L[(\alpha\ell)^2] = \mathbb{E}_\mu[\ell^2(\rho_h(\mathbf{x}, y))]$. And the corresponding hypothesis space is defined as

$$\mathcal{H}_r = \{h|\ell(\rho_h(\mathbf{x}, y)) \in \mathcal{L}_r\}.$$

Definition 2. *The empirical Rademacher complexity of loss space $\mathcal{L}_r$ on labeled data and all data are*

$$\widehat{\mathcal{R}}_n(\mathcal{L}_r) = \mathbb{E}_\sigma \sup_{\ell\in\mathcal{L}_r} \frac{1}{n}\sum_{i=1}^n \sigma_i\ell(h(\mathbf{x}_i, y_i)),$$

$$\widehat{\mathcal{R}}(\mathcal{L}_r) = \mathbb{E}_\sigma \sup_{\ell\in\mathcal{L}_r} \frac{1}{n+u}\sum_{i=1}^{n+u} \sigma_i\ell(h(\mathbf{x}_i, y_i^\circ)),$$

where $\sigma_1, \sigma_2, \cdots, \sigma_{n+u}$ are $\{\pm 1\}$-valued independent Rademacher random variables with probability $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Their deterministic counterparts are $\mathcal{R}_n(\mathcal{L}_r) = \mathbb{E}_\mu\widehat{\mathcal{R}}_n(\mathcal{L}_r)$ and $\mathcal{R}(\mathcal{L}_r) = \mathbb{E}_\mu\widehat{\mathcal{R}}(\mathcal{L}_r)$.

Definition 3. *The empirical local Rademacher complexity of hypothesis space $\mathcal{H}_r$ on all data is defined as*

$$\widehat{\mathcal{R}}(\mathcal{H}_r) = \mathbb{E}_\sigma \sup_{h\in\mathcal{H}_r} \frac{1}{n+u}\sum_{i=1}^{n+u} \sigma_i h(\mathbf{x}_i, y_i^\circ),$$

where $\sigma_1, \sigma_2, \cdots, \sigma_{n+u}$ are $\{\pm 1\}$-valued independent Rademacher random variables with probability $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. The expected counterpart on all data is $\mathcal{R}(\mathcal{H}_r) = \mathbb{E}_\mu\widehat{\mathcal{R}}(\mathcal{H}_r)$.

Remark 2. *For $i \in \{1, \cdots, n\}$, labels of y_i° corresponds to labels in $\mathcal{D}_l$. For $i \in \{n + 1, \cdots, n + u\}$, y_i° are pseudo-labeled, as a consequence $\widehat{\mathcal{R}}(\mathcal{H}_r)$ actually is not computable.*

3 Shaper Generalization Error Bound

In this section, we present a shaper generalization error bound for multi-class classification by using the local Rademacher complexity and unlabeled data and consider the supervised case in Corollary 1. Proof details are deferred in Section 7.

Theorem 1. *For any $\ell \in \mathcal{L}_r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, $\forall k > 1, \|\mathbf{W}\| \leq 1$ and $\forall\, \delta \in (0, 1)$, the following holds with probability at least $1 - \delta$,*

$$L(\ell) \leq \max\Big\{\frac{k}{k-1}\widehat{L}(\ell), \widehat{L}(\ell) + \frac{c_1}{n} + \frac{c_2}{n+u} + \frac{c_3K\sum_{j>\theta}\lambda_j(\mathbf{W})}{\sqrt{n+u}}\Big\},$$

where $c_1 = (3 + 4k)\log(1/\delta)$, $c_2 = 32k\theta$ and $c_3 = 64kL$, $\lambda_j(\mathbf{W})$ is the j largest singular value of matrix $\mathbf{W}$.

Note that the bound partially depend on $\sum_{j>\theta}\lambda_j(\mathbf{W})$ which represents the tail sum of singular values, while the global Rademacher complexity bounds depend on the trace. The convergence rate of above bound presented in Theorem 1 is $\mathcal{O}(1/n)$ or $\mathcal{O}(K\sum_{j>\theta}\lambda_j(\mathbf{W})/\sqrt{n+u})$, much faster than common rate $\mathcal{O}(K/\sqrt{n})$. When there is no unlabeled data, namely $u = 0$, the result in Corollary 1 reduces to the state-of-the-art multi-class bound [Li *et al.*, 2018].

