Supplementary Materials: Accelerated Stochastic Gradient Descent for Minimizing Finite Sums

Atsushi Nitanda
Tokyo Institute of Technology and NTT DATA Mathematical Systems Inc.
Tokyo, Japan
nitanda@msi.co.jp

1 Proof of the Proposition 1

We now prove the Proposition 1 that gives the condition of compactness of sublevel set.

Proof. Let $B^d(r)$ and $S^{d-1}(r)$ denote the ball and sphere of radius r, centered at the origin. By affine transformation, we can assume that X_* contains the origin O, $X_* \subset B^d(1)$, and $X_* \cap S^{d-1}(1) = \phi$. Then, we have that for $\forall x \in S^{d-1}(1)$,

$$(\nabla f(x), x) \geq f(x) - f(O) > 0,$$

where we use convexity for the first inequality and $O \in X_* \wedge x \notin X_*$ for the second inequality. We denote the minimum value of $(\nabla f(x), x)$ on $S^{d-1}(1)$ by α. Since $(\nabla f(x), x)$ is positive continuous, we have $\alpha > 0$. For $\forall r \geq 1$ and $\forall x \in S^{d-1}(r)$, we set $\hat{x} = x/r \in S^{d-1}(1)$, then it follows that

$$\begin{aligned} f(x) &\geq f(\hat{x}) + (\nabla f(\hat{x}), x - \hat{x}) \\ &\geq f(\hat{x}) + (r-1)(\nabla f(\hat{x}), \hat{x}) \\ &\geq f_* + (r-1)\alpha \end{aligned}$$

This inequality implies that if $r > 1 + \frac{c - f_*}{\alpha}$, then we have $f(x) > c$ for $\forall x \in S^{d-1}(r)$. Therefore, sublevel set $\{x \in \mathbb{R}^d; f(x) \leq c\}$ is a closed bounded set. □

2 Proof of the Lemma 1

To prove Lemma 1, the following lemma is required, which is also shown in [1].

Lemma A. *Let $\{\xi_i\}_{i=1}^n$ be a set of vectors in $\mathbb{R}^d$ and μ denote an average of $\{\xi_i\}_{i=1}^n$. Let I denote a uniform random variable representing a size b subset of $\{1, 2, \ldots, n\}$. Then, it follows that,*

$$\mathbb{E}_I \left\| \frac{1}{b} \sum_{i \in I} \xi_i - \mu \right\|^2 = \frac{n-b}{b(n-1)} \mathbb{E}_i \|\xi_i - \mu\|^2.$$

Proof. We denote a size b subset of $\{1, 2, \ldots, n\}$ by $S = \{i_1, \ldots, i_b\}$ and denote $\xi_i - \mu$ by $\tilde{\xi}_i$. Then,

$$\begin{aligned} \mathbb{E}_I \left\| \frac{1}{b} \sum_{i \in I} \xi_i - \mu \right\|^2 &= \frac{1}{C(n,b)} \sum_S \left\| \frac{1}{b} \sum_{j=1}^b \xi_{i_j} - \mu \right\|^2 \\ &= \frac{1}{b^2 C(n,b)} \sum_S \left\| \sum_{j=1}^b \tilde{\xi}_{i_j} \right\|^2 \\ &= \frac{1}{b^2 C(n,b)} \sum_S \left(\sum_{j=1}^b \|\tilde{\xi}_{i_j}\|^2 + 2 \sum_{j,k,j<k} \tilde{\xi}_{i_j}^T \tilde{\xi}_{i_k} \right), \end{aligned}$$

where $C(\cdot,\cdot)$ is a combination. By symmetry, an each $\tilde{\xi}_i$ appears $\frac{bC(n,b)}{n}$ times and an each pair $\tilde{\xi}_i^T\tilde{\xi}_j$ for $i<j$ appears $\frac{C(b,2)C(n,b)}{C(n,2)}$ times in $\sum_S$. Therefore, we have

$$\begin{aligned}\mathbb{E}_I\left\|\frac{1}{b}\sum_{i\in I}\xi_i-\mu\right\|^2 &= \frac{1}{b^2C(n,b)}\left(\frac{bC(n,b)}{n}\sum_{i=1}^n\|\tilde{\xi}_i\|^2+\frac{2C(b,2)C(n,b)}{C(n,2)}\sum_{i,j,i<j}\tilde{\xi}_i^T\tilde{\xi}_j\right)\\ &= \frac{1}{bn}\sum_{i=1}^n\|\tilde{\xi}_i\|^2+\frac{2(b-1)}{bn(n-1)}\sum_{i,j,i<j}\tilde{\xi}_i^T\tilde{\xi}_j.\end{aligned}$$

