**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 \( x \in B \),

\[
\|x_{t+1} - x\|^2 - \|x_t - x\|^2 \leq \|x'_{t+1} - x\|^2 - \|x_t - x\|^2
\]

\[
= \|x_t - \eta_t(\nabla f(x_t, \xi_t) + \lambda_t g(x_t)) - x\|^2 - \|x_t - x\|^2
\]

\[
\leq \eta_t^2 \|\nabla f(x_t, \xi_t) + \lambda_t g(x_t)\|^2 - 2\eta_t (x_t - x)^T (\nabla f(x_t, \xi_t) + \lambda_t g(x_t))
\]

\[
= \eta_t^2 \|\nabla f(x_t, \xi_t) + \lambda_t g(x_t)\|^2 - 2\eta_t (x_t - x)^T (\nabla f(x_t, \xi_t) + \lambda_t g(x_t)) + 2\eta_t (x_t - x)^T (\nabla f(x_t, \xi_t) - \nabla f(x_t))
\]

\[
\equiv \delta_t L(x_t, \lambda_t)
\]

Then we have

\[
(x_t - x)^T \nabla_x L(x_t, \lambda_t) \leq \frac{1}{2\eta_t} \left( \|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right) + \frac{\eta_t}{2} \|\nabla f(x_t, \xi_t) + \lambda_t g(x_t)\|^2 + \zeta_t(x)
\]

\[
\leq \frac{1}{2\eta_t} \left( \|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right) + \eta_t \|\nabla f(x_t, \xi_t)\|^2 + \eta_t \lambda_t^2 \|g(x_t)\|^2 + \zeta_t(x)
\]

\[
\leq \frac{1}{2\eta_t} \left( \|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right) + 2\eta_t \|\nabla f(x_t, \xi_t)\|^2 + 2\eta_t \lambda_t^2 \|g(x_t)\|^2 + \zeta_t(x)
\]

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

\[
|\lambda_{t+1} - \lambda|^2 - |\lambda_t - \lambda|^2 \leq |\lambda_t + \eta_t (g(x_t) - g(\lambda_t))|^2 - |\lambda_t - \lambda|^2
\]

\[
\leq \eta_t^2 |g(x_t) - g(\lambda_t)|^2 + 2\eta_t (\lambda_t - \lambda) (g(x_t) - g(\lambda_t)).
\]

Then we have

\[
(\lambda - \lambda_t) \nabla_{\lambda} L(x_t, \lambda_t) \leq \frac{1}{2\eta_t} \left( \|\lambda - \lambda_t\|^2 - \|\lambda_{t+1} - \lambda_t\|^2 \right) + \frac{\eta_t}{2} |g(x_t) - g(\lambda_t)|^2.
\]

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

\[
(\lambda - \lambda_t) \nabla_{\lambda} L(x_t, \lambda_t) \leq \frac{1}{2\eta_t} \left( \|\lambda - \lambda_t\|^2 - \|\lambda_{t+1} - \lambda_t\|^2 \right) + 2\eta_t C_2^2.
\]
B Proof of Lemma 2

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

$$L(x, \lambda_t) - L(x_t, \lambda_t) \geq (x - x_t)^T \nabla_x L(x_t, \lambda_t),$$
$$L(x_t, \lambda) - L(x_t, \lambda_t) \leq (\lambda - \lambda_t) \nabla_\lambda L(x_t, \lambda_t).$$

Using the inequalities in Lemma 1, we have

$$L(x_t, \lambda_t) - L(x, \lambda_t) \leq \frac{1}{2\eta_t} (\|x - x_t\|^2 - \|x - x_{t+1}\|^2) + 2\eta_t G_1^2 + \eta_t G_2^2 \lambda_t^2 + 2\eta_t \Delta_t + \zeta_t(x),$$
$$L(x_t, \lambda) - L(x_t, \lambda_t) \leq \frac{1}{2\eta_t} (\|\lambda - \lambda_t\|^2 - \|\lambda - \lambda_{t+1}\|^2) + 2\eta_t C_2^2,$$

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

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

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

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

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

$$\sum_{t=1}^T (f(x_t) - f(x^*)) + \frac{\|\sum_{t=1}^T g(x_t)\|^2}{2(\gamma T + 2G_2^2)} \leq \frac{G_2^2}{\gamma} + \frac{(G_1 + C_2)}{G_2^2} \gamma T + \frac{\gamma}{G_2^2} \sum_{t=1}^T \Delta_t + \sum_{t=1}^T \zeta_t(x^*)$$

C Proof of Lemma 3

Since $F(x)$ is strongly convex in $x$, we have

$$F(x) - F(x_t) \geq (x - x_t)^T \nabla F(x_t) + \frac{\beta}{2} \|x - x_t\|^2.$$ 