Corollary 1. *For any $\ell \in \mathcal{L}_r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, $\forall k > 1, \|W\| \leq 1$ and $\forall\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$L(\ell) \leq \max\Big\{\frac{k}{k-1}\widehat{L}(\ell), \widehat{L}(\ell) + \frac{c_1 + c_2}{n} + \frac{c_3K\sum_{j>\theta}\lambda_j(\mathbf{W})}{\sqrt{n}}\Big\},$$

where $c_1 = (3 + 4k)\log(1/\delta)$, $c_2 = 32k\theta$ and $c_3 = 64kL$.

3.1 Comparison with Other Bounds

VC dimension Bound

VC-dimension is a general distribution free complexity measure in statistical learning. VC-dimension was applied to generalization analysis in multi-class area [Allwein *et al.*, 2000] and a data-independent bound is derived: $L(\ell) \leq \widehat{L}(\ell) + \mathcal{O}\big(\sqrt{V}\log K/\sqrt{n}\big)$, where V is the VC-dimension.

Rademacher Complexity Bounds

In terms of the global Rademacher complexity, a data dependent margin-based multi-class classification bound was proposed in [Koltchinskii and Panchenko, 2002], of which the convergence rate is $\mathcal{O}\big(K^2/\sqrt{n}\big)$. Moreover, the bound was improved to $\mathcal{O}\big(K/\sqrt{n}\big)$ in [Maximov and Reshetova, 2016]. Lei et al. stated a bound reducing the dependence on the class size [Lei *et al.*, 2015], that is $\mathcal{O}\big((\log K)/\sqrt{n}\big)$. The global Rademacher complexity multi-class bounds were extended into semi-supervised case in [Maximov *et al.*, 2018], of which the convergence rate is $\mathcal{O}\big(\sqrt{K/n} + K\sqrt{K/u}\big)$.

Local Rademacher Complexity Bounds

As we all know, the local Rademacher complexity was firstly presented in binary settings and obtained shaper error bounds [Bartlett *et al.*, 2005]. Furthermore, it was extended into multi-class learning in [Li *et al.*, 2018] and derived a sharper data dependent bound only using labeled data.

In this paper, we propose a novel local Rademacher complexity bound for multi-class learning using both labeled and unlabeled samples in Theorem 1. When there is no unlabeled samples available, we derive Corollary 1, achieving similar bounds as [Li *et al.*, 2018]. (1) In common case, we observe that the bound shown in Theorem 1 is at most of order $\mathcal{O}\big(K/\sqrt{n+u} + 1/n\big)$, while convergence rate of others is usually at $\mathcal{O}(K/\sqrt{n})$, such that **faster convergence rate** is obtained in the common case. (2) In the special case that the singular values decrease exponentially, the convergence rate is $\mathcal{O}\big((c_1+c_2)\log^2 K/n\big)$ in [Li *et al.*, 2018], while it is $\mathcal{O}\big(c_1/n\big)$ in Theorem 1, thus much **smaller constant** on is derived in the special case. Therefore, the proposed linear multi-class approach improves current multi-class bounds with **faster convergence rate** or **smaller constant**. Table 1 reports statistical properties of related approaches and ours.

Bounds	Common Case	Special Case
[Allwein *et al.*, 2000]	$\mathcal{O}\big(\frac{\sqrt{V}\log K}{\sqrt{n}}\big)$	
[Cortes *et al.*, 2013]	$\mathcal{O}\big(\frac{K}{\sqrt{n}}\big)$	
[Maximov *et al.*, 2018]†	$\mathcal{O}\big(\sqrt{\frac{K}{n}} + K\sqrt{\frac{K}{u}}\big)$	
[Li *et al.*, 2018]	$\mathcal{O}\big((c_1+c_2)\frac{\log^2 K}{n}\big)$	
Theorem 1†	$\mathcal{O}\big(\frac{K}{\sqrt{n+u}} + \frac{1}{n}\big)$	$\mathcal{O}\big(\frac{c_1}{n}\big)$

Table 1: Comparison of multi-class classification error bounds, including one VC-dimension bound, two global Rademacher complexity bounds, and two local Rademacher complexity bounds. Here $n \ll u$, $K \ll n$ and † represents making use of unlabeled data.

4 Algorithms

4.1 Previous Works

Consider a similarity matrix $\mathbf{S}$ on entire $n+u$ examples and the weight S_{ij} represents the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, for example, kernel weights $S_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$. Motivated by continuity assumption that similar points share same label, the cost function for multi-class classification is

$$E(h) = \sum_{i,j=1}^{n+u} S_{ij}\|h(\mathbf{x}_i) - h(\mathbf{x}_j)\|_2^2 = \mathrm{trace}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}),$$

where $\mathbf{X} \in \mathbb{R}^{d\times(n+u)}$, graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{S}$ and $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n+u} S_{ij}$.

