Supplement for "SADAGRAD: Strongly Adaptive Stochastic Gradient Methods"

Zaiyi Chen [*1 2] **Yi Xu** [*2] **Enhong Chen** [1] **Tianbao Yang** [2]

1. Proof of Proposition 1

Proposition 1. *Let $\epsilon > 0$ be fixed, $H_0 = \gamma I$, $\gamma \geq \max_t \|\mathbf{g}_t\|_\infty$, $\mathrm{E}[F(\mathbf{w}_1) - F(\mathbf{w}_*)] \leq \epsilon_0$ and iteration number T satisfies $T \geq \frac{2}{\epsilon}\max\left\{\frac{\epsilon_0(\gamma+\max_i \|g_{1:T,i}\|_2)}{\eta\lambda}, \eta\sum_{i=1}^d \|g_{1:T,i}\|_2\right\}$. Algorithm 1 gives a solution $\widehat{\mathbf{w}}_T$ such that $\mathrm{E}[F(\widehat{\mathbf{w}}_T) - F_*] \leq \epsilon$.*

Proof. Let $\psi_0(\mathbf{w}) = 0$ and $\|\mathbf{x}\|_H = \sqrt{\mathbf{x}^\top H\mathbf{x}}$. First, we can see that $\psi_{t+1}(\mathbf{w}) \geq \psi_t(\mathbf{w})$ for any $t \geq 0$. Define $\mathbf{z}_t = \sum_{\tau=1}^t \mathbf{g}_t$ and $\Delta_\tau = (\partial F(\mathbf{w}_t) - \mathbf{g}_t)^\top(\mathbf{w}_t - \mathbf{w})$. Let ψ_t^* be defined by

$$\psi_t^*(g) = \sup_{\mathbf{x}\in\Omega} g^\top\mathbf{x} - \frac{1}{\eta}\psi_t(\mathbf{x})$$

Taking the summation of objective gap in all iterations, we have

$$\begin{aligned}
&\sum_{t=1}^T(F(\mathbf{w}_t) - F(\mathbf{w})) \leq \sum_{t=1}^T \partial F(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w})\\
&= \sum_{t=1}^T \mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}) + \sum_{t=1}^T \Delta_t\\
&= \sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_t - \sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w} - \frac{1}{\eta}\psi_T(\mathbf{w}) + \frac{1}{\eta}\psi_T(\mathbf{w})\\
&+ \sum_{t=1}^T \Delta_t\\
&\leq \frac{1}{\eta}\psi_T(\mathbf{w}) + \sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_t + \sum_{t=1}^T \Delta_t\\
&+ \sup_{\mathbf{x}\in\Omega}\left\{-\sum_{t=1}^T \mathbf{g}_t^\top\mathbf{x} - \frac{1}{\eta}\psi_T(\mathbf{x})\right\}\\
&= \frac{1}{\eta}\psi_T(\mathbf{w}) + \sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_t + \psi_T^*(-\mathbf{z}_T) + \sum_{t=1}^T \Delta_t
\end{aligned}$$

Note that

$$\begin{aligned}
&\psi_T^*(-\mathbf{z}_T) = -\sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_{T+1} - \frac{1}{\eta}\psi_T(\mathbf{w}_{T+1})\\
&\leq -\sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_{T+1} - \frac{1}{\eta}\psi_{T-1}(\mathbf{w}_{T+1})\\
&\leq \sup_{\mathbf{x}\in\Omega} -\mathbf{z}_T^\top\mathbf{x} - \frac{1}{\eta}\psi_{T-1}(\mathbf{x}) = \psi_{T-1}^*(-\mathbf{z}_T)\\
&\leq \psi_{T-1}^*(-\mathbf{z}_{T-1}) - \mathbf{g}_T^\top\nabla\psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2
\end{aligned}$$

where the last inequality uses the fact that $\psi_t(\mathbf{w})$ is 1-strongly convex w.r.t $\|\cdot\|_{\psi_t} = \|\cdot\|_{H_t}$ and consequentially $\psi_t^*(\mathbf{w})$ is η-smooth wr.t. $\|\cdot\|_{\psi_t^*} = \|\cdot\|_{H_t^{-1}}$. Thus, we have

$$\begin{aligned}
&\sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_t + \psi_T^*(-\mathbf{z}_T)\\
&\leq \sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_t + \psi_{T-1}^*(-\mathbf{z}_{T-1}) - \mathbf{g}_T^\top\nabla\psi_{T-1}^*(-\mathbf{z}_{T-1})\\
&+ \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2\\
&= \sum_{t=1}^{T-1} \mathbf{g}_t^\top\mathbf{w}_t + \psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2
\end{aligned}$$