Since, $0=\|\sum_{i=1}^n\tilde{\xi}_i\|^2=\sum_{i=1}^n\|\tilde{\xi}_i\|^2+2\sum_{i,j,i<j}\tilde{\xi}_i^T\tilde{\xi}_j$, we have

$$\mathbb{E}_I\left\|\frac{1}{b}\sum_{i\in I}\xi_i-\mu\right\|^2=\left(\frac{1}{bn}-\frac{b-1}{bn(n-1)}\right)\sum_{i=1}^n\|\tilde{\xi}_i\|^2=\frac{n-b}{b(n-1)}\frac{1}{n}\sum_{i=1}^n\|\tilde{\xi}_i\|^2.$$

This finishes the proof of Lemma. □

We now prove the Lemma 1.

Proof of Lemma 1 . We set $v_j^1=\nabla f_j(x_k)-\nabla f_j(\tilde{x})+\tilde{v}$. Using Lemma A and

$$v_k=\frac{1}{b}\sum_{j\in I_k}v_j^1,$$

conditional variance of v_k is as follows

$$\mathbb{E}_{I_k}\|v_k-\nabla f(x_k)\|^2=\frac{1}{b}\frac{n-b}{n-1}\mathbb{E}_j\|v_j^1-\nabla f(x_k)\|^2,$$

where expectation in right hand side is taken with respect to $j\in\{1,\dots,n\}$. By Corollary 3 in [2], it follows that,

$$\mathbb{E}_j\|v_j^1-\nabla f(x_k)\|^2\le 4L(f(x_k)-f(x_*)+f(\tilde{x})-f(x_*)).$$

This completes the proof of Lemma 1. □

3 Stochastic gradient descent analysis

Below is the proof of Lemma 3.

Proof of Lemma 3 . It is clear that y_k is equal to $x_k-\eta v_k$. Since $f(x)$ is L-smooth and $\eta=\frac{1}{L}$, we have,

$$\begin{aligned}f(y_k) &\le f(x_k)+(\nabla f(x_k),y_k-x_k)+\frac{L}{2}\|y_k-x_k\|^2\\ &= f(x_k)-\frac{1}{L}(\nabla f(x_k),v_k)+\frac{1}{2L}\|v_k\|^2.\end{aligned}$$

v_k is an unbiased estimator of gradient $\nabla f(x_k)$, that is, $\mathbb{E}_{I_k}[v_k]=\nabla f(x_k)$. Hence, we have

$$\mathbb{E}_{I_k}\|v_k\|^2=\|\nabla f(x_k)\|^2+\mathbb{E}_{I_k}\|v_k-\nabla f(x_k)\|^2.$$

Using above two expressions, we get

$$\begin{aligned}\mathbb{E}_{I_k}[f(y_k)] &= f(x_k)-\frac{1}{L}\|\nabla f(x_k)\|^2+\frac{1}{2L}\mathbb{E}_{I_k}\|v_k\|^2\\ &= f(x_k)-\frac{1}{2L}\|\nabla f(x_k)\|^2+\frac{1}{2L}\mathbb{E}_{I_k}\|v_k-\nabla f(x_k)\|^2.\end{aligned}$$

□

4 Stochastic mirror descent analysis

We give the proof of Lemma 4.

Proof of Lemma 4 . The following are basic properties of Bregman divergence.

$$\langle \nabla V_x(y), u - y\rangle = V_x(u) - V_y(u) - V_x(y), \tag{1}$$

$$V_x(y) \geq \frac{1}{2}\|x - y\|^2. \tag{2}$$

Using (1) and (2), we have

$$\begin{aligned}
\alpha_k\langle v_k, z_{k-1} - u\rangle &= \alpha_k\langle v_k, z_{k-1} - z_k\rangle + \alpha_k\langle v_k, z_k - u\rangle \\
&= \alpha_k\langle v_k, z_{k-1} - z_k\rangle - \langle \nabla V_{z_{k-1}}(z_k), z_k - u\rangle \\
&\underset{(1)}{=} \alpha_k\langle v_k, z_{k-1} - z_k\rangle + V_{z_{k-1}}(u) - V_{z_k}(u) - V_{z_{k-1}}(z_k) \\
&\underset{(2)}{\leq} \alpha_k\langle v_k, z_{k-1} - z_k\rangle - \frac{1}{2}\|z_{k-1} - z_k\|^2 + V_{z_{k-1}}(u) - V_{z_k}(u) \\
&\leq \frac{1}{2}\alpha_k^2\|v_k\|^2 + V_{z_{k-1}}(u) - V_{z_k}(u),
\end{aligned}$$

where for the second equality we use stochastic mirror descent step, that is, $\alpha_k v_k + \nabla V_{z_{k-1}}(z_k) = 0$ and for the last inequality we use the Fenchel-Young inequality $\alpha_k\langle v_k, z_{k-1} - z_k\rangle \leq \frac{1}{2}\alpha_k^2\|v_k\|^2 + \frac{1}{2}\|z_{k-1} - z_k\|^2$.