To make use of unlabeled samples, some novel algorithms are developed for multi-class learning in semi-supervised settings. Liu and Chang proposed transductive semi-supervised multi-class learning by minimizing cost function $E(h)$ to propagate labels [Liu and Chang, 2009]. Instead of transductive learning, manifold regularization is introduced into semi-supervised multi-class learning, which minimizes the cost function together with empirical error and the penalty of estimator complexity [Li and Guo, 2015; Li *et al.*, 2017b].

4.2 Optimization

Theoretical analysis demonstrates that both the smaller tail sum of singular values $\sum_{j>\theta}\lambda_j(\mathbf{W})$ and the number of unlabeled samples can improve generalization performance. Therefore, instead of traditional regularized empirical risk minimization, we consider minimizing the combination of the empirical loss, the penalty term on estimator complexity, the local Rademacher complexity and Laplacian regularization term in the following form

$$\begin{aligned}\arg\min_{h\in\mathcal{H}_r} \quad & \frac{1}{n}\sum_{i=1}^{n}\ell(h(\mathbf{x}_i), y_i) + \tau_A\|\mathbf{W}\|_F^2 \\ & +\tau_I\mathrm{trace}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) + \tau_S\sum_{j>\theta}\lambda_j(\mathbf{W}),\end{aligned} \tag{1}$$

where $\ell(h(\mathbf{x}_i), y_i) = |1-(h(\mathbf{x}_i, y_i) - \max_{y'\neq y_i} h(\mathbf{x}_i, y'))|_+$, $\lambda_j(\mathbf{W})$ denotes the j-th largest singular value of $\mathbf{W} \in \mathbb{R}^{d\times K}$, and τ_A, τ_I and τ_S are regularization parameters. τ_I can be set zero when there is no unlabeled data available.

Similar to the minimization of matrix trace, minimization of the tail sum of singular values is also nonconvex, thus the optimization problem in Eq. (1) is nonconvex. But minimizing the sum of a part of singular values is quite different from minimizing the sum of all singular values. Generalized SVT algorithms with two-step updating were designed [Lu *et al.*, 2015; Xu *et al.*, 2016] to minimize the sum of a part of singular values. Based on generalized SVT methods, we devise a proximal stochastic sub-gradient singular value thresholding multi-class learning framework, named `PS3VT` which is shown in Algorithm 1. The algorithm `PS3VT` updates $\mathbf{W}$ twice in each iteration, firstly updating $\mathbf{W}$ according to first-order sub-gradient of terms except for the tail sum of singular values, and then updating $\mathbf{W}$ with a closed form solution given by singular value thresholding (SVT).

4.3 SVT with Proximal Gradient

For the sake of simplification, we rewrite optimization (1) as

$$\begin{aligned}\arg\min_{h\in\mathcal{H}_r}\ & \tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})+g(\mathbf{W}) \qquad \text{where}\\ g(\mathbf{W}) =& \frac{1}{n}\sum_{i=1}^{n}\overbrace{|1-([\mathbf{W}^T\mathbf{x}_i]_{y_i}-\max_{y'\neq y_i}[\mathbf{W}^T\mathbf{x}_i]_{y'})|_+}^{\omega(\mathbf{W},\mathbf{x}_i)} \\ &+\tau_A\|\mathbf{W}\|_F^2+\tau_I\text{trace}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}).\end{aligned} \tag{2}$$

In each iteration, to obtain a tight surrogate of Eq. (2), we keep $\tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})$ while relaxing $g(\mathbf{W})$ only, that leads

$$\begin{aligned}\mathbf{W}^{t+1} =& \arg\min_{\mathbf{W}}\tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})+g(\mathbf{W})\\ =& \arg\min_{\mathbf{W}}\tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})+g(\mathbf{W}^t)\\ &+\langle\nabla g(\mathbf{W}^t),\mathbf{W}-\mathbf{W}^t\rangle+\frac{\mu}{2}\|\mathbf{W}-\mathbf{W}^t\|_F^2 \\ =& \arg\min_{\mathbf{W}}\tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})\\ &+\frac{\mu}{2}\|\mathbf{W}-(\mathbf{W}^t-\frac{1}{\mu}\nabla g(\mathbf{W}^t)\|_F^2,\end{aligned} \tag{3}$$

where $\frac{1}{\mu}$ actually is the step size to update gradients.

