Corrected Stochastic Dual Coordinate Ascent for Top-k SVM

Yoshihiro HIROHASHI†, Nonmember and Tsuyoshi KATO†††,††††(a), Member

SUMMARY  Currently, the top-k error ratio is one of the primary methods to measure the accuracy of multi-category classification. Top-k multi-class SVM was designed to minimize the empirical risk based on the top-k error ratio. Two SDCA-based algorithms exist for learning the top-k SVM, both of which have several desirable properties for achieving optimization. However, both algorithms suffer from a serious disadvantage, that is, they cannot attain the optimal convergence in most cases owing to their theoretical imperfections. As demonstrated through numerical simulations, if the modified SDCA algorithm is employed, optimal convergence is always achieved, in contrast to the failure of the two existing SDCA-based algorithms. Finally, our analytical results are presented to clarify the significance of these existing algorithms.

key words: top-k SVM, empirical risk minimization, convex optimization, stochastic dual coordinate ascent

1. Introduction

One of the most popular performance measurements for multi-category classification is the top-k error ratio [17], which generalizes the classical error ratio. The classical error assumes that the classifier outputs a single predicted category, and counts the testing data in which the predicted category differs from the true one. Meanwhile, the top-k error presumes that the classifier outputs k categories and counts the data where the true category is excluded from the k predicted ones. Because the categories of data in many large benchmarking datasets are often ambiguous, the top-k error is one of the best criterions for evaluating the performance of multi-category classifiers, and has been deployed as a practical performance measure in a wide variety of application domains [8], [14], [15], [23].

Modern machine learning methods are typically based on empirical risk minimization (ERM). In the ERM framework, the average losses plus a regularization term are minimized with respect to the model parameters. The optimal values of the model parameters minimize the average losses equivalent to the error ratio. However, direct minimization of the error ratio is usually avoided, because it reduces the challenge to a non-convex and complicated optimization problem. Crammer and Singer [5] presented the max-hinge loss that is a convex surrogate for the classical error ratio, which may not be a direct solution to what is required for multi-category classifiers benchmarked with the top-k error ratio. Lapin et al [12] devised a breakthrough extension of the max-hinge loss that can be a convex surrogate for the top-k error ratio. Their surrogate loss is called the top-k hinge loss. Its remarkable properties have attracted several researchers, and motivated them to analyze and extend the idea [1], [3], [4], [25], [29].

Lapin et al [12] chose the stochastic dual coordinate ascent (SDCA) algorithm [21] for ERM in which the empirical risk is measured using the top-k hinge loss function. The SDCA is an optimization framework that finds the value of the dual variables that maximize the Fenchel duality [2] of the regularized empirical risk. The dual variables for the top-k support vector machine (SVM) is an \( m \times n \) matrix, where \( m \) and \( n \) are the numbers of categories and examples, respectively. In each iteration of the SDCA, one of \( n \) columns is randomly selected, and the value of the selected column is optimized while those of the other columns remain constant. To find the optimal value of the selected column at each iteration, Lapin et al [12] attempted to develop a projection-based algorithm. Sublinear convergence is guaranteed, if the selected column is optimized successfully at every iteration [21].

Our analysis reveals that Lapin et al [12]'s theory suffers from a serious defect that makes optimal convergence impossible. They found a property for the convex conjugate of a particular class of convex functions that they termed a compatible function, and attempted to derive the convex conjugate of the top-k hinge loss function using the compatible property. However, in this study, we found that the top-k hinge loss does not belong to the class of the compatible functions. From the observed error, it turns out that the feasible region used by Lapin et al is a mere subset of the true feasible region.

Chu et al [4] developed another optimization algorithm that directly utilizes the dual problem discussed by [22]. Chu et al [4] assert that a particular subset of the dual variables can be frozen to zero. However, there is no guarantee that, at least, one of the optimal solutions satisfies the added constraints (See Fig. 1). Hence, neither Chu et al’s algorithm nor Lapin et al’s algorithm can attain the optimal convergence in cases where the wrongly narrower feasible regions do not intersect with the set of the optimal solutions.
In this paper, we present a new SDCA-based algorithm for learning the top-\(k\) SVM. The two existing studies have attempted to update a selected column by the solution optimal to the sub-problem in each iteration. In contrast, our algorithm employs an approximate optimization for the sub-problem (Sect. 5). Nevertheless, the sublinear convergence can be guaranteed by rigorous analysis of the top-\(k\) hinge loss function. Numerical simulations demonstrate that the proposed algorithm successfully converges to the optimum, whereas the two existing SDCA algorithms fail due to the aforementioned theoretical faults (Sect. 6). In Sect. 7, we reveal the significant contributions of the theories developed in the two existing studies to knowledge, before concluding the paper. This paper is a journal version of our conference paper [10]. Proofs for some propositions in this paper are in the conference paper.