By repeating this process, we have

$$\begin{aligned}
&\sum_{t=1}^T \mathbf{g}_t^\top\mathbf{w}_t + \psi_T^*(-\mathbf{z}_T)\\
&\leq \psi_0^*(-\mathbf{z}_0) + \frac{\eta}{2}\sum_{t=1}^T \|\mathbf{g}_t\|_{\psi_{t-1}^*}^2\\
&= \frac{\eta}{2}\sum_{t=1}^T \|\mathbf{g}_t\|_{\psi_{t-1}^*}^2
\end{aligned}$$

Then

$$\sum_{t=1}^T(F(\mathbf{w}_t) - F(\mathbf{w})) \leq \frac{1}{\eta}\psi_T(\mathbf{w}) + \frac{\eta}{2}\sum_{t=1}^T \|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 + \sum_{t=1}^T \Delta_t \quad (1)$$

Following the analysis in (Duchi et al., 2011), we have

$$\sum_{t=1}^{T} \|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 \leq 2 \sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2$$

Thus

$$\begin{aligned}
&\sum_{t=1}^{T}(F(\mathbf{w}_t) - F(\mathbf{w})) \\
&\leq \frac{\gamma\|\mathbf{w}-\mathbf{w}_1\|_2^2}{2\eta} + \frac{(\mathbf{w}-\mathbf{w}_1)^\top \mathrm{diag}(s_T)(\mathbf{w}-\mathbf{w}_1)}{2\eta} \\
&+ \eta \sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2 + \sum_{t=1}^{T} \Delta_t \\
&\leq \frac{\gamma + \max_i \|\mathbf{g}_{1:T,i}\|_2}{2\eta}\|\mathbf{w}-\mathbf{w}_1\|_2^2 + \eta \sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2 \\
&+ \sum_{t=1}^{T} \Delta_t
\end{aligned}$$

Now by the value of $T \geq \frac{2}{\epsilon}\max\left\{\frac{\epsilon_0(\gamma+\max_i \|\mathbf{g}_{1:T,i}\|_2)}{\eta\lambda}, \eta\sum_{i=1}^{d}\|g_{1:T,i}\|_2\right\}$, we have

$$\begin{aligned}
\frac{(\gamma + \max_i \|g_{1:T,i}\|_2)}{2\eta T} &\leq \frac{\lambda\epsilon}{4\epsilon_0} \\
\frac{\eta\sum_{i=1}^{d}\|g_{1:T,i}\|_2}{T} &\leq \frac{\epsilon}{2}
\end{aligned}$$

Dividing by T on both sides and setting $\mathbf{w} = \mathbf{w}_*$, following the inequality (3) and the convexity of $F(\mathbf{w})$ we have

$$F(\widehat{\mathbf{w}}) - F_* \leq \frac{\lambda\epsilon}{4\epsilon_0}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 + \frac{\epsilon}{2} + \frac{1}{T}\sum_{t=1}^{T}\Delta_t$$

Let $\{\mathcal{F}_t\}$ be the filtration associated with Algorithm 1 in the paper. Noticing that T is a random variable with respect to $\{\mathcal{F}_t\}$, we cannot get rid of the last term directly. Define the Sequence $\{X_t\}_{t\in\mathbb{N}_+}$ as

$$X_t = \frac{1}{t}\sum_{i=1}^{t}\Delta_i = \frac{1}{t}\sum_{i=1}^{t}\langle \mathbf{g}_i - \mathrm{E}[\mathbf{g}_i], \mathbf{w}_i - \mathbf{w}_*\rangle \qquad (2)$$

where $\mathrm{E}[\mathbf{g}_i] \in \partial F(\mathbf{w}_i)$. Since $\mathrm{E}\left[\mathbf{g}_{t+1} - \mathrm{E}[\mathbf{g}_{t+1}]\right] = 0$ and $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} \eta\mathbf{w}^\top\left(\frac{1}{t}\sum_{\tau=1}^{t}\mathbf{g}_\tau\right) + \frac{1}{t}\psi_t(\mathbf{w})$, which is measurable with respect to $\mathbf{g}_1,\ldots,\mathbf{g}_t$ and $\mathbf{w}_1,\ldots,\mathbf{w}_t$, it is easy to see $\{\Delta_t\}_{t\in N}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$, e.g. $\mathrm{E}[\Delta_t|\mathcal{F}_{t-1}] = 0$. On the other hand, since $\|\mathbf{g}_t\|_2$ is upper bounded (e.g., by G), following the statement of T in the theorem, $T \leq N = \frac{4}{\epsilon^2}\max\{(\frac{2G\epsilon_0}{\theta\lambda})^2, \theta^2 d^2 G^2\} < \infty$ always holds. Then following Lemma 1 below we have that $\mathrm{E}[X_T] = 0$.