Proposition 1 (Theorem 6 of [Xu *et al.*, 2016]). *Let $\mathbf{Q}\in\mathbb{R}^{d\times K}$ with rank r and its SVD decomposition is $\mathbf{Q}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}\in\mathbb{R}^{d\times r}$ and $\mathbf{V}\in\mathbb{R}^{K\times r}$ have orthogonal columns, $\mathbf{\Sigma}$ is diagonal. Then,*

$$\mathcal{D}_\tau^\theta(\mathbf{Q})=\arg\min_{\mathbf{W}}\left\{\frac{1}{2}\|\mathbf{W}-\mathbf{Q}\|_F^2+\tau\sum_{j>\theta}\lambda_j(\mathbf{W})\right\},$$

is given by $\mathcal{D}_\tau^\theta=\mathbf{U}\mathbf{\Sigma}_\tau^\theta\mathbf{V}^T$, where $\mathbf{\Sigma}_\tau^\theta$ is diagonal with

$$(\mathbf{\Sigma}_\tau^\theta)_{jj}=\begin{cases}\mathbf{\Sigma}_{jj}, & i\le\theta,\\ \max(0,\Sigma_{jj}-\tau), & i>\theta.\end{cases}$$

4.4 Sub-Gradient

According to Eq. (3) and hinge loss is non-smooth, we need to solve the sub-gradient of $g(\mathbf{W})$. It is similar to solve the SVM optimization problem in the primal, which has been solved by Pegasos algorithm [Shalev-Shwartz *et al.*, 2011]. We also consider sub-gradient descend method in PS3VT. The sub-gradient of hinge loss in Eq. (2) is

$$\nabla\omega(\mathbf{W},\mathbf{x}_i)=\begin{cases}\mathbf{0}, & [\mathbf{W}^T\mathbf{x}_i]_{y_i}-\max\limits_{y'\neq y_i}[\mathbf{W}^T\mathbf{x}_i]_{y'}\ge 1,\\ [0,\cdots,\underbrace{-\mathbf{x}_i}_{y_i},\cdots,\underbrace{\mathbf{x}_i}_{y'},\cdots,0]_{d\times K}, & \text{else.}\end{cases}$$

Bacause gradient descend (GD) and stochastic gradient descend (SGD) are suitable for different situations, we explore them individually. In each iteration, for GD the gradient updates on the entire dataset

$$\nabla g(\mathbf{W})=\frac{1}{n}\sum_{i=1}^{n}\nabla\omega(\mathbf{W},\mathbf{x}_i)+2\tau_A\mathbf{W}+2\tau_I\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}.$$

Algorithm 1 Proximal Stochastic Sub-gradient Singular Value Thresholding (PS3VT)

Input: $\mathbf{X},\mathbf{y},\mathbf{W}^1,T,\theta,\mu,\tau_A,\tau_I,\tau_S$
Output: $\mathbf{W}$

Compute Laplacian matrix $\mathbf{L}$.
for $t=1,2,\cdots,T$ **do**
Choose sample $\mathbf{x}_{i_t}\in\mathcal{D}_l$ uniformly at random.
Compute sub-gradient $\nabla g(\mathbf{W}^t)$ sample $\mathbf{x}_{t_i}$,

$$\nabla g(\mathbf{W}^t,\mathbf{x}_{i_t})=\nabla\omega(\mathbf{W}^t,\mathbf{x}_{t_i})+2\tau_A\mathbf{W}^t+2\tau_I\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^t$$

Compute SVD decomposition

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T=\mathbf{W}^t-\frac{1}{\mu}\nabla g(\mathbf{W}^t,\mathbf{x}_{i_t})$$

Update $\mathbf{W}^{t+1}$ using Proposition 1

$$\mathbf{W}^{t+1}=\mathbf{U}\mathbf{\Sigma}_{\frac{\tau_S}{\mu}}^\theta\mathbf{V}^T$$

Normalize $\mathbf{W}^{t+1}$ by

$$\mathbf{W}^{t+1}=\min\left\{1,1/\|\mathbf{W}^{t+1}\|\right\}\mathbf{W}^{t+1}$$

end for

For SGD, the gradient updates on a random sample $\mathbf{x}'$

$$\nabla g(\mathbf{W},\mathbf{x}')=\nabla\omega(\mathbf{W},\mathbf{x}')+2\tau_A\mathbf{W}+2\tau_I\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}.$$

Applying Proposition 1, the update of $\mathbf{W}^t$ in equation (3) becomes $\mathbf{W}^{t+1}=\mathcal{D}_\tau^\theta(\mathbf{Q})$ where

$$\mathbf{Q}=\mathbf{W}^t-\frac{1}{\mu}\nabla g(\mathbf{W}^t)$$

and $\tau=\frac{\tau_S}{\mu}$. The updates combine gradient descent and SVT.

