MILEAGE: Multiple Instance LEArning with Global Embedding – Supplemental Materials

Dan Zhang[1] DANZHANG2008@GMAIL.COM
Jingrui He[2] JINGRUI.HE@GMAIL.COM
Luo Si[3] LSI@CS.PURDUE.EDU
Richard D. Lawrence[4] RICKLAWR@US.IBM.COM

[1]Facebook Incorporation, Menlo Park, CA 94025
[2]Computer Science Department, Stevens Institute of Technology, Hoboken, NJ 07030
[3]Computer Science Department, Purdue University, West Lafayette, IN 47907
[4]IBM T.J. Watson Research Center, Yorktown Heights, NY 10562

Proof of Theorem 1: It is clear that $P(\mathbf{w}^{(t)}, \gamma^{(t)}) \leq \gamma^{(t)}\Omega(\mathbf{w}^{(t-1)})$, since $(\mathbf{w} = \mathbf{w}^{(t-1)}, \zeta = 0)$ is a feasible solution for problem (10). So,

$$\gamma^{(t)}\zeta^{(t)} + \frac{1}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|^2 \leq \gamma^{(t)}(\Omega(\mathbf{w}^{(t-1)}) - \Omega(\mathbf{w}^{(t)})). \quad (1)$$

The LHS is lower bounded by $\gamma^{(t)}\mathbf{g}(\mathbf{w}^{(t-1)})^T(\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}) + \frac{1}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|^2$ [1]. Then, we would have $P(\mathbf{w}^{(t)}, \gamma^{(t)}) \geq \gamma^{(t)}g(\mathbf{w}^{(t-1)})^T(\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}) + \frac{1}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|^2 \geq -\frac{\gamma^{(t)2}}{2}\|g(\mathbf{w}^{(t-1)})\|^2$. So, $-\frac{\gamma_0^2}{2}R^2 \leq -\frac{\gamma^{(t)2}}{2}\|g(\mathbf{w}^{(t-1)})\|^2 \leq P(\mathbf{w}^{(t)}, \gamma^{(t)}) \leq \gamma_0 \max_t \Omega(\mathbf{w}^{(t)}) = \gamma_0 D$. It is clear that in problem (5), $R = C\max\{\max_{i,j}\|\mathbf{B}_{ij}\|, \max_i \|\mathbf{B}_i\|\}$. □

Proof of Theorem 2: The algorithm terminates under two conditions. We show under either of the two conditions, the algorithm terminates after finite steps.

1) Condition 1: As shown in step (8) of Table 1, if $\gamma^{(t)}$ is below a specific threshold ϵ_1, the algorithm terminates. It is clear that the method cannot execute the step (7) to step (9) more than $\log\frac{\epsilon_1}{\gamma_0}/\log(\eta)$ times.

2) Condition 2: Suppose the conditions stated in step (4) hold. By rearranging Eq.(1), we can get: $\zeta^{(t)} + \Omega(\mathbf{w}^t) - \Omega(\mathbf{w}^{t-1}) \leq -\frac{1}{2\gamma^{(t)}}\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|^2$. So, $|\zeta^{(t)} + \Omega(\mathbf{w}^t) - \Omega(\mathbf{w}^{t-1})| \geq \frac{1}{2\gamma^{(t)}}\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|^2 \geq \frac{1}{2\gamma_0}\theta^2$. It means that, each execution of step (4) to step (6) will decrease $F(\mathbf{w})$ by at least $\frac{m}{2\gamma_0}\theta^2$. Since $F(\mathbf{w})$ is upper bounded by E and lower bounded by 0. Step (4) to step (6) cannot be executed for at most $\frac{2E\gamma_0}{m\theta^2}$ steps.

By summarizing these two conditions, we can conclude that the total number iterations should not exceed $\log\frac{\epsilon_1}{\gamma_0}/\log(\eta) + \frac{2E\gamma_0}{m\theta^2}$.

