Supplementary Material — Distributed Primal-Dual Optimization for Online Multi-Task Learning

Peng Yang, Ping Li
Cognitive Computing Lab, Baidu Research USA
{yangpeng1985521}@gmail.com

November 15, 2019

Proof of Theorem

We provide the detailed proof of theorems in this work.

Proof of Theorem 1

We start with additional annotations by denoting the functions and variables as $\mathcal{L}_t(W, A) = F_t(W)+\lambda\text{tr}(A^\top W)-\rho[\|A\|_2-1]_+,\ \nabla_W\mathcal{L}_t = \nabla_W\mathcal{L}_t(W_t, A_t),\ \nabla_A\mathcal{L}_t = \nabla_A\mathcal{L}_t(W_t, A_t),\ P_t = \|W_t - W\|_F^2,\ \ Q_t = \|A_t - A\|_F^2$. For clarify, we divide the proof into two individual steps:

Step 1: Due to the convexity of $\mathcal{L}_t(W, A)$ with respect to W, for any $W \in \mathbb{R}^{d\times m}$ we have

$$
\begin{aligned}
&\mathcal{L}_t(W_t, A) - \mathcal{L}_t(W, A)\\
\le&(W_t - W)\nabla_W\mathcal{L}_t = -\frac{1}{\eta_t}(W_{t+1} - W_t)(W_t - W)\\
=&\frac{1}{2\eta_t}\left(\|W_t - W_{t+1}\|_F^2 + \|W - W_t\|_F^2 - \|W - W_{t+1}\|_F^2\right)\\
=&\frac{1}{2\eta_t}\left(P_t - P_{t+1}\right) + \frac{\eta_t}{2}\|\nabla_W\mathcal{L}_t\|_F^2,
\end{aligned}
$$

where the first equality is due to the update rule of primal variable. Similarly, the concavity of $\mathcal{L}_t(W, A)$ with respect to A yields to

$$
\mathcal{L}_t(W, A) - \mathcal{L}_t(W, A_t) \le \frac{1}{2\eta_t}(Q_t - Q_{t+1}) + \frac{\eta_t}{2}\|\nabla_A\mathcal{L}_t\|_F^2.
$$

Since $(\mathbf{v}, \mathbf{u})$ are unit vectors and $\max(\|\mathbf{w}\|_2, \|\mathbf{a}\|_2) \leq D$,

$$\begin{aligned}\|\nabla_A \mathcal{L}_t\|_F^2 &= \sum_{i=1}^{m} \|\lambda \mathbf{w}_t^i - \rho[\mathbf{u}\mathbf{v}^\top]_i\|_2^2 \\ &\leq \sum_{i=1}^{m} \|\lambda \mathbf{w}_t^i - \rho \mathbf{u}\|_2^2 \leq m(\lambda D + \rho)^2\end{aligned}$$

Moreover, given that $\max_t \|\nabla f_t(\mathbf{w}^i)\|_2 \leq \beta$ and $\max_t \|\nabla_u h(u)|_{u=f_t(\mathbf{w}^i)}\|_2 \leq \kappa$, we obtain

$$\|\nabla_W \mathcal{L}_t\|_F^2 = \sum_{i=1}^{m} \|\gamma_t^i \nabla f_t(\mathbf{w}_t^i) + \lambda \mathbf{a}_t^i\|_2^2 \leq m(\kappa\beta + \lambda D)^2.$$

Summing the above inequalities over $t = 1, \ldots, T$ will give

$$\begin{aligned}&\sum_{t=1}^{T} \frac{1}{2\eta_t}(P_t - P_{t+1}) \\ =&\frac{1}{2\eta_1}P_1 - \frac{1}{2\eta_T}P_{T+1} + \sum_{t=1}^{T-1}(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t})P_{t+1} \\ \leq&\frac{1}{2\eta_1}mD^2 + (\frac{1}{2\eta_T} - \frac{1}{2\eta_1})mD^2 = \frac{m\sqrt{T}}{2}D^2.\end{aligned}$$

where the inequality holds since $\max_t P_t \leq m \max_t \|\mathbf{w} - \mathbf{w}_t\|^2 = mD^2$. Similarly, $\sum_{t=1}^{T} \frac{1}{2\eta_t}(Q_t - Q_{t+1}) \leq \frac{m\sqrt{T}}{2}D^2$. In addition, $\sum_{t=1}^{T} \frac{\eta_t}{2} = \sum_{t=1}^{T} \frac{1}{2\sqrt{t}} \leq \sqrt{T}$. Summarizing the inequalities above, we obtain an upper bound

$$\sum_{t=1}^{T} \mathcal{L}_t(W_t, A) - \mathcal{L}_t(W, A_t) \leq \ m\sqrt{T}\left(D^2 + (\kappa\beta + \lambda D)^2 + (\rho + \lambda D)^2\right). \quad (1)$$

