6. Appendix A: Convergence Proof

The proofs of Theorem 1, 2 are similar to that in (Lacoste-Julien et al., 2013). To be self-contained, we provide proofs in the following.

6.1. Proof for Theorem 1

The dual problem (14) has (generalized) Hessian for i-th block of variable $\boldsymbol{\alpha}^i$ being upper bounded by

$$\nabla^2_{\boldsymbol{\alpha}^i} G(\boldsymbol{\alpha}) \preceq Q_i I.$$

where $Q_i = \|\boldsymbol{x}_i\|^2$. Since the active set includes the most-violating pair (19) that defines the Frank-Wolfe direction $\boldsymbol{\alpha}^t_{FW}$ satisfying (18), the update given by solving the active-set subproblem (21) has

$$\begin{aligned}
&G(\boldsymbol{\alpha}^{t+1}) - G(\boldsymbol{\alpha}^t) \\
&\leq \gamma \langle \nabla_{\boldsymbol{\alpha}^i} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^{it}_{FW} - \boldsymbol{\alpha}^{it} \rangle + \frac{Q_i \gamma^2}{2} \|\boldsymbol{\alpha}^{it}_{FW} - \boldsymbol{\alpha}^{it}\|^2 \\
&\leq \gamma \langle \nabla_{\boldsymbol{\alpha}^i} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^{it}_{FW} - \boldsymbol{\alpha}^{it} \rangle + \frac{Q_i R^2 \gamma^2}{2}
\end{aligned}$$

for any $\gamma \in [0,1]$, where $\|\boldsymbol{\alpha}^t_{FW} - \boldsymbol{\alpha}^{it}\|^2 \leq R^2 = 4C^2$ since both $\boldsymbol{\alpha}^t_{FW}$, $\boldsymbol{\alpha}^{it}$ lie within the domain (16). Taking expectation w.r.t. i (uniformly sampled from $[N]$), we have

$$\begin{aligned}
&E[G(\boldsymbol{\alpha}^{t+1})] - G(\boldsymbol{\alpha}^t) \\
&\leq \frac{\gamma}{N} \langle \nabla_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^t_{FW} - \boldsymbol{\alpha}^t \rangle + \frac{QR^2\gamma^2}{2N}
\end{aligned} \tag{31}$$

where $Q = \sum_{i=1}^N Q_i$. Then denote $\boldsymbol{\alpha}^*$ as an optimal solution, by convexity and the definition of Frank-Wolfe direction we have

$$\begin{aligned}
\langle \nabla_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^t_{FW} - \boldsymbol{\alpha}^t \rangle &\leq \langle \nabla_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^* - \boldsymbol{\alpha}^t \rangle \\
&\leq G^* - G(\boldsymbol{\alpha}^t),
\end{aligned}$$

where $G^* := G(\boldsymbol{\alpha}^*)$. Together with (31), we have

$$\Delta G^{t+1} - \Delta G^t \leq \frac{-\gamma}{N} \Delta G^t + \frac{QR^2\gamma^2}{2N} \tag{32}$$

for any $\gamma \in [0,1]$, where $\Delta G^t := E[G(\boldsymbol{\alpha}^t)] - G^*$. By choosing $\gamma = \frac{2N}{t+2N}$, the recurrence (32) leads to the result

$$\Delta G^t \leq \frac{2(QR^2 + \Delta G^0)}{t/N + 2},$$

which can be verified via induction as in the proof of Lemma C.2 of (Lacoste-Julien et al., 2013).

[4] http://research.microsoft.com/en-us/um/people/manik/code/SLEEC/download.html

6.2. Proof for Theorem 2

The approximation criteria (24) searches active label from one out of ν partitions of $[K]$. Suppose in the t-th iteration, a subset not containing most-violating label (20) was chosen, we have

$$G(\boldsymbol{\alpha}^{t+1}) - G(\boldsymbol{\alpha}^t) \leq 0 \tag{33}$$

and suppose a subset containing most-violating label was chosen, we have

$$\begin{aligned}
&G(\boldsymbol{\alpha}^{t+1}) - G(\boldsymbol{\alpha}^t) \\
&\leq \gamma \langle \nabla_{\boldsymbol{\alpha}^i} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^{it}_{FW} - \boldsymbol{\alpha}^{it} \rangle + \frac{Q_i R^2 \gamma^2}{2} + \gamma \epsilon_d
\end{aligned} \tag{34}$$

where ϵ_d is the error caused by sampling (25). Since (33), (34) happen with probabilities $1-1/\nu$ and $1/\nu$ respectively, we have expected descent amount

