CS331-HW10-Lukang-Sun

November 12, 2021

p1. (see Figure 1.) a = [matrix([[0.1]]), matrix([[0.424466]]), matrix([[0.77981303]]), matrix([[0.20033184]]), matrix([[0.51116473]]), matrix([[0.2604399]]), matrix([[0.97100656]]), matrix([[0.21263449]]), matrix([[0.26417151]]), matrix([[0.15995097]])]

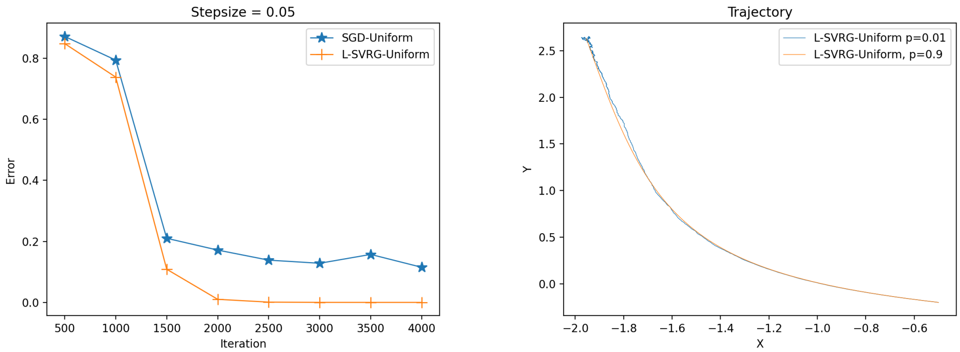
b = [matrix([[0.4231786]]), matrix([[0.524466]]), matrix([[0.17981303]]), matrix([[0.50033184]]), matrix([[0.71116473]]), matrix([[0.0604399]]), matrix([[0.37100656]]), matrix([[0.91263449]]), matrix([[0.66417151]]), matrix([[0.65995097]])] ,$f = \frac{1}{10}\sum_{i=1}^{10} f_i(x,y), f_i(x,y) = \sin(x+a[i]) + \cos(y+b[i])$, for the SGD method, I use SGD-US and L-SVRG-US. Initial point is init = matrix([[-0.5],[-0.2]]).
(1)L-SVRG-US will converge to the optimal point, while SGD-US will only converge to a neighborhood of the optimal point, in graph (a), we can see the results verifies prediction.
(2)in L-SVRG-US, p is bigger the trajectories will be more smooth, while p is smaller, the trajectories will fluctuate, in graph (b), we can see the results verifies prediction.

p2.

Proof. The proof is almost the same as the proof of theorem 121. Since f is L-smooth, we have

$$\begin{aligned} f\left(x^{k+1}\right) - f^{\inf} &\leq f\left(x^k\right) - f^{\inf} + \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &= f\left(x^k\right) - f^{\inf} - \gamma\left\langle \nabla f\left(x^k\right), g^k \right\rangle + \frac{L\gamma^2}{2}\left\|g^k\right\|^2 \end{aligned}$$

By applying expectation to both sides and subsequently using unbiasedness of g^k and the assumed bound on the second moment of the stochastic gradient,



(a) (b)

Figure 1: (a) shows $\mathrm{E}\left[||\nabla f||^2\right]$ changes in terms of iteration number K when set step size $\gamma = 0.05$,(b) shows the trajectories of L-SVRG-US with different p.

we get

$$\begin{aligned}
&\mathrm{E}\left[f\left(x^{k+1}\right)-f^{\inf}\right] \leq \mathrm{E}\left[f\left(x^{k}\right)-f^{\inf}\right]-\gamma\,\mathrm{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^{2}\right]+\frac{L\gamma^{2}}{2}\mathrm{E}\left[\left\|g^{k}\right\|^{2}\right] \\
&\leq \mathrm{E}\left[f\left(x^{k}\right)-f^{\inf}-\gamma\left\|\nabla f\left(x^{k}\right)\right\|^{2}+\frac{L\gamma^{2}}{2}\left[2A\left(f\left(x^{k}\right)-f^{\inf}\right)+B_{1}\sigma^{k}+B_{2}\left\|\nabla f\left(x^{k}\right)\right\|^{2}+C\right]\right] \\
&= \mathrm{E}\left[\left(1+LA\gamma^{2}\right)\left(f\left(x^{k}\right)-f^{\inf}\right)+\frac{LB_{1}\gamma^{2}}{2}\sigma^{k}-\left(\gamma-\frac{LB_{2}\gamma^{2}}{2}\right)\left\|\nabla f\left(x^{k}\right)\right\|_{2}^{2}+\frac{LC\gamma^{2}}{2}\right]
\end{aligned}$$