Step 2: In the rest of proof, we will show that the objective value $\sum_{t=1}^{T} \Phi_t(W_t)$ converges to $\sum_{t=1}^{T} \Phi_t(W^*)$ where $\Phi_t(W) = F_t(W) + \lambda\|W\|_*$, and W^* is the optimal solution in hindsight. To achieve this goal, we have to prove that

$$\sum_{t=1}^{T} \Phi_t(W_t) - \Phi_t(W^*) \leq \sum_{t=1}^{T} L_t(W_t, A_t^*) - L_t(W^*, A_t),$$

where $A_t^* = \operatorname{argmax}_{\|A\|_2 \leq 1} \operatorname{tr}(A^\top W_t)$. Due to $\|W_t\|_* = tr(A_t^{*\top} W_t)$, it is easy to verify that

$$L_t(W_t, A_t^*) = F_t(W_t) + \lambda\|W_t\|_* = \Phi_t(W_t).$$

Thus, it remains to prove

$$\lambda \operatorname{tr}(A_t^\top W^*) - \rho[\|A_t\|_2 - 1]_+ \leq \lambda\|W^*\|_*, \quad (2)$$

which consists of two conditions:

1) When $\|A_t\|_2 \leq 1$, (2) yields to $\lambda \operatorname{tr}(A_t^\top W^*) \leq \lambda\|W^*\|_*$, which can be verified since

$\max_{\|A\|_2 \leq 1} \mathrm{tr}(A^\top W) \leq \|W\|_*$.
2) When $\|A_t\|_2 > 1$, $[\|A_t\|_2 - 1]_+ > 0$. Since $\mathrm{tr}(A^\top B) \leq \|A\|_2 \|B\|_*$ for any matrices A and B, we have

$$\begin{aligned} &\lambda \mathrm{tr}(A_t^\top W^*) - \lambda \|W^*\|_* \\ \leq& \lambda \|A_t\|_2 \|W^*\|_* - \lambda \|W^*\|_* = \lambda \|W^*\|_* (\|A_t\|_2 - 1) \\ \leq& \rho (\|A_t\|_2 - 1)_+, \end{aligned}$$

which is hold as we assume $\lambda \|W^*\|_* \leq \rho$. Substituting the upper bound (1), we complete the proof of Theorem 1.

Proof of Theorem 2

For any round T, it can be formulated as $T = t\tau + s$, where $t \geq 0$ indicates the number of central update occurred at round T and $s \in [0, \tau]$ is number of local update occurred after the t-th central update. The local update is performed when $s \in [1, \tau)$, while the central update is conducted when $s = 0$ or $s = \tau$.

Similar with Theorem 1, We have the loss function $\mathcal{L}_t(\mathbf{W}, \mathbf{A})$ with respect to the t-th central update:

$$\begin{aligned} &\mathcal{L}_t(\mathbf{W}_t, \mathbf{A}) - \mathcal{L}_t(\mathbf{W}, \mathbf{A}_t) \\ \leq& \frac{1}{2\eta^t} [(P_t - P_{t+1}) + (Q_t - Q_{t+1})] + \frac{\eta^t}{2} [\|\nabla_{\mathbf{W}} \mathcal{L}_t\|_F^2 + \|\nabla_{\mathbf{A}} \mathcal{L}_t\|_F^2], \end{aligned} \tag{3}$$

where η^t denotes the learning rate for central update from the round t to the round $t+1$.

We next study the local updates during τ synchronize intervals. For any worker i, after t-th central update, the s-th local update ($s \in [1, \tau]$) has a following formulation,

$$\mathbf{w}^i_{t(s)} = \mathbf{w}^i_{t(s-1)} - \eta_{t(s)} \left(\gamma^i_{t(s)} \nabla f_{t(s)}(\mathbf{w}^i_{t(s-1)}) + \lambda \mathbf{a}^i_{t(s-1)} \right).$$

Cumulating the local updates over $s = 1, \ldots, \tau$, we obtain that,

$$\mathbf{w}^i_{t+1(0)} = \mathbf{w}^i_{t(0)} - \sum_{s=1}^{\tau} \eta_{t(s)} (\gamma^i_{t(s)} \nabla f_{t(s)}(\mathbf{w}^i_{t(s-1)})) + \lambda \mathbf{a}^i_{t(s-1)}),$$

where we let $\mathbf{w}^i_{t(\tau)} = \mathbf{w}^i_{t+1(0)}$.

Since $\eta_T = 1/\sqrt{\lceil T/\tau \rceil}$, it indicates that $\eta^t = \eta_{t(1)} = \eta_{t(2)} = \ldots = \eta_{t(\tau)}$. Thus, it infers the equation with respect to the central update from t to $t+1$,

$$\begin{aligned} \|\nabla_{\mathbf{W}} \mathcal{L}_t\|_F^2 =& \|\frac{\mathbf{W}_t - \mathbf{W}_{t+1}}{\eta^t}\|_F^2 = \sum_{i=1}^{m} \|\frac{1}{\eta^t} \mathbf{w}^i_{t+1(0)} - \mathbf{w}^i_{t(0)}\|^2 \\ =& \sum_{i=1}^{m} \| \sum_{s=1}^{\tau} \frac{\eta_{t(s)}}{\eta^t} (\gamma^i_{t(s)} \nabla f_{t(s)}(\mathbf{w}^i_{t(s-1)}) + \lambda \mathbf{a}^i_{t(s-1)}) \|^2 \\ \leq& m\tau^2 \| \max_{s: s \in [1,\tau]} \left(\gamma^i_{t(s)} \nabla f_{t(s)}(\mathbf{w}^i_{t(s-1)}) + \lambda \mathbf{a}^i_{t(s-1)} \right) \|^2 \\ =& m\tau^2 (\kappa\beta + \lambda D)^2. \end{aligned}$$