Following the same analysis as in Lemma 1, we have

$$(x_t - x)^T \nabla F(x_t) \leq \frac{1}{2\eta_t} (\|x - x_t\|^2 - \|x - x_{t+1}\|^2) + \frac{\eta_t}{2} \|\nabla f(x_t, \xi_t) + p(x_t)\lambda_0 \nabla g(x_t)\|^2$$
$$+ \zeta_t(x) - \frac{\beta}{2} \|x - x_t\|^2$$
$$\leq \frac{1}{2\eta_t} (\|x - x_t\|^2 - \|x - x_{t+1}\|^2) + \eta_t G_1^2 + \eta_t \lambda_0^2 G_2^2 + \zeta_t(x) - \frac{\beta}{2} \|x - x_t\|^2,$$

where

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

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

$$\sum_{t=1}^T F(x_t) - F(x) \leq \sum_{t=1}^T \frac{1}{2} \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\beta}{2} \right) \|x - x_t\|^2$$
$$+ \sum_{t=1}^T \eta_t (G_1^2 + \lambda_0^2 G_2^2) + \sum_{t=1}^T \zeta_t(x) - \frac{\beta}{4} \sum_{t=1}^T \|x - x_t\|^2.$$
Since $\eta_t = 1/(2\beta t)$, we have
\[
\sum_{t=1}^T (F(x_t) - F(x)) \leq \frac{(G_1^2 + \lambda_2^2 G_2^2)(1 + \ln T)}{2\beta} + \sum_{t=1}^T \zeta_t(x) - \frac{\beta}{4} \sum_{t=1}^T \|x - x_t\|^2
\]
We complete the proof by letting $x = x^* = \arg\min_{x \in \mathcal{K}} f(x)$.

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 $\mathcal{F} = (\mathcal{F}_t)_{1 \leq i \leq n}$ 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
\[
\Sigma_n^2 = \sum_{i=1}^n \mathbb{E} [X_i^2 | \mathcal{F}_{i-1}],
\]
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,
\[
\Pr \left[ \max_{i=1, \ldots, n} S_i > \sqrt{2\nu t} + \sqrt{\frac{2}{3}} Kt \text{ and } \Sigma_n^2 \leq \nu \right] \leq e^{-t}.
\]

**Proof of Lemma 4.** Define martingale difference $X_t = (x - x_t)^T (\nabla f(x_t) - \nabla f(x_t, \xi_t))$ and martingale $\Lambda_T = \sum_{t=1}^T X_t$. Define the conditional variance $\Sigma_T^2$ as
\[
\Sigma_T^2 = \sum_{t=1}^T \mathbb{E} \left[ X_t^2 \right] \leq 4G_1^2 \sum_{t=1}^T \|x_t - x\|^2 = 4G_1^2 D_T.
\]
Define $K = 4G_1$. We have
\[
\Pr \left( \Lambda_T \geq 2 \sqrt{4G_1^2 D_T \tau} + \sqrt{2K \tau/3} \right)
\]
\[
\begin{align*}
&= \Pr \left( \Lambda_T \geq 2 \sqrt{4G_1^2 D_T \tau} + \sqrt{2K \tau/3}, \Sigma_T^2 \leq 4G_1^2 D_T \right) \\
&= \Pr \left( \Lambda_T \geq 2 \sqrt{4G_1^2 D_T \tau} + \sqrt{2K \tau/3}, \Sigma_T^2 \leq 4G_1^2 D_T, D_T \leq \frac{4}{T} \right) \\
&+ \sum_{i=1}^m \Pr \left( \Lambda_T \geq 2 \sqrt{4G_1^2 D_T \tau} + \sqrt{2K \tau/3}, \Sigma_T^2 \leq 4G_1^2 D_T, \frac{4}{T} 2^{i-1} < D_T \leq \frac{4}{T} 2^i \right) \\
&\leq \Pr \left( D_T \leq \frac{4}{T} \right) + \sum_{i=1}^m \Pr \left( \Lambda_T \geq \sqrt{2 \times 4G_1^2 \frac{4}{T} 2^i \tau} + \sqrt{2K \tau/3}, \Sigma_T^2 \leq 4G_1^2 \frac{4}{T} 2^i \right) \\
&\leq \Pr \left( D_T \leq \frac{4}{T} \right) + me^{-\tau}.
\end{align*}
\]
where we use the fact $\|x_t - x\|^2 \leq 4$ for any $x \in \mathcal{B}$, and the last step follows the Bernstein inequality for martingales. We complete the proof by setting $\tau = \ln(m/\delta)$.
References