4.5 Time Complexity

The time complexity of each iteration consists of two parts: sub-gradient and SVT in each iteration:

Sub-Gradient
Laplacian regularization related term $\mathbf{X}\mathbf{L}\mathbf{X}^T$ is computed before iterations, and the computing $\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}$ needs $\mathcal{O}(d^2K)$. Time complexity is $\mathcal{O}(d^2K)$ for SGD and $\mathcal{O}(ndK+d^2K)$ for GD in each iteration.

Singular Values Thresholding (SVT)
In this part, SVD decomposition determines the time complexity, which is $\mathcal{O}(\min(d^2K,dK^2))$ for both SGD and GD.

In each iteration, update of matrix $\mathbf{W}$ follows above two steps, of which time complexity is determined by updating sub-gradient for SGD it is $\mathcal{O}(d^2K)$ and for GD it is $\mathcal{O}(ndK+d^2K)$. It is apparent that for the time complexity of GD in each iteration is associated with labeled sample size n, which is unfeasible when dealing with large datasets. While SGD update gradients on a random sample in each iteration with early stopping [Camoriano *et al.*, 2016], thus it is more suitable for large scale dataset. The total time complexity of SGD is $\mathcal{O}(d^2Kt)$, where t is the number of iterations.

Parameters	Optimization objectives
$\tau_I = 0, \tau_S = 0$	Linear-MC [Koltchinskii *et al.*, 2001]
$\tau_I = 0, \tau_S > 0$	LRC-MC [Li *et al.*, 2018]
$\tau_I > 0, \tau_S = 0$	SS-MC [Li *et al.*, 2015]
$\tau_I > 0, \tau_S > 0$	PS3VT

Table 2: Connections with other algorithms

4.6 Connections with Other Algorithms

PS3VT is a generalized multi-class learning framework, of which different parameters settings specialize PS3VT to other multi-class classification algorithms. Let $\tau_A > 0$ and hinge loss in terms of multi-class maximum margin. Connections with other algorithms are reported in Table 2.

(1) When both of τ_I and τ_I are zeros, the form is max-margin multi-class problem studied in [Koltchinskii *et al.*, 2001], which constitutes a strong baseline.

(2) When $\tau_I = 0$ and $\tau_S > 0$, the local Rademacher complexity was applied on multi-class area, which was firstly introduced into kernel-based multi-class learning [Li *et al.*, 2018], while ours in linear space.

(3) When $\tau_I > 0$ and $\tau_S = 0$, the problem becomes semi-supervised multi-class learning, studied in [Li *et al.*, 2015].

5 Experiments

In this section, we study the empirical behavior of our proposed algorithm PS3VT with several experiments in terms of test error, the convergence and relation between error rate and rate of labeled data. For each dataset, we compute the adjacency matrix $\mathbf{S}$ by 10-NN graph with similarity $S_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2}$, where $\sigma = \sum_{i=1}^{n+u} \|\mathbf{x}_i - \mathbf{x}_j\|/(n+u)$. Laplacian graph is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix with $D_{ii} = \sum_{j=1}^{n+u} S_{ij}$. For fair comparison, before a method runs on any dataset, we employ 5-folds cross validation to obtain the optimal parameter set by grid search over candidate sets complexity parameter $\tau_A \in \{10^{-15}, 10^{-14}, \cdots, 10^{-6}\}$, unlabeled samples parameter $\tau_I \in \{0, 10^{-15}, 10^{-14}, \cdots, 10^{-6}\}$, local Rademacher complexity parameter $\tau_S \in \{0, 10^{-10}, 10^{-9}, \cdots, 10^{-1}\}$, step size $\frac{1}{\mu} \in \{10^1, 10^2, \cdots, 10^5\}$ and tail parameter $\theta \in \{0.5, 0.6, \cdots, 0.9\} \times \min(|K|, |d|)$. Experiment 2 and Experiment 3 run on a randomly selected dataset to compare the convergences and influence of unlabeled samples.

Note that, in the algorithm, we use cross-validation to tune those regularization parameters, which significantly influences the empirical performance of methods but also cause a large computational burden. Beyond cross-validation, we consider more efficient model selection tools [Li *et al.*, 2017a; Liu *et al.*, 2017; Liu *et al.*, 2019] to tune parameters in future.

5.1 Comparison of Test Error

We run PS3VT and the compared methods on 15 multi-class datasets and report the results in Table 3. Labeled and unlabeled samples are given by stratified random sampling from train data that 30% as labeled samples and the rest as unlabeled ones. To obtain stable results, we run methods on each dataset 30 times with randomly partition such that 70% data for training and 30% data for testing. Further, those multiple test errors allow the estimation of the statistical significance of difference among methods. The statistical significance in Table 3 refers to 95% level of significance under t-test.