Choose any $M > 0$ and define

$$\Delta^{k+1} \stackrel{\text{def}}{=} f\left(x^{k+1}\right)-f^{\inf}+M\gamma^{2}\sigma^{k+1}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta^{k+1}\right] \leq \mathrm{E}\left[\left(1+LA\gamma^{2}+2M\tilde{A}\gamma^{2}\right)\left(f\left(x^{k}\right)-f^{inf}\right)+\left(\frac{LB_{1}}{2}+M\tilde{B}_{1}\right)\gamma^{2}\sigma^{k}\right] \\
&-\mathrm{E}\left[\left(\gamma-\frac{LB_{2}\gamma^{2}}{2}-M\tilde{B}_{2}\gamma^{2}\right)\left\|\nabla f\left(x^{k}\right)\right\|^{2}+\frac{LC\gamma^{2}}{2}+M\tilde{C}\gamma^{2}\right] \\
&= \mathrm{E}\left[a\left[f\left(x^{k}\right)-f^{\inf}+\frac{\frac{LB_{1}}{2}+M\tilde{B}_{1}}{a}\gamma^{2}\sigma^{k}\right]-b\left\|\nabla f\left(x^{k}\right)\right\|^{2}+c\right]
\end{aligned}$$

where

$$\begin{aligned} a &\stackrel{\text{def}}{=} 1 + LA\gamma^2 + 2M\tilde{A}\gamma^2 \\ b &\stackrel{\text{def}}{=} \gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2 \\ c &\stackrel{\text{def}}{=} \frac{LC\gamma^2}{2} + M\tilde{C}\gamma^2 \end{aligned}$$

In order to turn (1) into a recursion which has Δ^k on the right hand side, we need to make sure that

$$\frac{\frac{LB_1}{2} + M\tilde{B}_1}{a} \leq M$$

Fortunately, it is easy to see (prove this!) that we can make sure this holds by an appropriate choice of M. In particular, the last inequality holds if we choose

$$M \stackrel{\text{def}}{=} \frac{LB_1}{2\left(1 - \tilde{B}_1\right)} = \frac{L\theta}{2}$$

With this choice of M, we can continue from (314) and obtain the recursion

$$\mathrm{E}\left[\Delta^{k+1}\right] \quad \leq \mathrm{E}\left[a\Delta^k - b\left\|\nabla f\left(x^k\right)\right\|^2\right] + c \tag{1}$$

By applying expectation to both sides of this, and using the tower property of expectation, we get the recursion

$$\begin{aligned} \mathrm{E}\left[\Delta^{k+1}\right] &= \mathrm{E}\left[\mathrm{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right]\right] \\ &\leq \quad a\mathrm{E}\left[\Delta^k\right] - b\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + c \end{aligned} \tag{2}$$

We now apply Lemma 120 to recursion (2) with $X_k = \mathrm{E}\left[\Delta^k\right]$ and $Y_k = b\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]$. If we set $x = x^k$ with probability p_k (where p_k is as in Lemma 120), which means that $Y = Y_k$ with probability p_k, we conclude that

$$\begin{aligned} b\mathrm{E}\left[\|\nabla f(x)\|^2\right] &= \mathrm{E}[Y] \\ &\leq \frac{a^K}{S_K}\Delta^0 + c \\ &\leq \frac{a^K}{K}\Delta^0 + c \end{aligned} \tag{3}$$

where the last inequality follows since $a \geq 1$, which implies that $S_K \geq K$.

We now evaluate the expressions for b and c in (3). First,

$$\begin{aligned} b \quad &= \gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2 \\ &= \gamma - \frac{\gamma}{2}\left(LB_2\gamma + L\theta\tilde{B}_2\gamma\right) \\ &\geq \frac{\gamma}{2} \end{aligned} \tag{4}$$

where the last inequality holds by setting

$$\gamma \leq \frac{1}{L\left(B_2 + \theta\tilde{B}_2\right)}$$

Moreover,

$$c = \frac{LC}{2}\gamma^2 + M\tilde{C}\gamma^2 = \frac{L}{2}(C + \theta\tilde{C})\gamma^2 \tag{5}$$

By plugging the bound (4) on b and expression (5) for c into (3), we obtain the result.

$\square$

p3.