Now taking the expectation we have that

$$\begin{aligned}
&\mathrm{E}[F(\widehat{\mathbf{w}}) - F_*] \\
&\leq \mathrm{E}\left[\frac{\lambda\epsilon}{4\epsilon_0}\|\mathbf{w}-\mathbf{w}_1\|_2^2\right] + \frac{\epsilon}{2} + \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Delta_t\right] \\
&\leq \frac{\epsilon}{2\epsilon_0}\mathrm{E}\left[F(\mathbf{w}_1) - F(\mathbf{w}_*)\right] + \frac{\epsilon}{2} + 0 = \epsilon
\end{aligned}$$

Then we finish the proof. □

Lemma 1. *Let $\{\Delta_t\}_{t\in\mathbb{N}_+}$ be a martingale difference sequence w.r.t the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, T is a stopping time such that $\{T = t\} \in \mathcal{F}_t$ for all $t \in \mathbb{N}$. If $0 < T \leq N < \infty$, then we have*

$$\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Delta_t\right] = 0.$$

Proof.

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Delta_t\right] &= \mathrm{E}\left[\mathrm{E}[\frac{1}{T}\sum_{t=1}^{T}\Delta_t|\mathcal{F}_N]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\mathbb{I}(T=n)\mathrm{E}[\frac{1}{T}\sum_{t=1}^{T}\Delta_t|\mathcal{F}_n]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\mathrm{E}[\frac{\mathbb{I}(T=n)}{T}\sum_{t=1}^{T}\Delta_t|\mathcal{F}_n]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\mathrm{E}[\frac{\mathbb{I}(T=n)}{n}\sum_{t=1}^{n}\Delta_t|\mathcal{F}_n]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\frac{\mathbb{I}(T=n)}{n}\mathrm{E}[\sum_{t=1}^{n}\Delta_t|\mathcal{F}_n]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\frac{\mathbb{I}(T=n)}{n}\sum_{t=1}^{n}\mathrm{E}[\Delta_t|\mathcal{F}_n]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\frac{\mathbb{I}(T=n)}{n}\sum_{t=1}^{n}\mathrm{E}\big[\mathrm{E}[\Delta_t|\mathcal{F}_n]\big|\mathcal{F}_{t-1}\big]\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{T}\frac{\mathbb{I}(T=n)}{n}\sum_{t=1}^{n}\mathrm{E}[\Delta_t|\mathcal{F}_{t-1}]\right] \\
&= 0
\end{aligned}$$

where $\mathbb{I}(T = n)$ is the indicator function. The first equation follows from the definition of conditional expectation and $T \leq N$; the second equation follows from the fact that $\sum_{n=1}^{T}\mathbb{I}(T = n) = 1$; the third and fifth equations follow from the definition of stopping time ($(T = n) \in \mathcal{F}_n$); the seventh and last equations follow from the definition of martingale difference sequence; and eighth equation follows from Theorem 5.1.6 in (Durrett, 2010). □

2. Proof of Theorem 1

Theorem 1. *Consider SCO (1) with a property (3) and a given* $\epsilon > 0$. *Assume* $H_0 = \gamma I$ *in Algorithm 1 and* $\gamma \geq \max_{k,\tau} \|\mathbf{g}_\tau^k\|_\infty$, $F(\mathbf{w}_0) - F_* \leq \epsilon_0$ *and* t_k *is the minimum number such that* $t_k \geq \frac{2}{\sqrt{\lambda\epsilon_k}} \max\left\{\frac{2(\gamma+\max_i \|g_{1:t_k,i}^k\|_2)}{\theta}, \theta\sum_{i=1}^d \|g_{1:t_k,i}^k\|_2\right\}$. *With* $K = \lceil \log_2(\epsilon_0/\epsilon)\rceil$, *we have* $\mathrm{E}[F(\mathbf{w}_K) - F_*] \leq \epsilon$.