On the other hand, as weight matrix $\mathbf{S}$ is stochastic with its element satisfying $0 \leq [S]_{ij} \leq 1$,

$$\|\mathbf{A}_{t+1}^{(i)}\|_2 = \|\mathbf{A}_{t+1} \times \text{Diag}([\mathbf{S}]_i)\|_2 \leq \|\mathbf{A}_{t+1}\|_2 \|\text{Diag}([\mathbf{S}]_i)\|_2 \leq \|\mathbf{A}_{t+1}\|_2,$$

which deduces that

$$[\|\mathbf{A}_{t+1}^{(i)}\|_2 - 1]_+ \leq [\|\mathbf{A}_{t+1}\|_2 - 1]_+.$$

Similar with $\|\nabla_{\mathbf{W}} \mathcal{L}_t\|_F^2$, we can bound $\|\nabla_{\mathbf{A}} \mathcal{L}_t\|_F^2$ as below,

$$\begin{aligned}
\|\nabla_{\mathbf{A}} \mathcal{L}_t\|_F^2 &= \|\sum_{s=1}^{\tau}(\lambda \mathbf{W}_{t(s)} - \rho\partial[\|\mathbf{A}_t^{(i)}\|_2 - 1]_+)\|_F^2 \\
&\leq \|\sum_{s=1}^{\tau}(\lambda \mathbf{W}_{t(s)} - \rho\partial[\|\mathbf{A}_t\|_2 - 1]_+)\|_F^2 \\
&\leq m\tau^2(\lambda D + \rho)^2
\end{aligned}$$

Cumulate the objective function (3) over $c = 0, \ldots, t-1$, we can bound

$$\begin{aligned}
&\sum_{c=0}^{t-1} \frac{1}{2\eta^c}[(P_c - P_{c+1}) + (Q_c - Q_{c+1})] \\
=&\frac{1}{2\eta^0}(P_0 + Q_0) - \frac{1}{2\eta^{t-1}}(P_t + Q_t) + \sum_{c=0}^{t-2}(\frac{1}{2\eta^{c+1}} - \frac{1}{2\eta^c})(P_{c+1} + Q_{c+1}) \\
\leq&\frac{1}{\eta^0}mD^2 + (\frac{1}{\eta^{t-1}} - \frac{1}{\eta^0})mD^2 = m\sqrt{t}D^2,
\end{aligned}$$

$$\begin{aligned}
&\sum_{c=0}^{t-1} \frac{\eta^c}{2}[\|\nabla_{\mathbf{W}} \mathcal{L}_c\|_F^2 + \|\nabla_{\mathbf{A}} \mathcal{L}_c\|_F^2] \\
\leq&\sum_{c=0}^{t-1} \frac{\eta^c}{2} m\tau^2 \left((\kappa\beta + \lambda D)^2 + (\lambda D + \rho)^2\right) \\
\leq&m\tau^2\sqrt{t}\left((\kappa\beta + \lambda D)^2 + (\lambda D + \rho)^2\right),
\end{aligned}$$

where the last inequality holds due to $\sum_{c=0}^{t-1} \frac{\eta^c}{2} = \sum_{c=0}^{t-1} \frac{1}{2\sqrt{c+1}} \leq \sqrt{t}$.

Since $t = \sqrt{\frac{T-s}{\tau}}$, we can infer the bound with respect to the round T,

$$\sum_{c=0}^{t-1} \mathcal{L}_c(\mathbf{W}_c, \mathbf{A}) - \mathcal{L}_c(\mathbf{W}, \mathbf{A}_c) \leq \sqrt{T}m\tau^{3/2}\left((D/\tau)^2 + (\kappa\beta + \lambda D)^2 + (\lambda D + \rho)^2\right).$$

Following the step 2 in Theorem 1, we can conclude this proof.

References

[1] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent." in *COLT*, 2010, pp. 14–26.

[2] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[3] F. Bach, R. Jenatton, J. Mairal, G. Obozinski *et al.*, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[4] E. Hazan *et al.*, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[6] B. He and X. Yuan, "On the o(1/n) convergence rate of the douglas–rachford alternating direction method," *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.

[7] H. Wang and A. Banerjee, "Online alternating direction method," in *29th International Conference on Machine Learning, ICML 2012*, 2012.

[8] W. Zheng, A. Bellet, and P. Gallinari, "A distributed frank: Wolfe framework for learning low-rank matrices with the trace norm," *Machine Learning*, pp. 1–19, 2017.

[9] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization." in *ICML*, 2013, pp. 427–435.

[10] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.