Supplementary Materials

Proof of Lemma 3.1

\[ f(y_q, w_p^*) - f(y_q, w_q^*) = f(y_q, w_p^*) - f(y_p, w_p^*) \]
\[ + f(y_p, w_p^*) - f(y_q, w_p^*) \]
\[ + f(y_q, w_p^*) - f(y_q, w_q^*) \]  

(1)

(2)

(3)

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

For Eq 1,

\[ f(y_q, w_p^*) - f(y_p, w_p^*) = \left( R(w_p^*) - R(w_p^*) \right) + C \sum_{i=1}^{N} \left( l(y_{q,i}x_i^T w_p^*) - l(y_{p,i}x_i^T w_p^*) \right) \]

\[ \leq C \sum_{i=1}^{N} \left| l(y_{q,i}x_i^T w_p^*) - l(y_{p,i}x_i^T w_p^*) \right| \]

\[ = C \sum_{i=1}^{N} \{ y_{p,i} \neq y_{q,i} \} \left| l(x_i^T w_p^*) - l(-x_i^T w_p^*) \right| \]

\[ \leq C \sum_{i=1}^{N} \{ y_{p,i} \neq y_{q,i} \} \alpha (2|x_i w_p^*|) \]

\[ \leq 2C \sum_{i=1}^{N} \{ y_{p,i} \neq y_{q,i} \} \alpha B \]

\[ = 2C l_H(y_p, y_q) \alpha B \]

(4)

(5)

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

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

\[ f(y_q, w_p^*) - f(y_q, w_q^*) \leq 4l_H(y_p, y_q) \alpha B. \]

\[ \square \]
Proof of Theorem 3.1

Set \( w_0 \) as \( w_0^* \).

\[
T_{total} = \sum_{k=1}^{K} T_k \\
= O \left( \sum_{k=1}^{K} \frac{f(y_k, w_0^*) - f(y_k, w_k^*)}{\epsilon^p} \right) \\
= O \left( \sum_{k=1}^{K} \frac{4l_H(y_k, y_0)C\alpha B}{\epsilon^p} \right) \\
= O \left( \frac{NK}{\epsilon^p} \right) \\
= O \left( \frac{N\bar{T}}{\epsilon^p} \right)
\]

Eq 6 is true from lemma 1.

Proof of Theorem 3.2

\[
T_{total} = \sum_{k=1}^{K} T_k \\
= O \left( \sum_{k=1}^{K} \log \left( \frac{f(y_k, w_0^*) - f(y_k, w_k^*)}{\epsilon} \right) \right) \\
= O \left( K \log \left( \sum_{k=1}^{K} \frac{f(y_k, w_0^*) - f(y_k, w_k^*)}{K\epsilon} \right) \right) \\
= O \left( K \log \left( \frac{NK}{K\epsilon} \right) \right) \\
= O \left( K \log \left( \frac{N}{\epsilon} \right) \right)
\]

\( \bar{T} \) 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 logarithm and Eq 8 is true from lemma 1.

With naive zeros initialization, 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 \( \bar{T} \) and \( \epsilon \) does not scale with \( N \), then we have
\[
\frac{K \log \left( \frac{N}{\epsilon} \right)}{K \log \left( \frac{N}{\epsilon} \right)} = \log \left( \frac{N - N}{\epsilon} \right) = \Theta(\log N)
\]

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

If \( e_{p,q} \not\in T \), then obviously, \( T \) is still a minimum spanning tree after removing \( e_{p,q} \), so the cost of minimum spanning remains the same.

If \( e_{p,q} \in T \), let \( T' = (T \setminus e_{p,q}) \cup \{e_{p,k}, e_{q,k}\} \). Obviously, \( T' \) is still a spanning tree. Given that \( w_{p,k} + w_{q,k} = w_{p,q} \), so we have \( c(T') \) and \( c(T) \), which implies that \( T' \) 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 \( w_{p,q} < w_{p,0} + w_{q,0} \).

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

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 positive samples.

\[
|E| = K + \sum_{p=1}^{K} \sum_{q=p+1}^{K} 1\left\{ \sum_{i=1}^{N} 1\{y_{p,i} = y_{q,i} = 1\} > 0 \right\}
\]

\[
\leq K + \sum_{p=1}^{K} \sum_{q=p+1}^{K} \sum_{i=1}^{N} 1\{y_{p,i} = y_{q,i} = 1\}
\]

\[
= K + \sum_{i=1}^{N} \left( \sum_{p=1}^{K} \sum_{q=p+1}^{K} 1\{y_{p,i} = y_{q,i} = 1\} \right)
\]

\[
= K + \sum_{i=1}^{N} |\mathcal{L}_i|^2 / 2
\]
1 OVA-MST: algorithm

The detailed algorithm is shown in Algorithm 1.

Algorithm 1 OVA-MST
1: **Input**: \{y_k\}_{k \in [K]}
2: Construct label 0 and y_0
3: Initialize a dictionary \(d_k\) for each label \(k\).
4: for \(i = 1\) to \(N\) do
5:     for label pairs \(p, q\) s.t. \(p, q \in \mathcal{L}_i, p \neq q\) do
6:         \(d_p[q]++ = 1\)
7:         \(d_q[p]++ = 1\)
8:     end for
9:     for label \(q\) in \(d_p\) do
10:        \(d_p[q] = N_p + N_q - 2d_p[q]\)
11: end for
12: for \(k = 1\) to \(K\) do
13:     \(d_0[k] = N_k\) \(< \) \(N_k\) is the number of positive samples for label \(k\)
14: end for
15: Construct an empty undirected graph \(G(V, E)\) with \(K + 1\) vertices.
16: for \(p = 0\) to \(K\) do
17:     for label \(q\) in \(d_p\) s.t. \(q > k\) do
18:         Add edge \(e_{p,q}\) with weight \(d_p[q]\) to \(E\).
19:     end for
20: end for
21: Run Kruskal’s algorithm to find the minimum spanning tree \(T\) of \(G\).
22: **Output**: \(T\)

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