Multi-Class Learning using Unlabeled Samples:
Theory and Algorithm

Jian Li1,2, Yong Liu1*, Rong Yin1,2 and Weiping Wang1
1Institute of Information Engineering, Chinese Academy of Sciences
2School of Cyber Security, University of Chinese Academy of Sciences
{lilian, liuyong, yinrong, wangweiping}@iie.ac.cn

Abstract
In this paper, we investigate the generalization performance of multi-class classification, for which we obtain a sharper 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 important 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 $O(K/\sqrt{n})$ 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 $O(K/\sqrt{n+u+1/n})$, where $u$ is the number of unlabeled samples, which is much faster than the rate of common bounds $O(K/\sqrt{n})$. Further, motivated 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 $O(\log^2 K/n)$ 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 $O(\sqrt{K/n} + \sqrt{K/u})$. In this work, we derive a much shaper generalization bound for multi-class classification with fast rate $O(K/\sqrt{n+u} + 1/n)$, 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 \(D_1 = \{(x_1, y_1), \ldots, (x_n, y_n)\}\) are i.i.d. sampled from distribution \(\mu\) over \(\mathcal{X} \times \mathcal{Y}\) and unlabeled ones \(D_u = \{(x_1, \ldots, x_n)\}\) are i.i.d. drawn according to the marginal distribution \(\mu_X\) of \(\mu\) over \(\mathcal{X}\), typically \(n \ll u\). Further, we consider multi-class classification with \(K \geq 2\) categories, where \(x_i \in \mathbb{R}^d\) and \(y_i \in \{1, 2, \ldots, K\}\).

2.1 Maximum Margin Multi-Class Classification

To evaluate the probability of \(x\) belonging to each category, we wish to learn a scoring rule from the hypothesis space \(\mathcal{H}\): 
\[
    h(x) = W^T x,
\]
where \(h \in \mathcal{H}\), \(W \in \mathbb{R}^{d \times K}\) and \(x \in \mathbb{R}^d\), thus \(h\) is a vector-valued function with mapping \(\mathcal{X} \rightarrow \mathbb{R}^K\). The predictor uses the following mapping to predict labels:
\[
    x \rightarrow \arg\max_y h(x, y).
\]
The empirical Rademacher complexity of loss \(\mathcal{L}_r\) on labeled data and all data are
\[
    \mathcal{R}_n(\mathcal{L}_r) = \mathbb{E}_\sigma \sup_{\ell \in \mathcal{L}_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(h(x_i, y_i)),
\]
\[
    \mathcal{R}(\mathcal{L}_r) = \mathbb{E}_\sigma \sup_{\ell \in \mathcal{L}_r} \frac{1}{n + u} \sum_{i=1}^{n+u} \sigma_i \ell(h(x_i, y'_i)),
\]
where \(\sigma_1, \sigma_2, \ldots, \sigma_{n+u}\) are \([-1, 1]\)-valued independent Rademacher random variables with probability \(P(\sigma_i = +1) = P(\sigma_i = -1) = 1/2\). Their deterministic counterparts are \(\mathcal{R}_n(\mathcal{L}_r) = \mathbb{E}_\sigma \mathcal{R}_n(\mathcal{L}_r)\) and \(\mathcal{R}(\mathcal{L}_r) = \mathbb{E}_\sigma \mathcal{R}(\mathcal{L}_r)\).

Definition 2. The empirical local Rademacher complexity of hypothesis space \(\mathcal{H}_r\) on all data is defined as
\[
    \mathcal{R}_H(\mathcal{H}_r) = \mathbb{E}_\sigma \sup_{h \in \mathcal{H}_r} \frac{1}{n + u} \sum_{i=1}^{n+u} \sigma_i h(x_i, y'_i),
\]
where \(\sigma_1, \sigma_2, \ldots, \sigma_{n+u}\) are \([-1, 1]\)-valued independent Rademacher random variables with probability \(P(\sigma_i = +1) = P(\sigma_i = -1) = 1/2\). The expected counterpart on all data is \(\mathcal{R}(\mathcal{H}_r) = \mathbb{E}_\sigma \mathcal{R}(\mathcal{H}_r)\).

