Supplementary Material for “Robust AUC Optimization under the Supervision of Clean Data”

Chenkang Zhang\(^1\),\(^*\), Haobing Tian\(^1\), Lang Zhang\(^1\), and Pengju Jiao\(^1\)

\(^1\)China Mobile (Suzhou) Software Technology Company Limited, 215163 Suzhou, China
\(^*\)zhangchenkang@cmss.chinamobile.com

Proof of Theorem 1

In this section, we prove the convergence of our RAUCO algorithm.

We first introduce some definitions. Let \(G_\theta\) and \(G_w\) be the stochastic gradients with respect to \(\theta\) and \(w\) generated by Algorithm 1 in \(t\)-iteration. Importantly, for the stochastic gradient with respect to \(w_i, i \in [n]\) is not calculated by Algorithm 1 in \(t\)-iteration, the \(i\)-th element of \(G_w\) is 0. For the vector \(e_i \in \{0, 1\}^n\), if \(i\)-th element of \(G_w\) is 0, \(e_i^t = 0\); otherwise, \(e_i^t = 1\). Also, for the operator \(\otimes\), if \(A \otimes B = C\) and \(A, B, C \in \mathbb{R}^n\), we have \(A_i \ast B_i = C_i, i \in [n]\). Finally, \(\mathcal{P}_S(\cdot)\) is the projection operation to the set \(S\).

Then, we provide some necessary assumptions and the definition of the projected gradient.

Assumption 1 (Lipschitz Smooth). For our objective function \(\mathcal{L}_\lambda(\theta, w)\), the sub-problems with respect to \(\theta\) and \(w\), i.e., \(\mathcal{L}_\lambda(\theta; w)\) and \(\mathcal{L}_\lambda(w; \theta)\), are both Lipschitz smooth with the maximum Lipschitz constant \(L\), i.e., \(\forall \theta, \theta'\), and \(\forall w, w' \in [0, 1]^n\), we have:

\[
\|\mathcal{L}_\lambda(\theta; w) - \mathcal{L}_\lambda(\theta'; w)\| \leq L \|\theta - \theta'\|
\]

(1)

Assumption 2. For \(\forall t \in [T]\), we have

\[
\mathbb{E}[\|G_w - e_i \otimes \nabla_w \mathcal{L}_\lambda(w'; \theta')\|^2] \leq (\sigma_w^t)^2, \quad \mathbb{E}[\|G_\theta - \nabla_\theta \mathcal{L}_\lambda(\theta'; w^{t+1})\|^2] \leq (\sigma_\theta^t)^2,
\]

(2)

where \(\sigma_w^t > 0, \sigma_\theta^t > 0\) are some constants and we define \(\sigma^t = \max\{\sigma_w^t, \sigma_\theta^t\}\).

Definition 1 (Projected Gradient). \(^1\) Let \(S\) be a closed convex set with dimension \(N\), and the projected gradient is defined as:

\[
\mathcal{K}(w, g, \alpha) = \frac{1}{\alpha}(w - \mathcal{P}_S(w - \alpha g))
\]

(3)

where \(w \in S, g \in \mathbb{R}^N\) and \(\alpha \in \mathbb{R}^+\).

Assumptions 1 and 2 are common assumptions in stochastic optimization. Assumption 1 provides the guarantee of the Lipschitz smoothness, and Assumption 2 bounds the difference between the stochastic gradient and the full gradient.

Finally, our theoretical result is as follows.

Theorem 1. When Assumptions 1 and 2 hold, \(\lambda\) reaches its maximum value \(\lambda_{\infty}\) and the stepsizes \(\{\alpha^t\}_{t=1}^\infty\) satisfy

\[
0 < \alpha^{t+1} \leq \alpha^t < \frac{2}{L}, \quad \sum_{t=1}^\infty \alpha^t = +\infty, \quad \sum_{t=1}^\infty \alpha^t(\sigma^t)^2 < +\infty,
\]

(4)

then there exists an index sub-sequence \(\mathcal{I}\) in Algorithm 1 such that

\[
\lim_{t \to \infty} \mathbb{E}\left[\frac{(\theta^{t+1}, w^{t+1}) - (\theta^t, w^t)}{\alpha^t}\right] = 0.
\]

(5)

The above theorem shows that Algorithm 1 approaches a stationary point of our objective function. It indicates that our algorithm can obtain a satisfactory solution theoretically.