**Notation:** We shall use the notation \(\pi(j; s) \in [m]\), which is the index of the \(j\)-th largest component in a vector \(s \in \mathbb{R}^m\). When using this notation, the vector, \(s\), is omitted if there is no danger of confusion. For a vector, \(s \in \mathbb{R}^m\), we can write \(s_{\pi(1)} \geq s_{\pi(2)} \geq \ldots \geq s_{\pi(m)}\). Let us define \(\pi(s) := [\pi(1; s), \ldots, \pi(m; s)]^\top\), and introduce a notation for a vector with permutated components as \(s_{\pi(s)} := [s_{\pi(1)}, \ldots, s_{\pi(m)}]_\pi\).

2. Related Work

The SDCA [21] is one of the possible approaches for minimizing the convex empirical risks. Many of these approaches are based on the gradient of the empirical risk with respect to the model parameters in the primal form. These approaches are divided into two groups: the full gradient-based and the stochastic gradient-based approaches. The simplest full gradient-based method is the steepest descent method, in which the solution is moved to the negative gradient direction at each iteration. However, each iteration of the conventional full gradient method is expensive. Meanwhile, the stochastic gradient-based approach [18] avails a dramatically cheaper per-iteration cost by exploiting the structure of the empirical risk function that is the average of many similar functions. In each iteration of the stochastic gradient-based approach, the steepest descent direction is approximated by the negative gradient of the loss suffered only for the selected example. The feasibility of the stochastic gradient method has been examined extensively as the back-propagation scheme for training neural networks [19]. A weak point of the stochastic gradient method is that, although the approximation by the stochastic gradient is actually identical in expectation to the steepest descent direction, the large variance of the approximation makes the convergence slow [9]. To address this weakness, several methods have been proposed for reducing the approximation error. SVRG [9], SAG [20], and SAGA [6] are popular stochastic gradient methods for reducing approximation errors. The SDCA, however, appears to be a quite different approach, because it performs optimization in the dual form. A strong similarity between the SDCA and the variance-reduction approaches have been found by rewriting the SDCA equivalently in the primal form [6], [9], suggesting that the SDCA may have some variance-reduction effects. A comparison of the SDCA with the variance-reduction approaches reveal that the SDCA possesses several other advantages, as described in the previous section. Thus, we may conclude that the SDCA has a great potential to minimize the empirical risk.

The top-\(k\) hinge loss was proposed as an extension of Crammer and Singer’s max hinge loss [5] that was a convex upper-bound surrogate for 0/1 loss designed for multi-category SVM. Before the emergence of the top-\(k\) hinge loss, some researchers attempted to tackle the category ambiguity problem mentioned in the previous section. Tsochantaridis et al. [27] developed the structured SVM by refining the max hinge loss so that each category-wise loss can be specified manually. McAuley et al [16] adopted Tsochantaridis et all’s concept to cope with the category ambiguity problem in the multi-category classification task; however, their idea is not a direct solution to the problem. Usunier et al. [28] devised another convex upper-bound surrogate scheme to improve the top-\(k\) ranking performance. Usunier et al’s loss is formulated similarly to the top-\(k\) hinge loss; however, the bound of Usunier et al’s loss is looser for the 0/1 loss than that of the top-\(k\) hinge loss. Several works, such as Barrada et al [1] and Tan [26], presented alternatives to the top-\(k\) hinge loss. The truncated top-\(k\) hinge loss [13] and the robust top-\(k\) hinge loss [3] were designed for alleviating the sensitivity to the outliers. However, the convexity of those loss functions were compromised.

3. Lapin et al [12]’s Theories

In this section, we review the theories postulated by Lapin et al in developing an algorithm for learning the top-\(k\) SVM, and subsequently discuss the shortcomings of their theories. Without the loss of generality, a natural number in \([m]\) can be assigned to each category. The linear multi-category classifier has a model parameter \(W := [w_1, \ldots, w_m] \in \mathbb{R}^{d \times m}\) to predict an unknown input \(x \in \mathbb{R}^d\). The classifier computes the prediction score vector \(s := W^\top x\) to take top-\(k\) outputs \(\pi(1; s), \ldots, \pi(k; s) \in [m]\). An ideal loss to obtain a better
top-

k error is the top-

k 0/1 loss, defined as

$$
\Phi_{01}(s; y) := \begin{cases} 
0 & \text{if } y \in [\pi(1; s), \ldots, \pi(k; s)], \\
1 & \text{if } y \notin [\pi(1; s), \ldots, \pi(k; s)] 
\end{cases}
$$

(1)

for a ground-truth $y \in [m]$. The top-

k 0/1 loss suffers from two disadvantages. One is non-convexity, as a result of which learning is intractable. The other is the ambiguity; multiple permutations $\pi(s)$ exist when two equivalent entries are contained in the score vector $s \in \mathbb{R}^m$. The top-

k hinge, defined as

$$
\Phi_{utk}(s; y) := \max \left\{ 0, 1 - \sum_{j=1}^k (1_m - e_{yj} + s - s_j 1_m)_{[n]} \right\}
$$

(2)

is a convex surrogate that simultaneously eliminates both disadvantages.