By taking expectation with respect to I_k and using $\mathbb{E}_{I_k}\|v_k\|^2 = \|\nabla f(x_k)\|^2 + \mathbb{E}_{I_k}\|v_k - \nabla f(x_k)\|^2$, we have

$$\alpha_k\langle \nabla f(x_k), z_{k-1} - u\rangle \leq V_{z_{k-1}}(u) - \mathbb{E}_{I_k}[V_{z_k}(u)] + \frac{1}{2}\alpha_k^2\|\nabla f(x_k)\|^2 + \frac{1}{2}\alpha_k^2\mathbb{E}_{I_k}\|v_k - \nabla f(x_k)\|^2.$$

This finishes the proof of Lemma 4. □

5 Proof of the Lemma 2

We now prove the Lemma 2 that is the key to the analysis of our method.

Proof. We denote $V_{z_k}(x_*)$ by V_k for simplicity. We get

$$\begin{aligned}
&\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - x_*\rangle \\
&\leq V_k - \mathbb{E}_{I_{k+1}}[V_{k+1}] + L\alpha_{k+1}^2(f(x_{k+1}) - \mathbb{E}_{I_{k+1}}[f(y_{k+1})]) + \alpha_{k+1}^2\mathbb{E}_{I_{k+1}}\|v_{k+1} - \nabla f(x_{k+1})\|^2 \\
&\leq V_k - \mathbb{E}_{I_{k+1}}[V_{k+1}] + L\alpha_{k+1}^2(f(x_{k+1}) - \mathbb{E}_{I_{k+1}}[f(y_{k+1})]) \\
&\quad + 4L\alpha_{k+1}^2\delta_{k+1}(f(x_{k+1}) - f(x_*) + f(y_0) - f(x_*)) \\
&= V_k - \mathbb{E}_{I_{k+1}}[V_{k+1}] + (1 + 4\delta_{k+1})L\alpha_{k+1}^2(f(x_{k+1}) - f(x_*)) - L\alpha_{k+1}^2\mathbb{E}_{I_{k+1}}[f(y_{k+1}) - f(x_*)] \\
&\quad + 4L\alpha_{k+1}^2\delta_{k+1}(f(y_0) - f(x_*)),
\end{aligned}$$

where for the first inequality we use Lemma 3 and 4 with $u = x_*$, for the second inequality we use Lemma 1.

By taking the expectation with respect to the history of random variables $I_1, I_2 \ldots$, we have,

$$\begin{aligned}
\alpha_{k+1}\mathbb{E}[\langle \nabla f(x_{k+1}), z_k - x_*\rangle] \leq\ & \mathbb{E}[V_k - V_{k+1}] + (1 + 4\delta_{k+1})L\alpha_{k+1}^2\mathbb{E}[f(x_{k+1}) - f(x_*)] \\
& - L\alpha_{k+1}^2\mathbb{E}[f(y_{k+1}) - f(x_*)] + 4L\alpha_{k+1}^2\delta_{k+1}(f(y_0) - f(x_*)),
\end{aligned} \tag{3}$$

and we get

$$
\begin{aligned}
\sum_{k=0}^{m} \alpha_{k+1}\mathbb{E}[f(x_{k+1}) - f(x_*)] &\leq \sum_{k=0}^{m} \alpha_{k+1}\mathbb{E}[(\nabla f(x_{k+1}), x_{k+1} - x_*)] \\
&= \sum_{k=0}^{m} \alpha_{k+1}(\mathbb{E}[(\nabla f(x_{k+1}), x_{k+1} - z_k)] + \mathbb{E}[(\nabla f(x_{k+1}), z_k - x_*)]) \\
&= \sum_{k=0}^{m} \alpha_{k+1}\left(\frac{1-\tau_k}{\tau_k}\mathbb{E}[(\nabla f(x_{k+1}), y_k - x_{k+1})] + \mathbb{E}[(\nabla f(x_{k+1}), z_k - x_*)]\right) \\
&\leq \sum_{k=0}^{m} \left(\alpha_{k+1}\frac{1-\tau_k}{\tau_k}\mathbb{E}[f(y_k) - f(x_{k+1})] + \alpha_{k+1}\mathbb{E}[(\nabla f(x_{k+1}), z_k - x_*)]\right). \quad (4)
\end{aligned}
$$