Before proving our theoretical results (i.e., Theorem 1), we introduce the following lemmas.

Lemma 1. \(^1\) Let \(S\) be a closed convex set with dimension \(N\), for any \(w \in S, g \in \mathbb{R}^N\) and \(\alpha > 0\), we have

\[
\langle g, \mathcal{K}(w, g, \alpha) \rangle \geq \|\mathcal{K}(w, g, \alpha)\|^2.
\]

(6)
Lemma 2. Let $S$ be a closed convex set with dimension $N$, for any $w \in S$, $\alpha > 0$ and $g_1, g_2 \in \mathbb{R}^N$, we have
\[
\|\mathcal{K}(w, g_1, \alpha) - \mathcal{K}(w, g_2, \alpha)\|_2 \leq \|g_1 - g_2\|_2.
\]

Lemma 3. For two nonnegative scalar sequences $\{a_i\}_{i=1}^\infty$ and $\{b_i\}_{i=1}^\infty$, if $\sum_{i=1}^\infty a_i = +\infty$ and $\sum_{i=1}^\infty a_i b_i < +\infty$, then
\[
\lim_{i \to \infty} \inf b_i = 0.
\]

Finally, the proof of Theorem 1 is as follows.

Proof. Here, we define some concepts of the gradient with respect to $w$:
\[
K_w^i = \mathcal{K}(w', e^i \otimes \nabla_w \mathcal{L}_\theta(w'; \theta^i)), \quad k_w^i = \mathcal{K}(w', G'_w, \alpha^i), \quad \delta_w^i = G'_w - e^i \otimes \nabla_w \mathcal{L}_\theta(w'; \theta^i).
\]

Next are some similar concepts of the gradient with respect to $\theta$:
\[
K^i_\theta = \nabla_\theta \mathcal{L}(\theta^i; w^{i+1}), \quad k^i_\theta = G^i_\theta, \quad \delta^i_\theta = G^i_\theta - \nabla_\theta \mathcal{L}(\theta^i; w^{i+1}).
\]

According to the optimization process, we have
\[
\mathcal{L}_\theta(\theta^{i+1}; w^{i+1}) - \mathcal{L}_\theta(\theta^i, w^i) = \mathcal{L}_\theta(\theta^{i+1}, w^{i+1}) - \mathcal{L}_\theta(\theta^i, w^{i+1}) + \mathcal{L}_\theta(\theta^i, w^{i+1}) - \mathcal{L}_\theta(\theta^i, w^i).
\]

Due to that the sub-problem $\mathcal{L}(w; \theta)$ with respect to $w$ is Lipschitz smooth with the Lipschitz constant $L$, we have
\[
\mathcal{L}_\theta(w^{i+1}; \theta^i) - \mathcal{L}_\theta(w^i; \theta^i) \leq \langle \nabla_w \mathcal{L}_\theta(w^i; \theta^i), w^{i+1} - w^i \rangle + \frac{L}{2} ||w^{i+1} - w^i||^2
\]

Then, by Lemma 1 with $w = w^i, g = G'_w$ and $\alpha = \alpha^i$, we obtain
\[
\mathcal{L}(w^{i+1}; \theta^i) - \mathcal{L}(w^i; \theta^i) \leq \langle \nabla \mathcal{L}_\theta(w^i; \theta^i), w^{i+1} - w^i \rangle + \frac{L}{2} ||w^{i+1} - w^i||^2 + \alpha^i \langle \delta^i_w, k^i_w \rangle
\]

where the last inequality follows from Lemma 2 with $w = w^i, g_1 = G'_w, g_2 = e^i \otimes \nabla_w \mathcal{L}_\theta(w^i; \theta^i)$ and $\alpha = \alpha^i$. Take expectations on both sides and then we get:
\[
\mathbb{E}[\mathcal{L}(w^{i+1}; \theta^i) - \mathcal{L}(w^i; \theta^i)] \leq \mathbb{E}[(\alpha^i + \frac{L}{2} ||k^i_w||^2 + \alpha^i \langle \delta^i_w, k^i_w \rangle]
\]

where the last equality is due to $\mathbb{E}[\delta^i_w] = 0$ and Eq. (2).

