Supplementary Material: Stochastic Gradient Descent with Only One Projection

Mehrdad Mahdavi[†], Tianbao Yang[‡], Rong Jin[†], Shenghuo Zhu[⋆], and Jinfeng Yi[†]

[†]Dept. of Computer Science and Engineering, Michigan State University, MI, USA
[‡]Machine Learning Lab, GE Global Research, CA, USA
[⋆]NEC Laboratories America, CA, USA
[†]{mahdavim,rongjin,yijinfen}@msu.edu,[‡]tyang@ge.com,[⋆]zsh@nec-labs.com

A Proof of Lemma 1

Following the standard analysis of gradient descent methods, we have for any $\mathbf{x} \in \mathcal{B}$,

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 - \|\mathbf{x}_t - \mathbf{x}\|_2^2 \leq \|\mathbf{x}'_{t+1} - \mathbf{x}\|_2^2 - \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
&= \|\mathbf{x}_t - \eta_t(\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) + \lambda_t \nabla g(\mathbf{x}_t)) - \mathbf{x}\|_2^2 - \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
&\leq \eta_t^2 \|\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) + \lambda_t \nabla g(\mathbf{x}_t)\|_2^2 - 2\eta_t (\mathbf{x}_t - \mathbf{x})^\top (\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) + \lambda_t \nabla g(\mathbf{x}_t)) \\
&\leq \eta_t^2 \|\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) + \lambda_t \nabla g(\mathbf{x}_t)\|_2^2 \\
&\quad - 2\eta_t (\mathbf{x}_t - \mathbf{x})^\top \underbrace{(\nabla f(\mathbf{x}_t) + \lambda_t \nabla g(\mathbf{x}_t))}_{\equiv \nabla_{\mathbf{x}} L(\mathbf{x}_t, \lambda_t)} + 2\eta_t \underbrace{(\mathbf{x} - \mathbf{x}_t)^\top (\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_t))}_{\equiv \zeta_t(\mathbf{x})},
\end{aligned}$$

Then we have

$$\begin{aligned}
&(\mathbf{x}_t - \mathbf{x})^\top \nabla_{\mathbf{x}} L(\mathbf{x}_t, \lambda_t) \leq \frac{1}{2\eta_t} \left(\|\mathbf{x}_t - \mathbf{x}\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 \right) + \frac{\eta_t}{2} \|\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) + \lambda_t \nabla g(\mathbf{x}_t)\|_2^2 + \zeta_t(\mathbf{x}) \\
&\leq \frac{1}{2\eta_t} \left(\|\mathbf{x}_t - \mathbf{x}\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 \right) + \eta_t \|\widetilde{\nabla} f(\mathbf{x}_t, \xi_t)\|_2^2 + \eta_t \lambda_t^2 \|\nabla g(\mathbf{x}_t)\|_2^2 + \zeta_t(\mathbf{x}) \\
&\leq \frac{1}{2\eta_t} \left(\|\mathbf{x}_t - \mathbf{x}\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 \right) \\
&\quad + 2\eta_t \underbrace{\|\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_t)\|_2^2}_{\equiv \Delta_t} + 2\eta_t \|\nabla f(\mathbf{x}_t)\|_2^2 + \eta_t \lambda_t^2 \|\nabla g(\mathbf{x}_t)\|_2^2 + \zeta_t(\mathbf{x})
\end{aligned}$$

By using the bound on $\|\nabla f(\mathbf{x}_t)\|_2$ and $\|\nabla g(\mathbf{x}_t)\|_2$, we obtain the first inequality in Lemma 1. To prove the second inequality in Lemma 1, we follow the same analysis, i.e.,

$$\begin{aligned}
&|\lambda_{t+1} - \lambda|^2 - |\lambda_t - \lambda|^2 \leq |\lambda_t + \eta_t (g(\mathbf{x}_t) - \gamma \lambda_t)|^2 - |\lambda_t - \lambda|^2 \\
&\leq \eta_t^2 |g(\mathbf{x}_t) - \gamma \lambda_t|^2 + 2\eta_t (\lambda_t - \lambda) \underbrace{(g(\mathbf{x}_t) - \gamma \lambda_t)}_{\equiv \nabla_\lambda L(\mathbf{x}_t, \lambda_t)}.
\end{aligned}$$

Then we have

$$(\lambda - \lambda_t) \nabla_\lambda L(\mathbf{x}_t, \lambda_t) \leq \frac{1}{2\eta_t} \left(|\lambda_t - \lambda|^2 - |\lambda_{t+1} - \lambda|^2 \right) + \frac{\eta_t}{2} |g(\mathbf{x}_t) - \gamma \lambda_t|^2.$$