Supposing we are given a training dataset of $n$ input-output pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times [m]$. Determining the value of $W$ from the dataset is formulated as the following regularized ERM problem with the following objective:

$$
P_{\text{utk}}(W) := \frac{\lambda}{2} ||W||_F^2 + \frac{1}{n} \sum_{i=1}^n \Phi_{\text{utk}}(W^\top x_i; y_i).
$$

(3)

where $\lambda > 0$ is a regularization constant. The Fenchel duality, expressed as

$$
D_{\text{utk}}(A) := \frac{\lambda}{2} ||W(A)||_F^2 - \frac{1}{n} \sum_{i=1}^n \Phi_{\text{utk}}(-a_i; y_i)
$$

(4)

, can be used to find the minimizer of the primal objective $P_{\text{utk}}(W)$. Therein, $a_i$ is the $i$-th column in $A$; $\Phi_{\text{utk}}(\cdot; y_i) : \mathbb{R}^m \to \mathbb{R}$ is the convex conjugate of the top-

k hinge loss function, $\Phi_{\text{utk}}(\cdot; y_i)$, where $\mathbb{R} := \mathbb{R} \cup \{+\infty\}$; $W(\cdot)$ is defined as $W(A) := \frac{1}{n} X A^\top$, and $X := [x_1, \ldots, x_n]$. The values of the hyperparameters, $\lambda$ and $k$, may be determined using the cross-validation method. Assuming the maximizer of the Fenchel duality is $A^* \in \mathbb{R}^{m\times n}$, the primal optimal solution can be recovered by $W(A^*)$.

**Proposition 3.1:** The convex conjugate of the top-

k hinge loss is given as

$$
\Phi_{\text{utk}}^*(v; y) = \begin{cases} 
v_y & \text{if } v \in \text{dom}(\Phi_{\text{utk}}^*(\cdot; y)), \\
+\infty & \text{if } v \notin \text{dom}(\Phi_{\text{utk}}^*(\cdot; y)), 
\end{cases}
$$

(5)

where $\text{dom}(\Phi_{\text{utk}}^*(\cdot; y))$ is the effective domain of $\Phi_{\text{utk}}^*(\cdot; y)$. The effective domain is provided by

$$
\text{dom}(\Phi_{\text{utk}}^*(\cdot; y)) = \left\{ v \in \mathbb{R}^m \mid \langle v, 1 \rangle = 0, \exists b_y \in \mathbb{R}, \quad v + (b_y - v) e_y \in \Delta_{k,m} \right\},
$$

(6)

where $\Delta_{k,m}$ is the top-

k simplex defined as $\Delta_{k,m} := \{ \beta \in \mathbb{R}^m \mid (1, \beta) \leq 1, \beta \leq \frac{1}{k} \mathbf{1}^\top \beta \}$.

The proof is presented in Sect. A.1. Unfortunately, Lapin et al wrongly derived the following set as the effective domain of the convex conjugate:

$$
\text{dom}(\Phi_{\text{utk}}^*(\cdot; y)) = \left\{ v \in \mathbb{R}^m \mid \langle v, 1 \rangle = 0, \quad v - v_y e_y \in \Delta_{k,m} \right\}.
$$

(7)

In Fig. 2, the optimization process of the SDCA on a small toy problem is demonstrated. The small toy problem in the above demonstration contains a set of $n = 4$ examples, each belonging to one of $m = 3$ categories; the parameter of the top-

k hinge is set to $k = 2$. In this problem, the dual variable $A$ is the $3 \times 4$ matrix:

$$
A = \begin{bmatrix}
A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} \\
A_{2,1} & A_{2,2} & A_{2,3} & A_{2,4} \\
A_{3,1} & A_{3,2} & A_{3,3} & A_{3,4}
\end{bmatrix}
$$

(8)

in which each column is three-dimensional. Each of the four panels in Fig. 2 (b) plots where each column is located in the 3D space at each iteration $t$. In this case, the feasible region of $i$-th column, $-\text{dom}(\Phi_{\text{utk}}^*(\cdot; y))$, is a rhombus that is depicted with a thick line in one of the four panels on the right. At each iteration, a chosen column is moved to the optimal position in the rhombus, which increases the dual objective, $D_{\text{utk}}(A)$. In Fig. 2 (a), the primal and the dual objective values, $P_{\text{utk}}(W(A^{(t)}))$ and $D_{\text{utk}}(A^{(t)})$, are plotted against the iteration number $t$ with blue and red curves, respectively. As
can be observed in Fig. 2 (a), both values converge to the same value, demonstrating successful optimization.