Similarly, when considering $\mathcal{L}_\theta(\theta^{i+1}; w^{i+1}) - \mathcal{L}_\theta(\theta^i, w^{i+1})$, we have:
\[
\mathcal{L}(\theta^{i+1}; w^{i+1}) - \mathcal{L}(\theta^i, w^{i+1}) \leq \langle \nabla_\theta \mathcal{L}(\theta^{i+1}; w^{i+1}), \theta^{i+1} - \theta^i \rangle + \frac{L}{2} ||\theta^{i+1} - \theta^i||^2
\]


Take expectations on both sides and then we get:

\[
E[\mathcal{L}_\lambda(\theta^{t+1}:w^{t+1}) - \mathcal{L}_\lambda(\theta^t:w^t)] \leq E[(-\alpha' + \frac{L(\alpha')^2}{2})(||k'_\theta||^2_2 + ||\delta'_\theta||^2_2 + ||\delta'_w||^2_2)]
\]

\[
\leq E[(-\alpha' + \frac{L(\alpha')^2}{2})(||k'_\theta||^2_2 + ||\delta'_w||^2_2)] + \alpha' (\sigma^t)^2
\]

Above all, we have:

\[
E[\mathcal{L}_\lambda(\theta^{t+1},w^{t+1}) - \mathcal{L}_\lambda(\theta^t,w^t)] = E[\mathcal{L}_\lambda(\theta^{t+1},w^{t+1}) - \mathcal{L}_\lambda(\theta^t,w^t)] + E[\mathcal{L}_\lambda(\theta^t,w^{t+1}) - \mathcal{L}_\lambda(\theta^t,w^t)]
\]

\[
\leq E[(-\alpha' + \frac{L(\alpha')^2}{2})(||k'_\theta||^2_2 + ||k'_w||^2_2)] + 2\alpha' (\sigma^t)^2
\]

Then, because \(0 < \alpha' < \frac{2}{L}\), we have:

\[
\alpha' E[||k'_\theta||^2_2] + \alpha' E[||k'_w||^2_2] \leq \frac{2}{2-L\alpha'} E[\mathcal{L}_\lambda(\theta^t,w^t) - \mathcal{L}_\lambda(\theta^{t+1},w^{t+1}) + 2\alpha' (\sigma^t)^2]
\]

Summing the above inequality over \(t\) and using Eq. (4), we have

\[
\sum_{t=1}^{\infty} \alpha' E[||\frac{(\theta^{t+1},w^{t+1}) - (\theta^t,w^t)}{\alpha'}||^2_2] = \sum_{t=1}^{\infty} \alpha' E[||\frac{\theta^{t+1} - \theta^t}{\alpha'}||^2_2] + \sum_{t=1}^{\infty} \alpha' E[||\frac{w^{t+1} - w^t}{\alpha'}||^2_2]
\]

\[
= \sum_{t=1}^{\infty} \alpha' E[||k'_\theta||^2_2] + \sum_{t=1}^{\infty} \alpha' E[||k'_w||^2_2] \leq \infty
\]

Hence, by Lemma 3, there must exist an index sub-sequence \(\mathcal{M}\) such that

\[
\lim_{t \to \infty} E[||\frac{(\theta^{t+1},w^{t+1}) - (\theta^t,w^t)}{\alpha'}||^2_2] = 0.
\]

\[\square\]

References

1. Ghadimi, S., Lan, G. & Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.* **155**, 267–305 (2016).

2. Xu, Y. & Yin, W. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM J. on Optim.* **25**, 1686–1716 (2015).