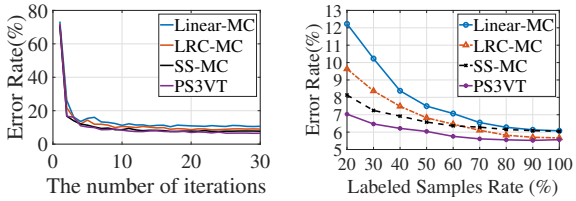

Figure 1: Error rate on *pendigits*. Left: Comparison of convergence for different approaches. Right: Influence of the labeled sample rate.

The results in Table 3 show: (1) Our method outperforms the others almost on all datasets except *iris* and *satimage*. (2) The classical linear margin-based multi-class classification in [Koltchinskii *et al.*, 2001] was defeated by other methods on all datasets. (3) Only combining the local Rademacher complexity or unlabeled samples can still obtain better empirical performance than the primal Linear-MC.

5.2 Comparison of Convergence

We explore the convergence of all methods on *pendigits*, under the same parameters setting and data partition in the above Experiment. The left of Figure 1 reports average test errors various on iterations on *pendigits*, showing our proposed PS3VT reaches a lower error rate than others, LRC-MC and SS-MC give better convergence than Linear-MC as well. The convergence speeds of all methods stay at the same level.

5.3 Influence of the Rate of Labeled Samples

The right of Figure 1 demonstrates the influence of the number of labeled samples. As the growth of the number of labeled samples, test errors of all methods decrease. PS3VT outperforms the others and get similar results with LRC-MC when all training data are labeled. Linear-MC is worse than the others and gets similar results with SS-MC when all training data are labeled. SS-MC gives better test errors than LRC-MC when the rate of labeled data is small, but the performance of LRC-MC has an advantage over which of SS-MC when the rate of labeled data is larger than 60%.

6 Conclusion

Motivated by the idea of taking advantage of unlabeled samples and local Rademacher complexity, we study the generalization behavior of multi-class classification. We combine linear multi-class estimator with the local Rademacher complexity and unlabeled samples, achieving a shaper multi-class generalization error bound with faster convergence rate or smaller constants. Driven by theoretical analysis, we propose a nonconvex optimization problem and design an efficient stochastic gradient descent algorithm to solve it. Further, our theoretical analysis and algorithm can be improved by Random Feature and extended to multi-label learning.

Approaches	Linear-MC	LRC-MC	SS-MC	PS3VT
iris	27.12±5.36	24.57±6.13	**23.53±5.04**	23.71±5.22
wine	8.77±3.22	8.33±5.22	8.20±4.12	**7.63±3.88**
glass	48.68±5.32	47.46±5.40	46.68±4.83	**46.28±5.18**
svmguide2	23.31±3.86	22.42±3.68	22.33±3.99	**21.37±3.46**
vowel	47.40±3.73	47.05±2.89	46.66±3.36	**45.74±3.15**
vehicle	33.78±2.17	29.74±2.41	29.67±2.73	**28.53±2.48**
dna	8.83±0.94	8.69±0.86	8.56±0.78	**8.56±0.78**
segment	26.69±2.20	26.84±2.37	26.28±2.30	**26.09±2.20**
satimage	15.94±0.83	**15.88±0.83**	15.92±0.87	15.89±0.82
pendigits	10.22±0.89	8.37±0.53	7.24±0.44	**6.46±0.37**
usps	7.19±0.42	7.09±0.41	7.10±0.41	**7.06±0.45**
shuttle	23.25±0.32	21.61±0.31	21.55±0.28	**21.48±0.28**
letter	28.31±0.54	26.98±0.49	26.92±0.52	**26.91±0.48**
poker	52.34±0.50	50.30±0.38	50.22±0.40	**50.11±0.45**
Sensorless	54.71±1.26	54.04±1.46	53.15±1.43	**52.50±1.22**

Table 3: Comparison of test err (%) among our proposed PS3VT and other methods listed in Table 2. For each dataset, we bold the optimal test error and underline results in other methods which show no significant difference from the optimal one.

7 Proof

Theorem 2. *For any $\ell \in \mathcal{L}_r : \mathcal{X} \times \mathcal{Y} \to [0,1]$, consider a sub-root function $\psi(r)$ with fixed point r^* and such that $\forall r > r^*$, $KL\mathcal{R}(\mathcal{H}_r) \leq \psi(r)$, then $\forall \ell \in \mathcal{L}_r$ and $\forall k > 1$, with probability at least $1-\delta$*

$$L(\ell) \leq \max\left\{\frac{k}{k-1}\widehat{L}(\ell), \widehat{L}(\ell) + c_4 r^* + \frac{c_1}{n}\right\},$$

where $c_1 = (3+4k)\log(1/\delta), c_4 = 32k$.