Proof of Theorem 1. We will show by induction that $\mathrm{E}[F(\mathbf{w}_k) - F_*] \leq \epsilon_k \triangleq \frac{\epsilon_0}{2^k}$ for $k = 0, 1, \ldots, K$, which leads to our conclusion when $k = K = \lceil \log(\epsilon_0/\epsilon)\rceil$.

The inequality holds obviously for $k = 0$. Conditioned on $\mathrm{E}[F(\mathbf{w}_{k-1}) - F_*] \leq \epsilon_{k-1}$, we will show that $\mathrm{E}[F(\mathbf{w}_k) - F_*] \leq \epsilon_k$. We will modify Proposition 1, then use it to the k-th epoch of Algorithm 2 conditioned on randomness in previous epoches. Let E_k denotes the expectation over all randomness before the last iteration of the k-th epoch and $\mathrm{E}_{k|1:k-1}$ denotes the expectation over the randomness in the k-th epoch given the randomness before k-th epoch. Given $\mathbf{w}_{k-1}$, we let $\mathbf{w}_{k-1}^*$ denote the optimal solution that is closest to $\mathbf{w}_{k-1}$ [1]. According to the proof of Propostion 1, We have

$$\begin{aligned}
&\mathrm{E}_{k|1:k-1}[F(\mathbf{w}_k) - F(\mathbf{w}_{k-1}^*)] \\
&\leq \mathrm{E}_{k|1:k-1}\left[\frac{\gamma + \max_i \|\mathbf{g}_{1:t_k,i}^k\|_2}{2\eta_k t_k}\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^*\|_2^2\right. \\
&\left.+\frac{\eta_k \sum_{i=1}^d \|\mathbf{g}_{1:t_k,i}^k\|_2}{t_k} + \sum_{t=1}^{t_k}\langle \mathrm{E}[\mathbf{g}_t^k] - \mathbf{g}_t^k, \mathbf{w}_t^k - \mathbf{w}_{k-1}^*\rangle\right]
\end{aligned}$$

By the value of $\eta_k = \theta\sqrt{\epsilon_k/\lambda}$ and $t_k \geq \max\left\{\frac{4(\gamma+\max_i \|g_{1:t_k,i}^k\|_2)}{\theta\sqrt{\lambda\epsilon_k}}, \frac{2\theta\sum_{i=1}^d \|g_{1:t_k,i}^k\|_2}{\sqrt{\lambda\epsilon_k}}\right\}$, we have

$$\begin{aligned}
\frac{(\gamma + \max_i \|g_{1:t_k,i}^k\|_2)}{2\eta_k t_k} &\leq \frac{\lambda}{8} \\
\frac{\eta_k \sum_{i=1}^d \|g_{1:t_k,i}^k\|_2}{t_k} &\leq \frac{\epsilon_k}{2}
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathrm{E}_{k|1:k-1}[F(\mathbf{w}_k) - F(\mathbf{w}_{k-1}^*)] \\
&\leq \mathrm{E}_{k|1:k-1}\left[\frac{\lambda}{8}\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^*\|_2^2 + \frac{\epsilon_k}{2}\right. \\
&\left.+ \sum_{t=1}^{t_k}\langle \mathrm{E}[\mathbf{g}_t^k] - \mathbf{g}_t^k, \mathbf{w}_t^k - \mathbf{w}_{k-1}^*\rangle\right]
\end{aligned}$$

[1] Since we only assume the condition (3) that does not necessarily imply the uniqueness of the optimal solutions.

Then following the similar arguments in Proposition 1, we have

$$\begin{aligned}
&\mathrm{E}_{k|1:k-1}[F(\mathbf{w}_k) - F(\mathbf{w}_{k-1}^*)] \\
&\leq \mathrm{E}_{k|1:k-1}\left[\frac{\lambda}{8}\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^*\|_2^2 + \frac{\epsilon_k}{2}\right]
\end{aligned}$$

Taking expectation over randomness in stages $1, \ldots, k-1$, we have

$$\begin{aligned}
&\mathrm{E}[F(\mathbf{w}_k) - F(\mathbf{w}_{k-1}^*)] \leq \mathrm{E}\left[\frac{\lambda}{8}\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^*\|_2^2\right] + \frac{\epsilon_k}{2} \\
&\leq \frac{1}{4}\mathrm{E}[F(\mathbf{w}_{k-1}) - F_*] + \frac{\epsilon_k}{2} \\
&\leq \frac{\epsilon_{k-1}}{4} + \frac{\epsilon_k}{2} = \epsilon_k
\end{aligned}$$

