Supplementary Materials

Proof of Lemma 3.1

$$\begin{aligned} f(y_q, w_p^*) - f(y_q, w_q^*) = & f(y_q, w_p^*) - f(y_p, w_p^*) && (1)\\ & + f(y_p, w_p^*) - f(y_q, w_p^*) && (2)\\ & + f(y_q, w_p^*) - f(y_q, w_q^*) && (3) \end{aligned}$$

By the optimality of w_p^*, we know that Eq 2 is smaller or equal to 0.

For Eq 1,

$$\begin{aligned} f(y_q, w_p^*) - f(y_p, w_p^*) = & \Big(\mathcal{R}(w_p^*) - \mathcal{R}(w_p^*)\Big) + C\sum_{i=1}^{N}\Big(l(y_{q,i}x_i^T w_p^*) - l(y_{p,i}x_i^T w_p^*)\Big) \\ \le & C\sum_{i=1}^{N}\Big|l(y_{q,i}x_i w_p^*) - l(y_{p,i}x_i^T w_p^*)\Big| \\ = & C\sum_{i=1}^{N}\mathbf{1}\{y_{p,i} \ne y_{q,i}\}\Big|l(x_i^T w_p^*) - l(-x_i^T w_p^*)\Big| \\ \le & C\sum_{i=1}^{N}\mathbf{1}\{y_{p,i} \ne y_{q,i}\}\alpha\big(2|x_i w_p^*|\big) && (4) \\ \le & 2C\sum_{i=1}^{N}\mathbf{1}\{y_{p,i} \ne y_{q,i}\}\alpha B && (5) \\ = & 2Cl_H(y_p, y_q)\alpha B \end{aligned}$$

Eq 4 is true we assume that $l(\cdot)$ is α-Lipschitz. Eq 5 is true since $|x_i^T w_p^*| \le \|x_i\|_2\|w_p^*\|_2$, and in our assumptions, we have $\|x_i\|_2 \le 1$ and $\|w_p^*\|_2 \le B$, so we have $|x_i^T w_p^*| \le B$

The same argument can also be applied for Eq 3, to sum these up, we can get the conclusion that

$$f(y_q, w_p^*) - f(y_q, w_q^*) \le 4l_H(y_p, y_q)C\alpha B.$$

$\square$

Proof of Theorem 3.1

Set w_0 as w_0^*.

$$
\begin{aligned}
T_{total} =& \sum_{k=1}^{K} T_k \\
=& O\left(\frac{\sum_{k=1}^{K}\left(f(y_k, w_0^*) - f(y_k, w_k^*)\right)}{\epsilon^p}\right) \\
=& O\left(\frac{\sum_{k=1}^{K} 4l_H(y_k, y_0)C\alpha B}{\epsilon^p}\right) \qquad (6) \\
=& O\left(\frac{\overline{N}K}{\epsilon^p}\right) \\
=& O\left(\frac{N\overline{L}}{\epsilon^p}\right)
\end{aligned}
$$

Eq 6 is true from lemma 1. □

Proof of Theorem 3.2

$$
\begin{aligned}
T_{total} =& \sum_{k=1}^{K} T_k \\
=& O\left(\sum_{k=1}^{K} \log\left(\frac{f(y_k, w_0^*) - f(y_k, w_k^*)}{\epsilon}\right)\right) \\
=& O\left(K \log\left(\sum_{k=1}^{K} \frac{f(y_k, w_0^*) - f(y_k, w_k^*)}{K\epsilon}\right)\right) \qquad (7) \\
=& O\left(K \log\left(\frac{\sum_{k=1}^{K} 4l_H(y_k, y_0)C\alpha B}{K\epsilon}\right)\right) \qquad (8) \\
=& O\left(K \log\left(\frac{\overline{N}K}{K\epsilon}\right)\right) \\
=& O\left(K \log\left(\frac{\overline{N}}{\epsilon}\right)\right) \qquad (9)
\end{aligned}
$$

$\overline{N}$ is the average number of samples per label, it usually does not scale with N, D or K for extreme classification problems. Eq 7 is true by the concavity of logrithm and Eq 8 is true from lemma 1.

With naive zeros initializaiton, we have $T'_{total} = O\left(K \log\left(\frac{N}{\epsilon}\right)\right)$

If we analyze the upper bound of T_{total} and T'_{total} and assume that $\overline{N}$ and ϵ does not scale with N, then we have

$$\frac{K\log\left(\frac{N}{\epsilon}\right)}{K\log\left(\frac{\overline{N}}{\epsilon}\right)}$$
$$=\log(\frac{N-\overline{N}}{\epsilon})$$
$$=\Theta(\log N)$$

So we can improve the upper bound of the total number of iterations by a factor of $\Theta(\log N)$ when using a solver with linear convergence rate. $\square$

Proof for Lemma 3.2

Denote the minimum spanning tree of G as $\mathcal{T}$(a set of edges). And the minimum spanning tree after removing $e_{p,q}$ from G.