Remark 2. For \(i \in \{1, \ldots, n\}\), labels of \(y'_i\) corresponds to labels in \(D_1\). For \(i \in \{n + 1, \ldots, n + u\}\), \(y'_i\) are pseudo-labeled, as a consequence \(\mathcal{R}(\mathcal{H}_r)\) 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 \(\ell \in \mathcal{L}_r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1], \forall k \geq 1, \|W\| \leq 1\) and \(\forall \delta \in (0, 1)\), the following holds with probability at least \(1 - \delta\):
\[
    L(\ell) \leq \max \left\{ \frac{k}{k - 1} \hat{L}(\ell), \right. \left. \frac{c_1}{n} + \frac{c_2}{n + u} + \frac{c_3 K \sum_{j > \theta} \lambda_j(W)}{\sqrt{n + u}} \right\},
\]
where \(c_1 = (3 + 4k) \log(1/\delta), c_2 = 32k \theta\) and \(c_3 = 64kL\), \(\lambda_j(W)\) is the \(j\) largest singular value of matrix \(W\).

Note that the bound partially depend on \(\sum_{j > \theta} \lambda_j(W)\) 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 \(O(1/n)\) or \(O(K \sum_{j > \theta} \lambda_j(W)/\sqrt{n + u})\), much faster than common rate \(O(K/\sqrt{n})\). 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 \(\ell \in \mathcal{L}_r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1], \forall k \geq 1, \|W\| \leq 1\) and \(\forall \delta \in (0, 1)\), with probability at least \(1 - \delta\),
\[
    L(\ell) \leq \max \left\{ \frac{k}{k - 1} \hat{L}(\ell), \right. \left. \frac{c_1 + c_2}{n} + \frac{c_3 K \sum_{j > \theta} \lambda_j(W)}{\sqrt{n}} \right\},
\]
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: $L(\ell) \leq \hat{L}(\ell) + O(\sqrt{V} \log K / \sqrt{n})$, 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 $O(K^2 / \sqrt{n})$. Moreover, the bound was improved to $O(K / \sqrt{n})$ in [Maximov and Reshetova, 2016]. Lei et al. stated a bound reducing the dependency on the class size [Lei et al., 2015], that is $O((\log K) / \sqrt{n})$. The global Rademacher complexity multi-class bounds were extended into semi-supervised case in [Maximov et al., 2018], of which the convergence rate is $O(\sqrt{K / n} + K \sqrt{K / n})$.

Local Rademacher Complexity Bounds

As we all know, the local Rademacher complexity was firstly presented in binary settings and obtained sharper 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. 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 $O(K / \sqrt{n} + \bar{u} + 1/n)$, while convergence rate of others is usually at $O(K / \sqrt{n})$, 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 $O((c_1 + c_2) \log^2 K / n)$ in [Li et al., 2018], while it is $O(c_1 / n)$ 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.

<table>
<thead>
<tr>
<th>Bounds</th>
<th>Common Case</th>
<th>Special Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Allwein et al., 2000]</td>
<td>$O(\sqrt{V} \log K / n)$</td>
<td></td>
</tr>
<tr>
<td>[Cortes et al., 2013]</td>
<td>$O(\sqrt{K / n})$</td>
<td></td>
</tr>
<tr>
<td>[Maximov et al., 2018]†</td>
<td>$O(\sqrt{K / n} + K \sqrt{K / n})$</td>
<td></td>
</tr>
<tr>
<td>[Li et al., 2018]</td>
<td>$O((c_1 + c_2) \log^2 K / n)$</td>
<td></td>
</tr>
<tr>
<td>Theorem 1†</td>
<td>$O(\sqrt{K / n} + \frac{1}{n})$</td>
<td>$O(\frac{c_1}{n})$</td>
</tr>
</tbody>
</table>