Therefore by induction, we have $\mathrm{E}[F(\mathbf{w}_K) - F_*] \leq \epsilon_K \leq \epsilon$. □

3. Proof of Theorem 2

Lemma 2. *Consider SCO (4) with the property (3). Let* $H_0 = \gamma I$ *in Algorithm 3 and* $\gamma \geq \max_t \|\mathbf{g}_t\|_\infty$. *For any* $\mathbf{w} \in \Omega$ *and its closest optimal solution* $\mathbf{w}_*$, *we have*

$$\begin{aligned}
&F(\widetilde{\mathbf{w}}_T) - F(\mathbf{w}) \leq \frac{G\|\mathbf{w}_1 - \mathbf{w}_{T+1}\|_2}{T} \\
&+ \frac{1}{T}\sum_{t=1}^T (\mathrm{E}[\mathbf{g}_t] - \mathbf{g}_t)^\top(\mathbf{w}_t - \mathbf{w}) \\
&+ \left[\frac{\eta\sum_{i=1}^d \|g_{1:T,i}\|_2}{T} + \frac{\gamma + \max_i \|g_{1:T,i}\|_2}{2\eta T}\|\mathbf{w} - \mathbf{w}_1\|_2^2\right]
\end{aligned}$$

where $\widetilde{\mathbf{w}}_T = \sum_{t=2}^{T+1}\mathbf{w}_t/T$.

Proof. This proof is similar to the proof of Proposition 1, but we do not take expectation here. For completeness, we give the proof here. Throughout the whole proof, we set the notation $\mathbf{g}_t$ as the stochastic gradient of $f(\mathbf{w}_t)$ and as a result $\mathrm{E}[\mathbf{g}_t] \in \partial f(\mathbf{w}_t)$. Let $\psi_0(\mathbf{w}) = 0$ and $\|\mathbf{x}\|_H = \sqrt{\mathbf{x}^\top H\mathbf{x}}$. First, we can see that $\psi_{t+1}(\mathbf{w}) \geq \psi_t(\mathbf{w})$ for any $t \geq 0$. Define $\mathbf{z}_t = \sum_{\tau=1}^t \mathbf{g}_t$ and $\Delta_\tau = (\partial f(\mathbf{w}_t) - \mathbf{g}_t)^\top(\mathbf{w}_t - \mathbf{w})$. Let ψ_t^* be defined by

$$\psi_t^*(g) = \sup_{\mathbf{x}\in\Omega} g^\top\mathbf{x} - \frac{1}{\eta}\psi_t(\mathbf{x}) - t\phi(\mathbf{x})$$

Taking the summation of objective gap in all iterations, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}(f(\mathbf{w}_t) - f(\mathbf{w}) + \phi(\mathbf{w}_t) - \phi(\mathbf{w})) \\
&\leq \sum_{t=1}^{T}(\partial f(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}) + \phi(\mathbf{w}_t) - \phi(\mathbf{w})) \\
&= \sum_{t=1}^{T}\mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}) + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}(\phi(\mathbf{w}_t) - \phi(\mathbf{w})) \\
&= \sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_t - \sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w} - \frac{1}{\eta}\psi_T(\mathbf{w}) - T\phi(\mathbf{w}) \\
&+ \frac{1}{\eta}\psi_T(\mathbf{w}) + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}\phi(\mathbf{w}_t) \\
&\leq \frac{1}{\eta}\psi_T(\mathbf{w}) + \sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_t + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}\phi(\mathbf{w}_t) \\
&+ \sup_{\mathbf{x}\in\Omega}\left\{-\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x} - \frac{1}{\eta}\psi_T(\mathbf{x}) - T\phi(\mathbf{x})\right\} \\
&= \frac{1}{\eta}\psi_T(\mathbf{w}) + \sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_t + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}\phi(\mathbf{w}_t) \\
&+ \psi_T^*(-\mathbf{z}_T) \qquad (3)
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\psi_T^*(-\mathbf{z}_T) \\
&= -\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_{T+1} - \frac{1}{\eta}\psi_T(\mathbf{w}_{T+1}) - T\phi(\mathbf{w}_{T+1}) \\
&\leq -\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_{T+1} - \frac{1}{\eta}\psi_{T-1}(\mathbf{w}_{T+1}) - (T-1)\phi(\mathbf{w}_{T+1}) \\
&- \phi(\mathbf{w}_{T+1}) \\
&\leq \sup_{\mathbf{x}\in\Omega}\left\{-\mathbf{z}_T^\top\mathbf{x} - \frac{1}{\eta}\psi_{T-1}(\mathbf{x}) - (T-1)\phi(\mathbf{x})\right\} \\
&- \phi(\mathbf{w}_{T+1}) \\
&= \psi_{T-1}^*(-\mathbf{z}_T) - \phi(\mathbf{w}_{T+1}) \\
&\leq \psi_{T-1}^*(-\mathbf{z}_{T-1}) - \mathbf{g}_T^\top\nabla\psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2 \\
&- \phi(\mathbf{w}_{T+1})
\end{aligned}
$$