The wrong effective domain \( \text{dom}(\Phi_{\text{irk}}(\cdot ; y)) \) is a line segment, not a rhombus. At each iteration of Lapin et al’s SDCA, a chosen column is optimized in the corresponding line segment, as demonstrated in Fig. 2 (d). As observed in Fig. 2 (c), a large disparity remains between the primal and the dual objectives. The algorithm never converges to the optimum, unless every column in the optimal \( A \) lies exactly in the corresponding line segment.

4. Chu et al [4]’s Theories

In this section, Chu et al [4]’s theories are briefly described. Their theory begins with a slightly different formulation of the regularized empirical risk:

\[
p_{\text{atk}}(w) := \frac{1}{2} ||w||^2 + \frac{1}{n} \sum_{i=1}^{n} \phi_{\text{atk}}((H_{y_i} \otimes x_i^\top)w ; y_i)
\]

where \( H_y := I_m - 1e_y^\top \),

\[
\phi_{\text{atk}}(s ; y) := \max \left\{ 0, \frac{1}{k} \sum_{j=1}^{k} \left( 1_m - e_y + s \right) x_{n(j,x)} \right\}
\]

and the operator \( \otimes \) denotes the Kronecker product. It is not difficult to show the equality \( p_{\text{atk}}(\text{vec}(W)) = p_{\text{atk}}(W) \) for all \( W \in \mathbb{R}^{d \times m} \). The Fenchel duality function of \( p_{\text{atk}}(w) \) is represented as follows:

\[
d_{\text{atk}}(A) := -\frac{1}{2\lambda n^2} \sum_{i=1}^{n} \left( H_{y_i} \otimes x_i \right) A_i^\top
\]

\[
-\frac{1}{n} \sum_{i=1}^{n} \phi_{\text{atk}}^*(-\alpha_i ; y_i),
\]

where \( H_y := I - 1e_y^\top \). This formulation is similar to the one presented in [22]. In maximizing \( d_{\text{atk}}(A) \), instead of a feasibility condition, \( A \in \text{dom}(-d_{\text{atk}}) \) must be given to the dual variable \( A \in \mathbb{R}^{mn} \). Nonetheless, Chu et al [4] insist that the \((y_i, i)\)-th entries in \( A \), such as \( \alpha_{y,i} \), for \( i \in [n] \) can be fixed as zero, and their algorithm has been developed on the assumption of this constant value. The constant values can be expressed in a matrix form as \( E_y \otimes A = O \), where \( E_y := [e_{y_1}, \ldots, e_{y_n}] \) and \( \otimes \) is the operator of the Hadamard product. The constraint \( E_y \otimes A = O \) inevitably makes the feasible region narrower than the true one. Fixing these \( n \) entries as zero would not be detrimental, only when the set of the optimal solutions contains a matrix with \( \alpha_{y,i} = 0 \), \( \forall i \in [n] \). However, unfortunately the empirical results in this study suggest that such a case is very rare.

5. A Corrected SDCA

As proved in Sect. 3 and 4, both the two existing SDCA-based methods have been founded on defective theories that prevent their algorithms from converging to the optimum. In this section, a corrected SDCA algorithm for learning the top-\( k \) SVM is presented. Sublinear convergence, an important property of SDCA, is retained not only when a selected column \( \alpha_i \) is optimized perfectly; however, the column is approximately optimized through the following procedure:

\[
\alpha_i^{(t)} := \alpha_i^{(t-1)} + s^{(t)}(u^{(t)} - \alpha_i^{(t-1)})
\]

where the superscript \( (t) \) indicates the value at \( t \)-th iteration, and

\[
s^{(t)} := \text{argmax}_{x \in [0,1]} D_{\text{atk}}(A^{(t-1)} + s^{(t)}(u^{(t)} - \alpha_i^{(t-1)})e_i^\top).
\]