Using (3), (4), and $V_{z_{k+1}}(x_*) \geq 0$, we have

$$
\begin{aligned}
&\sum_{k=0}^{m} \alpha_{k+1}\left(1 + \frac{1-\tau_k}{\tau_k} - (1+4\delta_{k+1})L\alpha_{k+1}\right)\mathbb{E}[f(x_{k+1}) - f(x_*)] \\
&\leq V_0 + \sum_{k=0}^{m} \alpha_{k+1}\frac{1-\tau_k}{\tau_k}\mathbb{E}[f(y_k) - f(x_*)] - L\sum_{k=0}^{m} \alpha_{k+1}^2\mathbb{E}[f(y_{k+1}) - f(x_*)] \\
&+4L\sum_{k=0}^{m} \alpha_{k+1}^2\delta_{k+1}(f(y_0) - f(x_*)).
\end{aligned}
$$

This completes the proof of Lemma 2. □

6 Modified AMSVRG for general convex problems

We now introduce a modified AMSVRG (described in Figure 1) that does not need the boundedness assumption for general convex problems. We set η, α_{k+1}, and τ_k as in (5). Let $b_{k+1} \in \mathbb{Z}_+$ be the

Algorithm 3$(w_0,\ (m_s)_{s\in\mathbb{Z}_+},\ \eta,\ (\alpha_{k+1})_{k\in\mathbb{Z}_+},\ (b_{k+1})_{k\in\mathbb{Z}_+},\ (\tau_k)_{k\in\mathbb{Z}_+})$

for $s \leftarrow 0,\ 1, \ldots$
  $y_0 \leftarrow w_s,\ \ z_0 \leftarrow w_0$
  $w_{s+1} \leftarrow$ **Algorithm1**$(y_0,\ z_0,\ m_s,\ \eta,\ (\alpha_{k+1})_{k\in\mathbb{Z}_+},\ (b_{k+1})_{k\in\mathbb{Z}_+},\ (\tau_k)_{k\in\mathbb{Z}_+})$
end

Figure 1: Modified AMSVRG

minimum values satisfying $4L\delta_{k+1}\alpha_{k+1} \leq p$ for small p $(e.g.\ 1/4)$. Let $m_s = \left\lceil 4\sqrt{\frac{LV_{z_0}(x_*)}{\epsilon}} \right\rceil$. From Theorem 1, we get

$$
\mathbb{E}[f(w_{s+1}) - f(x_*)] \leq \epsilon + a(f(w_s) - f(x_*)),
$$

where $a = \frac{5}{2}p$. Thus, it follows that,

$$
\begin{aligned}
\mathbb{E}[f(w_{s+1}) - f(x_*)] &\leq \sum_{t=0}^{s} a^t\epsilon + a^{s+1}(f(w_0) - f(x_*)) \\
&\leq \frac{1}{1-a}\epsilon + a^{s+1}(f(w_0) - f(x_*)).
\end{aligned}
$$

Hence, running the modified AMSVRG for $O\left(\log \frac{1}{\epsilon}\right)$ outer iterations achieves ϵ-accurate solution in expectation, and a complexity at each stage is

$$O\left(n + \sum_{k=0}^{m_s} b_{k+1}\right) \leq O\left(n + \frac{nm_s^2}{n+m_s}\right)$$
$$= O\left(n + \frac{nL}{\epsilon n + \sqrt{\epsilon L}}\right) = O\left(n + \min\left\{\frac{L}{\epsilon}, n\sqrt{\frac{L}{\epsilon}}\right\}\right),$$

where we used the monotonicity of b_{k+1} with respect to k for the first inequality. Note that $V_{z_0}(x_*)$ is constant (i.e. $V_{w_0}(x_*)$), and O hides this term. From the above analysis, we derive the following theorem.

Theorem 1. *Consider the modified AMSVRG under Assumptions 1. Let parameters be as above. Then the overall complexity for obtaining ϵ-accurate solution in expectation is*

$$O\left(\left(n + \min\left\{\frac{L}{\epsilon}, n\sqrt{\frac{L}{\epsilon}}\right\}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

References

[1] J. E. Freund. *Mathematical Statistics*. prentice Hall, 1962.

[2] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *arXiv:1403.4699*, 2014.