where the last inequality uses the fact that $\psi_t(\mathbf{w})$ is 1-strongly convex w.r.t $\|\cdot\|_{\psi_t} = \|\cdot\|_{H_t}$ and consequentially $\psi_t^*(\mathbf{w})$ is η-smooth w.r.t. $\|\cdot\|_{\psi_t^*} = \|\cdot\|_{H_t^{-1}}$. Thus, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_t + \psi_T^*(-\mathbf{z}_T) \\
&\leq \sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_t + \psi_{T-1}^*(-\mathbf{z}_{T-1}) - \mathbf{g}_T^\top\nabla\psi_{T-1}^*(-\mathbf{z}_{T-1}) \\
&+ \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2 - \phi(\mathbf{w}_{T+1}) \\
&= \sum_{t=1}^{T-1}\mathbf{g}_t^\top\mathbf{w}_t + \psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2 \\
&- \phi(\mathbf{w}_{T+1})
\end{aligned}
$$

By repeating this process, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{w}_t + \psi_T^*(-\mathbf{z}_T) \\
&\leq \psi_0^*(-\mathbf{z}_0) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 - \sum_{t=1}^{T}\phi(\mathbf{w}_{t+1}) \\
&= \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 - \sum_{t=1}^{T}\phi(\mathbf{w}_{t+1}) \qquad (4)
\end{aligned}
$$

Plugging inequality (4) in inequality (3), then

$$
\begin{aligned}
&\sum_{t=1}^{T}(F(\mathbf{w}_t) - F(\mathbf{w})) \\
&\leq \frac{1}{\eta}\psi_T(\mathbf{w}) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 + \sum_{t=1}^{T}\Delta_t + \phi(\mathbf{w}_1) \\
&- \phi(\mathbf{w}_{T+1})
\end{aligned}
$$

By adding $F(\mathbf{w}_{T+1}) - F(\mathbf{w}_1)$ on the both sides of above inequality and using the fact that $F(\mathbf{w}) = f(\mathbf{w}) + \phi(\mathbf{w})$, we get

$$
\begin{aligned}
&\sum_{t=2}^{T+1}(F(\mathbf{w}_t) - F(\mathbf{w})) \\
&\leq \frac{1}{\eta}\psi_T(\mathbf{w}) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 + \sum_{t=1}^{T}\Delta_t + f(\mathbf{w}_{T+1}) \\
&- f(\mathbf{w}_1)
\end{aligned}
$$

Following the analysis in (Duchi et al., 2011), we have

$$
\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 \leq 2\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2
$$

Thus

$$
\begin{aligned}
&\sum_{t=2}^{T+1}(F(\mathbf{w}_t)-F(\mathbf{w}))\\
&\leq \frac{\gamma\|\mathbf{w}-\mathbf{w}_1\|_2^2}{2\eta}+\frac{(\mathbf{w}-\mathbf{w}_1)^\top \mathrm{diag}(s_T)(\mathbf{w}-\mathbf{w}_1)}{2\eta}\\
&+\eta\sum_{i=1}^d\|\mathbf{g}_{1:T,i}\|_2+\sum_{t=1}^T\Delta_t+f(\mathbf{w}_{T+1})-f(\mathbf{w}_1)\\
&\leq \frac{\gamma+\max_i\|\mathbf{g}_{1:T,i}\|_2}{2\eta}\|\mathbf{w}-\mathbf{w}_1\|_2^2+\eta\sum_{i=1}^d\|\mathbf{g}_{1:T,i}\|_2\\
&+\sum_{t=1}^T\Delta_t+(\partial f(\mathbf{w}_{T+1}))^\top(\mathbf{w}_{T+1}-\mathbf{w}_1)\\
&\leq \frac{\gamma+\max_i\|\mathbf{g}_{1:T,i}\|_2}{2\eta}\|\mathbf{w}-\mathbf{w}_1\|_2^2+\eta\sum_{i=1}^d\|\mathbf{g}_{1:T,i}\|_2\\
&+\sum_{t=1}^T\Delta_t+G\|\mathbf{w}_{T+1}-\mathbf{w}_1\|_2
\end{aligned}
$$

where the last inequality hold using Cauchy-Schwartz Inequality and the fact that $\|\partial f(\mathbf{w}_{T+1})\|\leq G$. Dividing by T on both sides, then we finish the proof by using the convexity of $F(\mathbf{w})$. $\square$