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

$$E(h) = \sum_{i,j=1}^{n+u} S_{ij} \| h(x_i) - h(x_j) \|^2_2 = \text{trace}(W^T XLX^TW),$$

where $X \in \mathbb{R}^{d \times (n+u)}$, graph Laplacian $L = D - S$ and $D$ 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 $\sum_{j>\theta} \lambda_j(W)$ 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

$$\arg\min_{h \in \mathcal{H}_c} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) + \tau_A \| W \|^2_F + \tau_d \text{trace}(W^T XLX^TW) + \tau_S \sum_{j>\theta} \lambda_j(W),$$

where $\ell(h(x_i), y_i) = [1 - h(x_i, y_i) - \max_{y \neq y_i} h(x_i, y)]_+$, $\lambda_j(W)$ denotes the $j$-th largest singular value of $W \in \mathbb{R}^{d \times K}$, and $\tau_A, \tau_d$ and $\tau_S$ are regularization parameters. $\tau_d$ 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 $W$ twice in each iteration, firstly updating $W$ according to first-order sub-gradient of terms except for the tail sum of singular values, and then updating $W$ 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

$$\arg\min_{W} \sum_{j>\theta} \lambda_j(W) + g(W)$$

where

$$g(W) = \frac{1}{n} \sum_{i=1}^{n} [1 - (W^T X_i y_i - \max_{y' \neq y_i} [W^T X_i y'])]^+$$

(2)

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

$$W^{t+1} = \arg\min_{W} \sum_{j>\theta} \lambda_j(W) + g(W)$$

$$= \arg\min_{W} \sum_{j>\theta} \lambda_j(W) + g(W') + \langle \nabla g(W'), W - W' \rangle + \frac{\mu}{2} ||W - W'||_F^2$$

(3)

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

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

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

is given by $D^\theta_r(Q) = U\Sigma^\theta V^T$, where $\Sigma^\theta_r$ is diagonal with

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

4.4 Sub-Gradient

According to Eq. (3) and hinge loss is non-smooth, we need to solve the sub-gradient of $g(W)$. 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 PSVT.

The sub-gradient of hinge loss in Eq. (2) is

$$\nabla \omega(W, x_i) = \begin{cases} 
0, & [W^T x_i]_{y_i} - \max_{y' \neq y_i} [W^T x_i]_{y'} \geq 1, \\
[0, \ldots, \max_{y' \neq y_i} [W^T x_i]_{y'}, \ldots, 0]_d, & \text{else}.
\end{cases}$$

Because 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

$$\nabla g(W) = \frac{1}{n} \sum_{i=1}^{n} \nabla \omega(W, x_i) + 2\tau A W + 2\tau_1 XLX^T W.$$
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 $\tau_I$ and $\tau_I$ 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 $S$ by 10-NN graph with similarity $S_{ij} = e^{-||x_i-x_j||^2/\sigma^2}$, where $\sigma = \sum_{i=1}^{n+u}||x_i-x_j||/(n+u)$. Laplacian graph is defined as $L = D - W$, where $D$ 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 $\tau_I \in \{0, 10^{-5}, 10^{-4}, \cdots, 10^{-6}\}$, local Rademacher complexity parameter $\tau_S \in \{0, 10^{-10}, 10^{-9}, \cdots, 10^{-1}\}$, step size $\frac{1}{n} \in \{10^3, 10^2, \cdots, 10^5\}$ 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.