, and \( u^{(t)} \in -\partial \Phi_{\text{atk}}(z^{(t)} ; y_i) \), \( z^{(t)} := A^{(t-1)}k_i/\lambda n \), and \( k_i \) is the \( i \)-th column in a symmetric matrix \( X^\top X \). In cases where the top-\( k \) hinge, the value of \( s^{(t)} \) in (13) is given in a closed form as

\[
s^{(t)} = \text{Clip}_{[0,1]} \left( \frac{\lambda n \left( z^{(t)} - e_y, q_t \right)}{\|x_i\|^2 \|q_t\|^2} \right)
\]

where \( q_t := u^{(t)} - \alpha^{(t-1)} \), based on the rearrangement of the dual objective as shown below:

\[
D_{\text{atk}}(A^{(t-1)} + s q e_i^\top) = -\frac{1}{2\lambda n^2} \|x_i\|^2 \|q_t\|^2 s^2
\]

\[
-\frac{1}{n} \left( z^{(t)} - e_y, q_t \right) s - \frac{A}{2} \|W(A^{(t-1)})\|_F^2.
\]

Convergence Rate:

Shalev-Shwartz and Zhang [22] have theoretically proven that the SDCA converged sublinearly even when each column was updated with the approximated solution above. If denoting the dual objective gap at \( t \)-th iteration using \( e_D^{(t)} := \min W_{\text{atk}}(W) - D_{\text{atk}}(A^{(t)}) \), the expected dual objective gap can be bounded by

\[
\mathbb{E}[e_D^{(t)}] \leq \frac{G_{\text{max}}}{(2n + t - t_0)\lambda}
\]

where \( t_0 := \max \left\{ 0, \left[ n \log \left( 2n e_D^{(0)}/G_{\text{max}} \right) \right] \right\} \), and the quantity \( G_{\text{max}} \) is an upper-bound: \( G_{\text{max}} \geq R_2^2 \|q_t\|_2^2 \) where \( R_2 := \max_{n \in [n]} \|x_i\|_2 \). The following proposition implies that \( \|q_t\|_2^2 \) cannot exceed 16, \((\|q_t\|_2 \leq \|q_t\|_2 = (2 \cdot 2)^2 = 16) \), leading to an upper-bound: \( G_{\text{max}} := 16R_2^2 \).

**Proposition 5.1:** Let \( k \) satisfy \( 1 \leq k < m \). The top-\( k \) hinge loss is 2-Lipschitz continuous with respect to the norm \( \|\cdot\|_{\infty} \), namely, \( \forall y \in [m], \forall s, \forall s' \in \mathbb{R}^n \),

\[
\| \Phi_{\text{atk}}(s ; y) - \Phi_{\text{atk}}(s' ; y) \| \leq 2 \|s - s'\|_{\infty}.
\]

This inequality is tight. See Appendix A.2 for proof.

6. Numerical Examples

6.1 Comparison of SDCA Algorithms

Here, we illustrate the performance of the modified SDCA
algorithm in learning the top-$k$ SVM, compared with three existing SDCA-based algorithms, Chu I, Chu II, and Lapin. The algorithms, Chu I and Chu II, were developed by Chu et al [4]. Chu et al attempted to maximize $d_{unk}(A)$ with respect to a single column in $A$, for instance, $a_t$, while the other columns are fixed, for each SDCA update. In Chu I, Newton’s method is always used for the SDCA update, whereas, in Chu II, a variable-fixing method [11] replaces Newton’s method under some conditions. The algorithm Lapin is presented in the literature [12]. Their implementations are publicly available on the GitHub repositories. In their codes, a wrong loss function that shall be corrected for the comparison performed in our experiments.

We used three publicly available datasets, FMD, Letter, and News20. The feature vectors in the FMD are extracted from the fc7 layer in a deep neural network, VGG16 [24], that is pre-trained using ImageNet. The dimensionality is 4,096. The dataset contains 1,000 examples, each of which is labeled as one of ten categories. Letter has 15,000 feature vectors. The dimensionality is 16, and the number of categories is 26. The feature vectors in News20 were extracted using principal component analysis. The resultant vectors are 1,024 dimensional and each example is assigned to one of 20 categories. In this experiment, $k = 3$ and $\lambda = 1/n$ are used.

Figures 3 and 4 show the duality gaps $P(W(A)) - D(A)$ generated by the four SDCA algorithms on three datasets. The horizontal axes in each panel represent the CPU time and the number of epochs, respectively. We terminated each SDCA algorithm at the 1,000th epoch. The proposed algorithm, denoted by Corrected SDCA hereinafter, converged to the optimum on all three datasets. The CPU times were 35.02, 6.51, and 9.99 seconds, respectively, to ensure the duality gap is lower than $10^{-3}$ on the three datasets. Meanwhile, the three existing SDCA methods, Chu I, Chu II, and Lapin, could reduce the duality gap below $10^{-1}$ very quickly (within 0.66s) in all cases. However, the duality gaps were not decreased further, and eventually remained higher than $10^{-2}$ at the 1,000th epoch. If comparing the values of the primal objectives at the 1,000th epoch, the differences from that of Modified SDCA, say $P(W(A^{(1,000)})) - P(W(A^{(1,000)})_{cor})$ where $A^{(1,000)}_{cor}$ was the solution generated with Modified SDCA at 1,000th epoch, was seriously large. The minimum duality gaps among the three existing SDCA algorithms were 0.051, 0.020, and 0.013, for the three datasets, respectively. This indicates that, although the modified SDCA converged successfully in all the cases, none of the existing SDCA did. In the next section, we shall discuss what prevents the three existing SDCA from converging optimally.