Theorem 2. *For a given $\epsilon>0$, let $K=\lceil\log_2(\epsilon_0/\epsilon)\rceil$. Assume $H_0=\gamma I$ and $\gamma\geq\max_{k,\tau}\|\mathbf{g}_\tau^k\|_\infty$, $F(\mathbf{w}_0)-F_*\leq\epsilon_0$ and t_k is the minimum number such that $t_k\geq\frac{3}{\sqrt{\lambda\epsilon_k}}\max\left\{A_k,\frac{\sqrt{\lambda}G\|\mathbf{w}_1^k-\mathbf{w}_{t_k+1}^k\|_2}{\sqrt{\epsilon_k}}\right\}$, where $A_k=\max\left\{\frac{2(\gamma+\max_i\|g_{1:t_k,i}^k\|_2)}{\theta},\theta\sum_{i=1}^d\|g_{1:t_k,i}^k\|_2\right\}$. Algorithm 4 guarantees that $\mathrm{E}[F(\mathbf{w}_K)-F_*]\leq\epsilon$.*

Proof. This result is proved by revising Lemma 2 to hold for a bounded stopping time t_k of the supermartingale sequence X_t in (2).

Taking the expectation of Lemma 2, we have that

$$
\begin{aligned}
&\mathrm{E}[F(\widetilde{\mathbf{w}}_T)-F(\mathbf{w})]\leq \mathrm{E}\left[\frac{G\|\mathbf{w}_1-\mathbf{w}_{T+1}\|_2}{T}\right]\\
&+\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T(\mathrm{E}[\mathbf{g}_t]-\mathbf{g}_t)^\top(\mathbf{w}_t-\mathbf{w})\right]\\
&+\mathrm{E}\left[\frac{\eta\sum_{i=1}^d\|g_{1:T,i}\|_2}{T}\right.\\
&\left.+\frac{\gamma+\max_i\|g_{1:T,i}\|_2}{2\eta T}\|\mathbf{w}-\mathbf{w}_1\|_2^2\right]
\end{aligned}
$$

Then following the same arguments to Proposition 1, we have that

$$
\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T(\mathrm{E}[\mathbf{g}_t]-\mathbf{g}_t)^\top(\mathbf{w}_t-\mathbf{w})\right]=0
$$

Similar to the induction of Theorem 1, let $\eta_k=\theta\sqrt{\epsilon_k/\lambda}$ and the iteration number t_k in k-th epoch to be the smallest number satisfying following inequalities

$$
\begin{aligned}
\frac{(\gamma+\max_i\|g_{1:t_k,i}^k\|_2)}{2\eta_k t_k}&\leq\frac{\lambda}{12}\\
\frac{\eta_k\sum_{i=1}^d\|g_{1:t_k,i}^k\|_2}{t_k}&\leq\frac{\epsilon_k}{3}\\
\frac{G\|\mathbf{w}_1^k-\mathbf{w}_{t_k+1}^k\|}{t_k}&\leq\frac{\epsilon_k}{3}
\end{aligned}
$$

Thus conditioned on $1,\ldots,k-1$-th epoches, we have that

$$
\begin{aligned}
&E_{k|1:k-1}[F(\mathbf{w}_k)-F(\mathbf{w}_{k-1}^*)]\\
&\leq \mathrm{E}_{k|1:k-1}\left[\frac{\lambda}{12}\|\mathbf{w}_{k-1}-\mathbf{w}_{k-1}^*\|_2^2+\frac{2\epsilon_k}{3}\right]
\end{aligned}
$$

Taking expectation over randomness in stages $1,\ldots,k-1$, we have

$$
\begin{aligned}
&\mathrm{E}[F(\mathbf{w}_k)-F(\mathbf{w}_{k-1}^*)]\leq\mathrm{E}\left[\frac{\lambda}{12}\|\mathbf{w}_{k-1}-\mathbf{w}_{k-1}^*\|_2^2\right]+\frac{2\epsilon_k}{3}\\
&\leq\frac{1}{6}\mathrm{E}[F(\mathbf{w}_{k-1})-F_*]+\frac{2\epsilon_k}{3}\\
&\leq\frac{\epsilon_{k-1}}{6}+\frac{2\epsilon_k}{3}=\epsilon_k
\end{aligned}
$$