$$\begin{aligned}
&E[G(\boldsymbol{\alpha}^{t+1}) - G^*] - (G(\boldsymbol{\alpha}^t) - G^*) \\
&\leq \frac{\gamma}{N\nu} \langle \nabla_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}^t), \boldsymbol{\alpha}^t_{FW} - \boldsymbol{\alpha}^t \rangle + \frac{QR^2\gamma^2}{2N\nu} + \frac{\gamma\epsilon_d}{\nu} \\
&\leq \frac{-\gamma}{N\nu} (G(\boldsymbol{\alpha}^t) - G^*) + \frac{QR^2\gamma^2}{2N\nu} + \frac{\gamma\epsilon_d}{\nu}.
\end{aligned} \tag{35}$$

following the same reasoning of (31) and (32). For

$$\epsilon_d \leq \frac{QR^2\gamma}{2N},$$

we have

$$\begin{aligned}
&E[G(\boldsymbol{\alpha}^{t+1}) - G^*] - (G(\boldsymbol{\alpha}^t) - G^*) \\
&\leq \frac{-\gamma}{N\nu} (G(\boldsymbol{\alpha}^t) - G^*) + \frac{QR^2\gamma^2}{N\nu}.
\end{aligned} \tag{36}$$

Therefore, by choosing $\gamma = \frac{2}{t/(N\nu)+2}$, we have

$$\Delta G^t \leq \frac{4(QR^2 + \Delta G^0)}{t/(N\nu) + 2}$$

for t satisfying

$$0 \leq t \leq \nu QR^2/\epsilon_d.$$

7. Appendix B: Additional Statistics

Table 4. Default parameter setting used in SLEEC's code. One might need to refer to their webpage [6]for explanation of parameters.

num_learners	num_clusters	SVP_neigh
5	5	50
out_Dim	**w_thresh**	**sp_thresh**
75	0.75	0.5
cost	**NNtest**	**normalize**
0.1	20	1

Table 5. Statistics for heldout and test data set

Data Sets	Train Size	Heldout Size	Test Size.
LSHTC-wiki	2355436	5000	5000
EUR-Lex	15643	1738	1933
bibtex	5991	665	739
RCV1-regions	20835	2314	5000
LSHTC	83805	5000	5000
aloi.bin	90000	10000	8000
Dmoz	310562	34506	38340
ImageNet	1125264	10000	126140
sector	7793	865	961

8. Appendix C: Bounds for Approximation (25)

Let σ_{ki}^2 be the variance of $\bar{C}_k(\mathcal{D}_i)$. We have

$$\sigma_{ki}^2 \leq \hat{\sigma}_{ki}^2 = \frac{1}{\tilde{d}_i}\|\boldsymbol{x}_i\|_1\|\boldsymbol{x}_i\|_\infty R_w^2 \leq \frac{d_i}{\tilde{d}_i}\|\boldsymbol{x}_i\|_\infty^2 R_w^2 \quad (37)$$

, where R_w^2 is an upper bound on $\sum_{j:\boldsymbol{x}_{ij}\neq 0}(\boldsymbol{w}_{kj}^t)^2$.

For $\epsilon = O(\|\boldsymbol{x}_i\|_1 R_w)$, Bernstein-Type inequality gives

$$\Pr[|\bar{C}_k(\mathcal{D}_i) - \langle \boldsymbol{w}_k^t, \boldsymbol{x}_i\rangle| > \epsilon] \leq e^{-\frac{\epsilon^2}{2\hat{\sigma}_{ki}^2}} \quad (38)$$

Suppose we want to approximate $\langle \boldsymbol{w}_k^t, \boldsymbol{x}_i\rangle$ within ϵ_d for all $k \in [K]$ with failure probability at most δ. Combining (37), (38) and using union bound, we only need

$$\frac{d_i}{\tilde{d}_i} \lesssim \frac{\epsilon_d^2}{\log(\frac{K}{\delta})\|\boldsymbol{x}_i\|_\infty^2 R_w^2} \quad (39)$$

Also, look at the dual objective function in (14), initially we have $G(\boldsymbol{\alpha}) = G(\mathbf{0}) = 0$. Since our method is dual-descent, we have $G(\boldsymbol{\alpha}^t) \leq 0$, thus

$$\frac{1}{2}\sum_{k=1}^{K}\|\boldsymbol{w}_k^t\|_2^2 \leq -\sum_{i=1}^{N}\boldsymbol{e}_i^T\boldsymbol{\alpha}^i \leq CN \quad (40)$$

where the last inequality follows from (16).