6.2 Effect of Varying $k$ and $\lambda$

We investigated the effect of varying the values of $k$ in the top-$k$ hinge loss and the regularization parameter $\lambda$. When the value of $\lambda$ is constant, the minimal regularized empirical risk is smaller for larger $k$, because a larger $k$ yields a narrower feasible region for the dual problem. Despite the difference among the minimal regularized empirical risks, the Lipschitz coefficients of the top-$k$ hinge loss with differ-
ent $k$ are not changed, leading to the same theoretical bound of the runtime. Then, how does variation in the value of $k$ affect the convergence of our SDCA algorithm practically? Fig. 5 shows results for the experiment on three datasets, in which the duality gaps are plotted against the number of epochs, where the regularization parameter $\lambda$ is fixed as $1/n$. The empirical results suggest that top-$k$ hinge with larger $k$ tended to converge slightly faster, although the differences were not dramatic. We next varied the value of the regularization parameter $\lambda$ while $k$ was constant. Figure 6 shows the convergence behaviors of different $\lambda$ at $k = 3$. As the empirical loss was regularized more strongly, the duality gap decreased rapidly. These results coincide with the theoretical analysis discussed in Sect. 5 that states that the number of required iterations is inversely proportional to the value of $\lambda$. Thus, the theoretical results of the convergence analysis in relation to varying the value of $k$ and different $\lambda$ were confirmed.

7. Discussions

We have empirically and theoretically demonstrated why only the proposed modified SDCA algorithm converges optimally, in contrast with the failure of both existing SDCA-based algorithms [4], [12]. Subsequently, we shall elucidate the significance of the solutions derived from the theories of these algorithms.

Lapin et al [12] introduced a new concept named $y$-compatible in Definition 2 of their paper for $y \in [m]$. For simplicity, we here assume that $y = m$. Lapin et al [12] defined a convex function $\phi(\cdot ; y) : \mathbb{R}^m \to \mathbb{R}$ to be $y$-compatible, if $\forall y \in \mathbb{R}^{m-1}$,

$$\sup_{s \in \mathbb{R}^{m-1}} \left( \langle s^y, s^\perp \rangle - \phi \left( (s^y)^\top, 0 \right) ; y \right)$$

where we have used the notation $x^y \in \mathbb{R}^{m-1}$ to denote the $(m - 1)$-dimensional vector generated by excluding the $y$-th entry from a vector $x \in \mathbb{R}^m$. In Proposition 3 in [12], it is stated that the function,

$$\phi_{uk}(s ; y) := \max \left\{ 0, \frac{1}{k} \sum_{j=1}^{k} (1 - s_{y} + s_{m,j}) \right\}$$

is $y$-compatible, and the convex conjugate of the top-$k$ hinge loss has been derived with dependence on [12]'s Proposition 3. However, in this study, we have found a result that contradicts [12]'s Proposition 3.

**Proposition 7.1:** The function $\phi_{uk}(\cdot ; y) : \mathbb{R}^m \to \mathbb{R}$ defined in (19) is not $y$-compatible for $2 \leq k < m$.

The proof is provided in the supplementary materials of [10]. The inaccurate effective domain of the convex conjugate, described in (7), would be derived, if the function $\phi_{uk}(\cdot ; y) : \mathbb{R}^m \to \mathbb{R}$ was $y$-compatible, although this claim is rebutted in Proposition 7.1. As a matter of fact, the function

$$\Phi_{uk}^*(\nu ; y) = \begin{cases} \nu_y & \text{if } \nu \in \text{dom}(\Phi_{uk}^*(\cdot ; y)), \\ +\infty & \text{if } \nu \notin \text{dom}(\Phi_{uk}^*(\cdot ; y)), \end{cases}$$

is the convex conjugate of the following function that is no longer top-$k$ hinge loss function, and is defined as follows:

$$\Phi_{uk}(s ; y) = \max \left\{ 0, \frac{1}{k} \sum_{j=1}^{k} (1 - s_{y})1_{m-1} + s_{m,j} \right\}$$