By induction, it is straightforward to show that $\lambda_t \leq C_2/\gamma$, which yields the second inequality in Lemma 1, i.e.,

$$(\lambda - \lambda_t) \nabla_\lambda L(\mathbf{x}_t, \lambda_t) \leq \frac{1}{2\eta_t} \left(|\lambda_t - \lambda|^2 - |\lambda_{t+1} - \lambda|^2 \right) + 2\eta_t C_2^2.$$

B Proof of Lemma 2

Since $L_t(\mathbf{x}, \lambda)$ is convex in $\mathbf{x}$ and concave in λ, we have the following inequalities

$$L(\mathbf{x}, \lambda_t) - L(\mathbf{x}_t, \lambda_t) \geq (\mathbf{x} - \mathbf{x}_t)^\top \nabla_{\mathbf{x}} L(\mathbf{x}_t, \lambda_t),$$
$$L(\mathbf{x}_t, \lambda) - L(\mathbf{x}_t, \lambda_t) \leq (\lambda - \lambda_t) \nabla_\lambda L(\mathbf{x}_t, \lambda_t).$$

Using the inequalities in Lemma 1, we have

$$L(\mathbf{x}_t, \lambda_t) - L(\mathbf{x}, \lambda_t) \leq \frac{1}{2\eta_t} \left(\|\mathbf{x} - \mathbf{x}_t\|_2^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|_2^2 \right) + 2\eta_t G_1^2 + \eta_t G_2^2 \lambda_t^2 + 2\eta_t \Delta_t + \zeta_t(\mathbf{x}),$$
$$L(\mathbf{x}_t, \lambda) - L(\mathbf{x}_t, \lambda_t) \leq \frac{1}{2\eta_t} \left(|\lambda - \lambda_t|^2 - |\lambda - \lambda_{t+1}|^2 \right) + 2\eta_t C_2^2,$$

where $\zeta_t(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_t)^\top (\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_t))$ as abbreviated before. Since $\eta_1 = \cdots = \eta_T$, denoted by η, by taking summation of above two inequalities over $t = 1, \cdots, T$, we get

$$\sum_{t=1}^T L(\mathbf{x}_t, \lambda) - L(\mathbf{x}, \lambda_t) \leq \frac{\|\mathbf{x}\|_2^2}{2\eta} + \frac{\lambda^2}{2\eta} + 2\eta T(G_1^2 + C_2^2) + \sum_t \eta G_2^2 \lambda_t^2 + 2\eta \sum_{t=1}^T \Delta_t + \sum_{t=1}^T \zeta_t(\mathbf{x}).$$

By plugging the expression of $L(\mathbf{x}, \lambda)$, and due to $\|\mathbf{x}\|_2 \leq 1$, we have

$$\sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x})) + \lambda \sum_{t=1}^T g(\mathbf{x}_t) - \left(\frac{\gamma T}{2} + \frac{1}{2\eta} \right) \lambda^2$$
$$\leq \frac{1}{2\eta} + 2\eta T(G_1^2 + C_2^2) + \sum_t (\eta G_2^2 - \gamma/2) \lambda_t^2 + \sum_t \lambda_t g(\mathbf{x}) + 2\eta \sum_{t=1}^T \Delta_t + \sum_{t=1}^T \zeta_t(\mathbf{x}).$$

Let $\mathbf{x} = \mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{K}} f(\mathbf{x})$. By taking minimization over $\lambda \geq 0$ on left hand side and considering $\eta = \gamma/(2G_2^2)$, we have

$$\sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{[\sum_{t=1}^T g(\mathbf{x}_t)]_+^2}{2(\gamma T + 2G_2^2/\gamma)} \leq \frac{G_2^2}{\gamma} + \frac{(G_1^2 + C_2^2)}{G_2^2} \gamma T + \frac{\gamma}{G_2^2} \sum_{t=1}^T \Delta_t + \sum_{t=1}^T \zeta_t(\mathbf{x}^*)$$

C Proof of Lemma 3

Since $F(\mathbf{x})$ is strongly convex in $\mathbf{x}$, we have

$$F(\mathbf{x}) - F(\mathbf{x}_t) \geq (\mathbf{x} - \mathbf{x}_t)^\top \nabla F(\mathbf{x}_t) + \frac{\beta}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2.$$