Proof. Since ℓ is L-Lipschitz continuous, exploiting the contraction inequality [Koltchinskii, 2011] and applying Lemma 1 in [Maximov and Reshetova, 2016], we have

$$\begin{aligned}
\widehat{\mathcal{R}}(\mathcal{L}_r) &= \frac{1}{n+u}\mathbb{E}_\sigma \sup_{\ell\in\mathcal{L}_r} \sum_{i=1}^{n+u} \sigma_i \ell(\rho_h(\mathbf{x}_i, y_i^\circ)) \\
&\leq \frac{1}{n+u}\mathbb{E}_\sigma \sup_{h\in\mathcal{H}_r} \sum_{i=1}^{n+u} \sigma_i L\rho_h(\mathbf{x}_i, y_i^\circ) \\
&\leq L\sum_{j=1}^{K}\frac{1}{n+u}\mathbb{E}_\sigma \sup_{h\in\mathcal{H}_r} \sum_{i=1}^{n+u} \sigma_i h(\mathbf{x}_i, j) = L\sum_{j=1}^{K}\widehat{\mathcal{R}}(\mathcal{H}_r).
\end{aligned}$$

So we have $\mathcal{R}(\mathcal{L}_r) \leq KL\mathcal{R}(\mathcal{H}_r)$. According Lemma A.6 in [Oneto *et al.*, 2015], we have $\mathcal{R}_n(\mathcal{L}_r) = \mathcal{R}(\mathcal{L}_r)$.

With $\mathcal{R}_n(\mathcal{L}_r) \leq KL\mathcal{R}(\mathcal{H}_r) \leq \psi(r)$ and Theorem 3.3 in [Bartlett *et al.*, 2005] with $\alpha = 1$, we complete the proof. □

Theorem 3. *Let $\mathbf{W} = \mathbf{U\Sigma V}$ be SVD decomposition of $\mathbf{W}$,$\mathbf{U}$ and $\mathbf{V}$ are unitary matrices with size of $d \times d$ and $K \times K$ respectively, and $\mathbf{\Sigma}$ is a $d \times K$ matrix with singular values $\{\lambda_j\}$ on the diagonal in descending order. Assume $\|\mathbf{W}\| \leq 1$, such that the local Rademacher complexity $\mathcal{R}(\mathcal{H}_r)$ over all examples is upper bounded by*

$$\mathcal{R}(\mathcal{H}_r) \leq \frac{1}{KL}\sqrt{\frac{r\theta}{n+u}} + \frac{\sum_{j>\theta}\lambda_j}{\sqrt{n+u}}.$$

Proof. Let $\mathbb{E}_\mu\|\mathbf{x}_i^T\mathbf{x}_i\| \leq K\sqrt{r}/L$ by normalization. According to Definition 1 that $L(\ell^2) \leq r$, let

$$\begin{aligned}
L(\ell^2) &= \mathbb{E}_\mu\left[\ell(\rho_h(\mathbf{x},y))\right]^2 \leq L^2\mathbb{E}_\mu[h(\mathbf{x},y) - \max_{y'\neq y} h(\mathbf{x},y'))]^2 \\
&\leq L^2\mathbb{E}_\mu[h(\mathbf{x},y)]^2 \leq L^2\mathbb{E}_\mu[\mathbf{W}_{\cdot y}^T\mathbf{x}]^2 \\
&\leq L^2\mathbb{E}_\mu\|\mathbf{W}^T\mathbf{x}\|^2 \leq KL\sqrt{r}\mathbb{E}_\mu\|\mathbf{W}^T\mathbf{W}\| \leq r
\end{aligned}$$

So subset space of hypothesis $\mathcal{H}_r$ is satisfied $L(\ell^2) \leq r$ when $\mathbb{E}_\mu\|\mathbf{W}^T\mathbf{W}\| \leq \frac{\sqrt{r}}{KL}$. By proof in Theorem 5 of [Xu *et al.*, 2016] with $\mathbb{E}_\mu\|\mathbf{W}^T\mathbf{W}\| \leq \frac{\sqrt{r}}{KL}$, we complete the proof. □

Proof of Theorem 1. Applying Theorem 3 we have

$$KL\mathcal{R}(\mathcal{H}_r) \leq \sqrt{\frac{r\theta}{n+u}} + \frac{KL\sum_{j>\theta}\lambda_j}{\sqrt{n+u}}.$$