We refer to $\Phi_{uk}(\cdot ; y)$ as the pseudo top-$k$ hinge loss. Let us
denote the $P_{ptk}$ the regularized empirical risk with $\Phi_{ptk}(\cdot; y_i)$ replaced to $\Phi_{ptk}(\cdot; y_i)$ for $i = 1, \ldots, n$. It can be observed that,

$$\forall y \in [m], \forall s \in \mathbb{R}^m, \quad \Phi_{ptk}(s; y) \leq \Phi_{topk}(s; y),$$

implying that $\mathbb{V}W \in \mathbb{R}^{d \times m}, P_{ptk}(W) \leq P_{topk}(W)$. Therefore, the duality gap derived from the true top-$k$ hinge loss can remain positive, even when the duality gap from the pseudo top-$k$ hinge loss vanishes.

Next, we will discuss the effects of Chu et al’s SDCA algorithms [4]. As has been demonstrated, Chu et al’s algorithms could not provide an accurate solution in our experiments, which suggests that the unnecessary constraint $E_y \odot A = O$ should not be imposed for optimization. The observed failure to convergence of their algorithm has prompted us to analyze their theories, because of which we arrived at the following proposition.

**Proposition 7.2:** The maximization problem with objective $d_{ptk}(A)$, subject to a constraint, $E_y \odot A = O$, is the dual to the ERM problem for minimizing $P_{ptk}(W)$.

Proof is provided in the supplementary materials of [10]. From the above discussions, this study has unraveled a new fact: the two existing theories in [12] and [4] are not meant for learning the top-$k$ SVM; rather, they are more applicable to the ERM with the pseudo top-$k$ hinge (21).

8. Conclusions

In this paper, an SDCA-based algorithm was presented for solving the dual problem for learning the top-$k$ multiclass SVM. Due to the theoretical incompleteness of the previous studies [4], [12] tackling the same learning problem, this study is the first to provide an algorithm for this learning problem. The experimental results demonstrated that, unlike the existing algorithms failed to converge optimally, the proposed algorithms always converge optimally. Furthermore, our analysis revealed that both existing algorithms [4], [12] were more suitable for solving a different optimization problem that is not related to learning top-$k$ SVM. In future work, it is necessary to examine the pattern recognition performance of correctly learned top-$k$ SVM for application in various domains.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19K04661.

References

[1] L. Berrada, A. Zisserman, and M.P. Kumar, “Smooth loss functions for deep top-$k$ classification,” 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings, 2018.

[2] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, 1999.

[3] X. Chang, Y.-L. Yu, and Y. Yang, “Robust top-$k$ multiclass SVM for visual category recognition,” Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17, ACM Press, pp.75–83, 2017.

[4] D. Chu, R. Lu, J. Li, X. Yu, C. Zhang, and Q. Tao, “Optimizing top-$k$ multiclass SVM via semismooth newton algorithm,” IEEE Transactions on Neural Networks and Learning Systems, vol.29, no.12, pp.6264–6275, Dec. 2018.

[5] K. Crammer and Y. Singer, “On the algorithmic implementation of multiclass kernel-based vector machines,” J. Mach. Learn. Res., vol.2, pp.265–292, March 2002.

[6] A. Defazio, F. Bach, and S. Lacoste-Julien, “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives,” Proc. 27th International Conference on Neural Information Processing Systems, vol.1, pp.1646–1654, Dec. 2014.

[7] X. Glorot, A. Bordef, and Y. Bengio, “Deep sparse rectifier neural networks,” G. Gordon, D. Dunson, and M. Dudik, editors, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, vol.15, pp.315–323, Fort Lauderdale, FL, USA, 11–13 April 2011. PMLR.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp.770–778, June 2016.

[9] R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” Advances in Neural Information Processing Systems 26: Proceedings of a meeting held Dec. 5–8, 2013, Lake Tahoe, Nevada, United States, pp.315–323, 2013.

[10] T. Kato and Y. Hirohashi, “Learning weighted top-$k$ support vector machine,” W.S. Lee and T. Suzuki, editors, Proceedings of The 10th Asian Conference on Machine Learning, vol.101 of Proceedings of Machine Learning Research, pp.387–393, PMLR, 17–19 Nov. 2019.

[11] K.C. Kiwiel, “Variable fixing algorithms for the continuous quadratic knapsack problem,” Journal of Optimization Theory and Applications, vol.136, no.3, pp.445–458, Nov. 2007. doi:10.1007/s10957-007-9317-7.

[12] M. Lapin, M. Hein, and B. Schiele, “Top-$k$ multiclass svm,” Proceedings of the 28th International Conference on Neural Information Processing Systems, vol.1, NIPS’15, pp.325–333, Cambridge, MA, USA, 2015. MIT Press.