Therefore by induction, we have $\mathrm{E}[F(\mathbf{w}_K)-F_*]\leq\epsilon_K\leq\epsilon$. $\square$

4. Proof of Theorem 3

Theorem 3. *Under the same assumptions as Theorem 1 and $F(\mathbf{w}_0)-F_*\leq\epsilon_0$, where $\mathbf{w}_0$ is an initial solution. Let $\lambda_1\geq\lambda$, $\epsilon\leq\frac{\epsilon_0}{2}$, $K=\log_2\frac{\epsilon_0}{\epsilon}$ and $t_k^{(s)}\geq\frac{2}{\sqrt{\lambda_s\epsilon_k}}\max\left\{\frac{2(\gamma+\max_i\|g_{1:t_k,i}^k\|_2)}{\theta},\theta\sum_{i=1}^d\|g_{1:t_k,i}^k\|_2\right\}$. Then with at most a total number of $S=\lceil\log_2\left(\frac{\lambda_1}{\lambda}\right)\rceil+1$ calls of* SADAGRAD *and a worse-cast iteration complexity of $O(1/(\lambda\epsilon))$, Algorithm 5 finds a solution $\mathbf{w}^{(S)}$ such that $\mathrm{E}[F(\mathbf{w}^{(S)})-F_*]\leq\epsilon$.*

Proof. Since $\lambda_1/\lambda>1$, then $F(\mathbf{w}_0)-F_*\leq(\lambda_1/\lambda)\epsilon_0$. Following the proof of Theorem 1, we can show that

$$
\mathrm{E}[F(\mathbf{w}^{(1)})-F_*]\leq\frac{(\lambda_1/\lambda)\epsilon_0}{2^K}=\left(\frac{\lambda_1}{\lambda}\right)\epsilon
$$

with $K=\log_2\frac{\epsilon_0}{\epsilon}$ and $t_k^{(1)}\geq\frac{2}{\sqrt{\lambda(\frac{\lambda_1}{\lambda}\epsilon_k)}}\max\left\{\frac{2(\gamma+\max_i\|g_{1:t_k,i}^k\|_2)}{\theta},\theta\sum_{i=1}^d\|g_{1:t_k,i}^k\|_2\right\}$,

$k = 1, \ldots, K$. Next, since $\epsilon \leq \frac{\epsilon_0}{2}$, then we have $\mathrm{E}[F(\mathbf{w}^{(1)}) - F_*] \leq \left(\frac{\lambda_1}{\lambda}\right)\frac{\epsilon_0}{2} = \left(\frac{\lambda_2}{\lambda}\right)\epsilon_0$. By running SADAGRAD from $\mathbf{w}^{(1)}$, Theorem 1 ensures that

$$\mathrm{E}[F(\mathbf{w}^{(2)}) - F_*] \leq \frac{\mathrm{E}[F(\mathbf{w}^{(1)}) - F_*]}{2^K} \leq \frac{(\lambda_2/\lambda)\epsilon_0}{2^K} = \left(\frac{\lambda_2}{\lambda}\right)\epsilon$$

By continuing the process, with $S = \left\lceil \log_2\left(\frac{\lambda_1}{\lambda}\right)\right\rceil + 1$, we have

$$\mathrm{E}[F(\mathbf{w}^{(S)}) - F_*] \leq \left(\frac{\lambda_S}{\lambda}\right)\epsilon \leq \epsilon \tag{5}$$

The total number of iterations for the S calls of SADAGRAD is upper bounded by

$$\begin{aligned} T_{\text{total}} &= \sum_{s=1}^{S}\sum_{k=1}^{K} t_k^{(s)} \leq \sum_{s=1}^{S}\frac{C}{\lambda_s\epsilon_0}\sum_{k=1}^{K} 2^{k-1} \\ &= \frac{C}{\lambda_1\epsilon_0}\sum_{s=1}^{S} 2^{s-1}\sum_{k=1}^{K} 2^{k-1} \\ &= O\left(\frac{1}{\lambda\epsilon}\right) \end{aligned}$$

for some $C > 0$. □

Acknowledgement

We thank Prof. Qihe Tang from University of Iowa for his help on the proof of Lemma 1.

References

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12 (Jul):2121–2159, 2011.

Durrett, Rick. *Probability: theory and examples*. Cambridge university press, 2010.