That is $KL\mathcal{R}(\mathcal{H}_r) \leq A\sqrt{r} + B$ when we set

$$A = \sqrt{\frac{\theta}{n+u}}, \quad B = \frac{KL\sum_{j>\theta}\lambda_j}{\sqrt{n+u}}$$

Properties of sub-root function also show that $\psi(r) \leq r$ where $r > r^*$, thus there is $KL\mathcal{R}(\mathcal{H}_r) \leq r$.
Based on $KL\mathcal{R}(\mathcal{H}_r) \leq A\sqrt{r} + B$ and $KL\mathcal{R}(\mathcal{H}_r) \leq r$, we consider choose r^* by solution of $A\sqrt{r} + B = r$, such that

$$r^* \leq \frac{\theta}{n+u} + \frac{2KL\sum_{j>\theta}\lambda_j}{\sqrt{n+u}}$$

Applying the above into Theorem 2, we finish the proof. □

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No.61703396, No.61673293), the Youth Innovation Promotion Association CAS, the National Key Research and Development Program of China (No.2018YFC0823104, No.2016YFB1000604), the Science and Technology Project of Beijing (No.Z181100002718004) and the Excellent Talent Introduction of Institute of Information Engineering of CAS (Y7Z0111107).

References

[Allwein *et al.*, 2000] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[Balcan and Blum, 2010] Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):19, 2010.

[Bartlett *et al.*, 2005] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[Camoriano *et al.*, 2016] Raffaello Camoriano, Tomás Angles, Alessandro Rudi, and Lorenzo Rosasco. Nytro: When subsampling meets early stopping. In *Artificial Intelligence and Statistics*, pages 1403–1411, 2016.

[Cortes *et al.*, 2013] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 46–54, 2013.

[Daniely *et al.*, 2015] Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *Journal of Machine Learning Research*, 16(1):2377–2404, 2015.

[Koltchinskii and Panchenko, 2002] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30:1–50, 2002.

[Koltchinskii *et al.*, 2001] Vladimir Koltchinskii, Dmitriy Panchenko, and Fernando Lozano. Some new bounds on the generalization error of combined classifiers. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 245–251, 2001.

[Koltchinskii, 2011] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

[Lei *et al.*, 2015] Yunwen Lei, Urun Dogan, Alexander Binder, and Marius Kloft. Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2035–2043, 2015.

[Li and Guo, 2015] Xin Li and Yuhong Guo. Max-margin zero-shot learning for multi-class classification. In *Artificial Intelligence and Statistics*, pages 626–634, 2015.

[Li *et al.*, 2015] Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4211–4219, 2015.

[Li *et al.*, 2017a] Jian Li, Yong Liu, Hailun Lin, Yinliang Yue, and Weiping Wang. Efficient kernel selection via spectral analysis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2124–2130, 2017.

[Li *et al.*, 2017b] Xiao Li, Min Fang, and Jinqiao Wu. Zero-shot classification by transferring knowledge and preserving data structure. *Neurocomputing*, 238:76–83, 2017.

[Li *et al.*, 2018] Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1591–1600, 2018.

[Liu and Chang, 2009] Wei Liu and Shih-Fu Chang. Robust multi-class transductive learning with graphs. In *Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 381–388. IEEE, 2009.

[Liu *et al.*, 2017] Yong Liu, Shizhong Liao, Hailun Lin, Yinliang Yue, and Weiping Wang. Infinite kernel learning: generalization bounds and algorithms. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 2280–2286, 2017.

[Liu *et al.*, 2019] Yong Liu, Shizhong Liao, Shali Jiang, Lizhong Ding, Hailun Lin, and Weiping Wang. Fast cross-validation for kernel-based algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[Lu *et al.*, 2015] Canyi Lu, Changbo Zhu, Chunyan Xu, Shuicheng Yan, and Zhouchen Lin. Generalized singular value thresholding. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1805–1811, 2015.

[Maximov and Reshetova, 2016] Yu Maximov and Daria Reshetova. Tight risk bounds for multi-class margin classifiers. *Pattern Recognition and Image Analysis*, 26(4):673–680, 2016.

[Maximov *et al.*, 2018] Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61:761–786, 2018.

[Oneto *et al.*, 2011] Luca Oneto, Davide Anguita, Alessandro Ghio, and Sandro Ridella. The impact of unlabeled patterns in rademacher complexity theory for kernel classifiers. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 585–593, 2011.

[Oneto *et al.*, 2015] Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125, 2015.

[Shalev-Shwartz *et al.*, 2011] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[Xu *et al.*, 2016] Chang Xu, Tongliang Liu, Dacheng Tao, and Chao Xu. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016.