The problem (5) is upper bounded by nC, since $(\mathbf{w} = 0, \xi_i = 1), i = 1, \ldots, n$ is a feasible solution. So, the maximum number of required iterations is $\log\frac{\epsilon_1}{\gamma_0}/\log(\eta) + \frac{2nC\gamma_0}{m\theta^2}$. □

Proof of Theorem 3: The Rademacher Complexity of $\mathcal{F}_B$ can be calculated as:

$$\hat{R}_n(\mathcal{F}_B) = E_\sigma[sup_{f\in\mathcal{F}_B}|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{B}_i, \mathbf{B}_{i*})|]$$

$$=E_\sigma[sup_{\|\mathbf{w}\|\leq B}|\frac{2}{n}\sum_{i=1}^{n}\sigma_i(\lambda_i \max_{j\in\mathbf{B}_i}\mathbf{w}^T\mathbf{B}_{ij} + (1-\lambda_i)\mathbf{w}^T\mathbf{B}_i)|]$$

$$\leq\frac{2}{n}E_\sigma[|\sum_{i=1}^{n}\sigma_i\lambda_i \max_{j\in\mathbf{B}_i}\mathbf{w}_1^{*T}\mathbf{B}_{ij}|] + \frac{2}{n}E_\sigma[|\sum_{i=1}^{n}\sigma_i(1-\lambda_i)\mathbf{w}_2^{*T}\mathbf{B}_i|]$$

$$\leq\frac{2B}{n}E_\sigma[\|\sum_{i=1}^{n}\sigma_i\lambda_i\mathbf{B}_{ij_i^*}\|] + \frac{2B}{n}E_\sigma[\|\sum_{i=1}^{n}\sigma_i(1-\lambda_i)\mathbf{B}_i\|]$$

$$\leq\frac{2B}{n}\sqrt{\sum_{i=1}^{n}\lambda_i^2 K(\mathbf{B}_{ij_i^*}, \mathbf{B}_{ij_i^*})} + \frac{2B}{n}\sqrt{\sum_{i=1}^{n}(1-\lambda_i)^2 K(\mathbf{B}_i, \mathbf{B}_i)}$$

$$\leq\frac{2B}{n}(\max_{\varphi_{ij}\geq 0, \varphi_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\lambda_i^2\varphi_{ij}^2 K(\mathbf{B}_{ij}, \mathbf{B}_{ij})} + \sqrt{\sum_{i=1}^{n}(1-\lambda_i)^2 K(\mathbf{B}_i, \mathbf{B}_i)}),$$

where $j_i^* = \arg\max_j \mathbf{w}^T\mathbf{B}_{ij}$, $\mathbf{w}_1^* = \arg\max_{\mathbf{w}}|\sum_{i=1}^{n}\sigma_i\lambda_i\max_{j\in\mathbf{B}_i}\mathbf{w}^T\mathbf{B}_{ij}|$ and $\mathbf{w}_2^* = \arg\max_{\mathbf{w}}|\sum_{i=1}^{n}(1-\sigma_i)\lambda_i\max_{j\in\mathbf{B}_i}\mathbf{w}^T\mathbf{B}_{ij}|$. It can be seen that the Rademacher Complexity is composed of two different parts. The first part is mainly derived from the multiple instance setting, while the second part is from the traditional setting.

Proof of Theorem 5: It is clear that

$$\frac{2B}{n}\max_{\varphi_{ij}\geq 0, \varphi_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\lambda_i^2\varphi_{ij}^2 K(\mathbf{B}_{ij}, \mathbf{B}_{ij})}$$

$$+ \frac{2B}{n}\sqrt{\sum_{i=1}^{n}(1-\lambda_i)^2 K(\mathbf{B}_i, \mathbf{B}_i)}$$

$$\leq \max\{\frac{C_2}{C_1}a, 1 - \frac{C_2}{C_1}(1-a)\}C_1 + (1-a)C_2$$

[1] Here, to avoid confusion, $\mathbf{g}(\mathbf{w}^{(t-1)}) = \mathbf{g}_{t-1}$.

If $C_1 < C_2$, $\max\{\frac{C_2}{C_1}a, 1-\frac{C_2}{C_1}(1-a)\} = \frac{C_2}{C_1}a$, So, $\max\{\frac{C_2}{C_1}a, 1-\frac{C_2}{C_1}(1-a)\}C_1 + (1-a)C_2 = \frac{C_2}{C_1}aC_1 + (1-a)C_2 = C_2$.
If $C_2 \leq C_1$, $\max\{\frac{C_2}{C_1}a, 1-\frac{C_2}{C_1}(1-a)\} = 1-\frac{C_2}{C_1}(1-a)$, So, $\max\{\frac{C_2}{C_1}a, 1-\frac{C_2}{C_1}(1-a)\}C_1 + (1-a)C_2 = (1-\frac{C_2}{C_1}(1-a))C_1 + (1-a)C_2 = C_1$.

By summarizing these two conditions, we can get the conclusion.□