If $e_{p,q} \notin \mathcal{T}$, then obviously, $\mathcal{T}$ is still a minimum spanning tree after removing $e_{p,q}$, so the cost of minimum spanning remains the same.

If $e_{p,q} \in \mathcal{T}$, let $\mathcal{T}' = (\mathcal{T} \setminus e_{p,q}) \cup \{e_{p,k}, e_{q,k}\}$. Obviously, $\mathcal{T}'$ is still a spanning tree. Given that $w_{p,k} + w_{q,k} = w_{p,q}$, so we have $c(\mathcal{T}')$ and $c(\mathcal{T})$, which implies that $\mathcal{T}'$ is a minimum spanning tree of the graph after removing $e_{p,q}$.

The cost of minimum spanning tree remains the same in both cases, so we complete the proof. $\square$

Proof for Theorem 3.3

First, we keep edges $e_{k,0}\ \forall k \in [K]$. Then we only need to consider edges $e_{p,q}$ such that $w_{p,q} < w_{p,0} + w_{q,0}$.

$$|E| = K + \sum_{p=1}^{K}\sum_{q=p+1}^{K} \mathbf{1}\{l_H(y_p, y_q) < N_p + N_q\}$$

where N_p denotes the number of positive samples for label p.

Note that $l_H(y_p, y_q) < N_p + N_q$ iff label p and label q does not share any positve samples.

$$\begin{aligned}
|E| =& K + \sum_{p=1}^{K}\sum_{q=p+1}^{K} \mathbf{1}\Big\{\sum_{i=1}^{N}\mathbf{1}\{y_{p,i} = y_{q,i} = 1\} > 0\Big\} \\
\leq& K + \sum_{p=1}^{K}\sum_{q=p+1}^{K}\sum_{i=1}^{N} \mathbf{1}\{y_{p,i} = y_{q,i} = 1\} \\
=& K + \sum_{i=1}^{N}\Big(\sum_{p=1}^{K}\sum_{q=p+1}^{K} \mathbf{1}\{y_{p,i} = y_{q,i} = 1\}\Big) \\
=& K + \sum_{i=1}^{N} |\mathcal{L}_i|^2/2
\end{aligned}$$

$\square$

1 OVA-MST: algorithm

The detailed algorithm is shown in Algorithm 1.

Algorithm 1 OVA-MST

Input : $\{y_k\}_{k\in[K]}$
Construct label 0 and y_0
Initialize a dictionary d_k for each label k.
for $i = 1$ to N **do**
    for label pairs p, q s.t. $p, q \in \mathcal{L}_i$, $p \neq q$ **do**
        $d_p[q] += 1$
        $d_q[p] += 1$
for $p = 1$ to K **do** $\triangleleft$ convert d into weights.
    for label q in d_p **do**
        $d_p[q] = N_p + N_q - 2d_p[q]$
for $k = 1$ to K **do** $\triangleleft$ connect vertex 0 with all other edges
    $d_0[k] = N_k$ $\triangleleft$ N_k is the number of positve samples for label k
Construct an empty undirected graph $G(V, E)$ with $K + 1$ vertices.
for $p = 0$ to K **do**
    for label q in d_p s.t. $q > k$ **do**
        Add edge $e_{p,q}$ with weight $d_p[q]$ to E.
Run Kruskal's algorithm to find the *minimum spanning tree* $\mathcal{T}$ of G.
Output : $\mathcal{T}$

2 Implementation Details

All of our experiments are conducted on a server with 16 Intel Xeon E5-2690 @ 2.90GHz CPUs and 64GB memory.

2.1 Stopping criterion for DiSMEC, OVA-Naive and OVA-Primal++

For a subproblem, denote the number of its positive samples as N_{pos} and the number of negative samples as N_{neg}. LIBLINEAR's [1] default stopping criterion is $\|\nabla f(w)\|_2 \leq 0.01 \times \min\{N_{pos}, N_{neg}\}/N \|\nabla f(0)\|_2$.

In the setting of extreme classification, $\min\{N_{pos}, N_{neg}\}/N = O(1/N)$, which is very strict when N is large. So we modify the stopping criterion a little bit and stop when $\|\nabla f(w)\|_2 \leq \min\{\epsilon_1 \min\{N_{pos}, N_{neg}\}/N, \epsilon_2\}\|\nabla f(0)\|_2$. We set $\epsilon_1 = 1.0$ and $\epsilon_2 = 1e-4$ in our experiments.

References

[1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.