Following the same analysis as in Lemma 1, we have

$$\begin{aligned}(\mathbf{x}_t - \mathbf{x})^\top \nabla F(\mathbf{x}_t) &\leq \frac{1}{2\eta_t} \left(\|\mathbf{x} - \mathbf{x}_t\|_2^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|_2^2 \right) + \frac{\eta_t}{2} \|\widetilde{\nabla} f(\mathbf{x}_t, \xi_t) + p(\mathbf{x}_t) \lambda_0 \nabla g(\mathbf{x}_t)\|_2^2 \\ &\quad + \zeta_t(\mathbf{x}) - \frac{\beta}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \\ &\leq \frac{1}{2\eta_t} \left(\|\mathbf{x} - \mathbf{x}_t\|_2^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|_2^2 \right) + \eta_t G_1^2 + \eta_t \lambda_0^2 G_2^2 + \zeta_t(\mathbf{x}) - \frac{\beta}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2,\end{aligned}$$

where

$$p(\mathbf{x}) = \frac{\exp(\lambda_0 g(\mathbf{x})/\gamma)}{1 + \exp(\lambda_0 g(\mathbf{x})/\gamma)}.$$

Taking summation of above inequality over $t = 1, \cdots, T$ gives

$$\begin{aligned}\sum_{t=1}^T F(\mathbf{x}_t) - F(\mathbf{x}) &\leq \sum_{t=1}^T \frac{1}{2} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\beta}{2} \right) \|\mathbf{x} - \mathbf{x}_t\|_2^2 \\ &\quad + \sum_{t=1}^T \eta_t (G_1^2 + \lambda_0^2 G_2^2) + \sum_{t=1}^T \zeta_t(\mathbf{x}) - \frac{\beta}{4} \sum_{t=1}^T \|\mathbf{x} - \mathbf{x}_t\|_2^2.\end{aligned}$$

Since $\eta_t = 1/(2\beta t)$, we have

$$\sum_{t=1}^{T}(F(\mathbf{x}_t) - F(\mathbf{x})) \leq \frac{(G_1^2 + \lambda_0^2 G_2^2)(1 + \ln T)}{2\beta} + \sum_{t=1}^{T}\zeta_t(\mathbf{x}) - \frac{\beta}{4}\sum_{t=1}^{T}\|\mathbf{x} - \mathbf{x}_t\|_2^2$$

We complete the proof by letting $\mathbf{x} = \mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$.

D Proof of Lemma 4

The proof is based on the Berstein inequality for martingales [1] which is restated here for completeness.

Theorem 1. *(Bernsteins inequality for martingales). Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to the filtration $\mathcal{F} = (\mathcal{F}_i)_{1\leq i\leq n}$ and with $\|X_i\| \leq K$. Let*

$$S_i = \sum_{j=1}^{i} X_j$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_n^2 = \sum_{t=1}^{n} \mathrm{E}\left[X_t^2 | \mathcal{F}_{t-1}\right],$$

Then for all constants t, $\nu > 0$,

$$\Pr\left[\max_{i=1,\ldots,n} S_i > t \text{ and } \Sigma_n^2 \leq \nu\right] \leq \exp\left(-\frac{t^2}{2(\nu + Kt/3)}\right),$$

and therefore,

$$\Pr\left[\max_{i=1,\ldots,n} S_i > \sqrt{2\nu t} + \frac{\sqrt{2}}{3}Kt \text{ and } \Sigma_n^2 \leq \nu\right] \leq e^{-t}.$$

Proof of Lemma 4. Define martingale difference $X_t = (\mathbf{x} - \mathbf{x}_t)^\top(\nabla f(\mathbf{x}_t) - \widetilde{\nabla} f(\mathbf{x}_t, \xi_t))$ and martingale $\Lambda_T = \sum_{t=1}^{T} X_t$. Define the conditional variance Σ_T^2 as

$$\Sigma_T^2 = \sum_{t=1}^{T} \mathrm{E}_{\xi_t}\left[X_t^2\right] \leq 4G_1^2 \sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}\|_2^2 = 4G_1^2 D_T.$$

Define $K = 4G_1$. We have

$$\begin{aligned}
&\Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3\right) \\
&= \Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 D_T\right) \\
&= \Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 D_T, D_T \leq \frac{4}{T}\right) \\
&+ \sum_{i=1}^{m} \Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 D_T, \frac{4}{T}2^{i-1} < D_T \leq \frac{4}{T}2^i\right) \\
&\leq \Pr\left(D_T \leq \frac{4}{T}\right) + \sum_{i=1}^{m} \Pr\left(\Lambda_T \geq \sqrt{2 \times 4G_1^2 \frac{4}{T}2^i \tau} + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 \frac{4}{T}2^i\right) \\
&\leq \Pr\left(D_T \leq \frac{4}{T}\right) + me^{-\tau}.
\end{aligned}$$

where we use the fact $\|\mathbf{x}_t - \mathbf{x}\|_2^2 \leq 4$ for any $\mathbf{x} \in \mathcal{B}$, and the last step follows the Bernstein inequality for martingales. We complete the proof by setting $\tau = \ln(m/\delta)$.

□

References

[1] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240, 2003.