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.
<table>
<thead>
<tr>
<th>Approaches</th>
<th>Linear-MC</th>
<th>LRC-MC</th>
<th>SS-MC</th>
<th>PS3VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>iris</td>
<td>27.12±5.36</td>
<td>24.57±6.13</td>
<td><strong>23.53±5.04</strong></td>
<td>23.71±5.22</td>
</tr>
<tr>
<td>wine</td>
<td>8.77±3.22</td>
<td>8.33±5.22</td>
<td>8.20±4.12</td>
<td><strong>7.63±3.38</strong></td>
</tr>
<tr>
<td>glass</td>
<td>48.68±5.32</td>
<td>47.46±5.40</td>
<td>46.68±4.83</td>
<td><strong>46.28±5.18</strong></td>
</tr>
<tr>
<td>svmguide2</td>
<td>23.31±3.86</td>
<td>22.42±3.68</td>
<td>22.33±3.99</td>
<td><strong>21.37±3.46</strong></td>
</tr>
<tr>
<td>vowell</td>
<td>47.40±3.73</td>
<td>47.05±2.89</td>
<td>46.66±3.36</td>
<td><strong>45.74±3.15</strong></td>
</tr>
<tr>
<td>vehicle</td>
<td>33.78±2.17</td>
<td>29.74±2.41</td>
<td>29.67±2.73</td>
<td><strong>28.53±2.48</strong></td>
</tr>
<tr>
<td>dna</td>
<td>8.83±0.94</td>
<td>8.96±0.86</td>
<td>8.56±0.78</td>
<td><strong>8.56±0.78</strong></td>
</tr>
<tr>
<td>segment</td>
<td>26.69±2.20</td>
<td>26.84±2.37</td>
<td>26.28±2.30</td>
<td><strong>26.09±2.20</strong></td>
</tr>
<tr>
<td>satimage</td>
<td>15.94±0.83</td>
<td><strong>15.88±0.83</strong></td>
<td>15.92±0.87</td>
<td>15.89±0.82</td>
</tr>
<tr>
<td>pendants</td>
<td>10.22±0.89</td>
<td>8.37±0.53</td>
<td>7.24±0.44</td>
<td><strong>6.46±0.37</strong></td>
</tr>
<tr>
<td>usps</td>
<td>7.19±0.42</td>
<td>7.09±0.41</td>
<td>7.10±0.41</td>
<td><strong>7.06±0.45</strong></td>
</tr>
<tr>
<td>shuttle</td>
<td>23.25±0.32</td>
<td>21.61±0.31</td>
<td>21.55±0.28</td>
<td><strong>21.48±0.28</strong></td>
</tr>
<tr>
<td>letter</td>
<td>28.31±0.54</td>
<td>26.98±0.49</td>
<td>26.92±0.52</td>
<td><strong>26.91±0.48</strong></td>
</tr>
<tr>
<td>poker</td>
<td>52.34±0.50</td>
<td>50.30±0.38</td>
<td>50.22±0.40</td>
<td><strong>50.11±0.45</strong></td>
</tr>
<tr>
<td>Sensorless</td>
<td>54.71±1.26</td>
<td>54.04±1.46</td>
<td>53.15±1.43</td>
<td><strong>52.50±1.22</strong></td>
</tr>
</tbody>
</table>

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 $\ell \in L_r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$, consider a sub-root function $\psi(r)$ with fixed point $r^*$ and such that $\forall r > r^*, KLR(H_r) \leq \psi(r)$, then $\forall \ell \in L_r$ and $\forall k > 1$, with probability at least $1 - \delta$

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

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

**Proof.** Since $\ell$ is $L$-Lipschitz continuous, exploiting the contraction inequality [Kolchinskii, 2011] and applying Lemma 1 in [Maximov and Reshetova, 2016], we have

$$\mathcal{R}(L_r) = \frac{1}{n + u} \mathbb{E}_\sigma \sup_{h \in H_r} \sum_{i=1}^{n+u} \sigma_i \ell(\rho_h(x_i, y_i^*))$$

$$\leq \frac{1}{n + u} \mathbb{E}_\sigma \sup_{h \in H_r} \sum_{i=1}^{n+u} \sigma_i \nu \ell(\rho_h(x_i, y_i^*))$$

$$\leq \frac{1}{n + u} \mathbb{E}_\sigma \sup_{h \in H_r} \sum_{i=1}^{n+u} \sigma_i h(x_i, y_i) = L \sum_{j=1}^{K} \bar{R}(H_r).$$

So we have $\mathcal{R}(L_r) \leq KLR(H_r)$. According Lemma A.6 in [Oneto et al., 2015], we have $R_n(L_r) = \mathcal{R}(L_r)$.

With $R_n(L_r) \leq KLR(H_r) \leq \psi(r)$ and Theorem 3.3 in [Bartlett et al., 2005] with $\alpha = 1$, we complete the proof. $\square$

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

$$\mathcal{R}(H_r) \leq \frac{1}{KL} \sqrt{\frac{r \theta}{n + u} + \sum_{j>0} \lambda_j} \sqrt{n + u}.$$
References