[13] M. Lapin, M. Hein, and B. Schiele, “Loss functions for top-$k$ error: Analysis and insights,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp.1468–1477, June 2016.

[14] T. Lee, T. Gong, S. Padhy, A. Rouditchenko, and A. Ndrango, Label-efficient audio classification through multitask learning and self-supervision, 2019, arXiv:1910.12587.

[15] W. Li, S.M. Siniscalchi, N.F. Chen, and C.-H. Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp.6135–6139, March 2016.

[16] J.J. McAuley, A. Ramisa, and T.S. Caetano, “Optimization of robust loss functions for weakly-labeled image taxonomies: An ImageNet case study,” Lecture Notes in Computer Science, vol.6819, pp.355–368. Springer Berlin Heidelberg, 2011.

[17] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid, “Towards good practice in large-scale learning for image classification,” 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp.3482–3489, June 2012.

[18] H. Robbins and S. Monro, “A stochastic approximation method,” Annals of Mathematical Statistics, vol.22, no.3, pp.400–407, 1951.

[19] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, “Learning representations by back-propagating errors,” Nature, vol.323, no.6088, pp.533–536, Oct. 1986.

[20] M. Schmidt, N.L. Roux, and F. Bach, “Minimizing finite sums with the stochastic average gradient,” Mathematical Programming, vol.162, no.1-2, pp.83–112, June 2016. doi:10.1007/s10107-016-1030-6.

[21] S. Shalev-Shwartz and T. Zhang, “Stochastic dual coordinate ascent methods for regularized loss,” J. Mach. Learn. Res., vol.14, no.1,
Theorem 2 in [10] gives the convex conjugate of the hinge loss defined in our another paper [10]. Theorem 2.1 in [10] gives the convex conjugate of the weighted top-$k$ hinge loss defined in our another paper [10].

We here give a proof for the fact that the top-$k$ weighted top-$k$ hinge loss is 2-Lipschitz continuous with respect to some norm. Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be an $M$-Lipschitz continuous function and let $g : \mathbb{R}^m \rightarrow \mathbb{R}$ be an $L$-Lipschitz continuous function with respect to some norm $\|\|$. Then, the composition function $h \circ g : \mathbb{R}^m \rightarrow \mathbb{R}$ is $L \cdot M$-Lipschitz continuous function with respect to the norm $\|\|$. The proof is based on the following two lemmas.

Lemma 1: Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be an $M$-Lipschitz continuous function and let $g : \mathbb{R}^m \rightarrow \mathbb{R}$ be an $L$-Lipschitz continuous function with respect to some norm $\|\|$. Then, the composition function $h \circ g : \mathbb{R}^m \rightarrow \mathbb{R}$ is $L \cdot M$-Lipschitz continuous function with respect to the norm $\|\|$. The proof is based on the following two lemmas.

Lemma 2: Let $\Phi' : \mathbb{R}^m \rightarrow \mathbb{R}$ be defined as $\Phi'(a) := \frac{1}{k} \sum_{j=1}^{k} (a_j - c - d_j \pi_j)$. The function $\Phi'$ is 2-Lipschitz continuous with respect to $\|\|_{\infty}$. We shall show the two lemmas, followed by derivation of the inequality (17).

Proof for Lemma 1: For all $x, x' \in \mathbb{R}^m$,

$$LM\|x - x'\| \geq M \|g(x) - g(x')\| \geq \|h \circ g)(x) - (h \circ g)(x')\|. \quad (A.1)$$

Proof for Lemma 2: Take arbitrary $a, b \in \mathbb{R}^m$. Let $a' := a + c - d_y \pi_y$ and $b' := b + c - d_y \pi_y$. Then, we have $\Phi'(a) - \Phi'(b) \leq \frac{1}{k} \sum_{p=1}^{k} (a'_p - b'_p \pi_p)$
Tsuyoshi Kato  received his B.E., M.E., and PhD degrees from Tohoku University, Sendai, Japan, in 1998, 2000, and 2003, respectively. From 2003 to 2005, he was with the National Institute of Advanced Industrial Science Technology (AIST) as a postdoctoral fellow in the Computational Biology Research Center (CBRC) in Tokyo. From 2005 to 2008, he was an assistant professor at the Graduate School of Frontier Sciences, University of Tokyo. From 2008 to 2010, he was an associate professor at the Center for Informational Biology, Ochanomizu University. He is now an associate professor at the Graduate School of Science and Technology, Gunma University. His current scientific interests include pattern recognition, computer vision, water engineering and bioinformatics. He is a member of IPSJ and IEICEJ.