Proof. By lecture, we have

$$E\left[\|\nabla f(x)\|^2\right] \leq \frac{2\left(f\left(x^0\right) - f^{\inf}\right)}{K} \times \max\{\underbrace{\sqrt{\frac{4}{3}\frac{1-p}{p}\alpha(c+1)}}_{M_1}, \underbrace{L\left(B_2 + \frac{B_1\tilde{B}_2}{1-\tilde{B}_1}\right)}_{M_2}\}$$

$$M_2 \leq L\left(1 + \frac{3}{c+1} + \frac{4(1-p)}{pc}\right)$$

$$\alpha \stackrel{\text{def}}{=} \frac{(n-\tau)L_{\text{avg}}^2}{(n-1)\tau},$$

if we let $p = 0.5, c = 1$, then $M_1 = \sqrt{\frac{8}{3}}n^{-1/3}L_{avg}, M_2 \leq 6.5L$, so $\max\{M_1, M_2\} \leq 6.5L$, when n is large. So we need $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ steps to make the error less than ϵ^2, each step's computation is $n^{2/3}$, so, the total complexity is $\mathcal{O}\left(n^{2/3}/\varepsilon^2\right)$. $\square$

p4.

Theorem. *Let Assumption 8 (L-smoothness and $f \geq f^{inf}$), PL-condition ($\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^\star))$) and Assumption 11 ($\sigma^k$ assumption for non-convex functions) be satisfied. Choose constant stepsize γ satisfying*

$$0 < \gamma \leq \min\left\{\frac{1}{\frac{LB_2}{2} + M\tilde{B}_2}, \frac{\mu}{LA + 2MA + \mu LB_2 + 2\mu M\tilde{B}_2}\right\}$$

Then

$$\mathrm{E}\left[\Delta^k\right] \leq (1-\mu\gamma)^k\Delta^0 + \frac{c}{\mu\gamma}$$

where $c \overset{def}{=} \frac{LC\gamma^2}{2} + M\tilde{C}\gamma^2, M = \frac{LB_1}{2\left(a-2\mu b-\tilde{B}_1\right)}, a \overset{def}{=} 1 + LA\gamma^2 + 2M\tilde{A}\gamma^2, b \overset{def}{=} \gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2, \Delta^k \overset{def}{=} f\left(x^k\right) - f^{\inf} + M\gamma^2\sigma^k,$

Proof. Since f is L-smooth, we have

$$\begin{aligned} f\left(x^{k+1}\right) - f^{\inf} &\leq f\left(x^k\right) - f^{\inf} + \left\langle\nabla f\left(x^k\right), x^{k+1} - x^k\right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &\overset{(295)}{=} f\left(x^k\right) - f^{\inf} - \gamma\left(\nabla f\left(x^k\right), g^k\right\rangle + \frac{L\gamma^2}{2}\left\|g^k\right\|^2. \end{aligned} \tag{6}$$

By applying expectation to both sides of (6) and subsequently using unbiasedness of g^k and the assumed bound (297) (from lecture)on the second moment of the stochastic gradient, we get

$$\begin{aligned} \mathrm{E}\left[f\left(x^{k+1}\right) - f^{\inf} \mid x^k, \xi^k\right] &\overset{(6)}{\leq} f\left(x^k\right) - f^{\inf} - \gamma\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{L\gamma^2}{2}\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, \xi^k\right] \\ &\overset{(297)}{\leq} f\left(x^k\right) - f^{\inf} - \gamma\left\|\nabla f\left(x^k\right)\right\|^2 \\ &\quad + \frac{L\gamma^2}{2}\left[2A\left(f\left(x^k\right) - f^{\inf}\right) + B_1\sigma^k + B_2\left\|\nabla f\left(x^k\right)\right\|^2 + C\right] \\ &= \left(1 + LA\gamma^2\right)\left(f\left(x^k\right) - f^{\inf}\right) + \frac{LB_1\gamma^2}{2}\sigma^k \\ &\quad - \left(\gamma - \frac{LB_2\gamma^2}{2}\right)\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{LC\gamma^2}{2} \end{aligned} \tag{7}$$

Choose any $M > 0$ and define

$$\Delta^{k+1} \overset{\text{def}}{=} f\left(x^{k+1}\right) - f^{\inf} + M\gamma^2\sigma^{k+1} \tag{8}$$

by combining inequality (7) with assumption (298), we get

$$
\begin{aligned}
E\left[\Delta^{k+1} \mid x^k, \xi^k\right] \overset{(306)+(298)}{\leq} & \underbrace{\left(1 + LA\gamma^2 + 2M\tilde{A}\gamma^2\right)}_{=a}\left(f\left(x^k\right) - f^{\inf}\right) + \left(\frac{LB_1}{2} + M\tilde{B}_1\right)\gamma^2\sigma^k \\
& - \underbrace{\left(\gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2\right)}_{=b}\left\|\nabla f\left(x^k\right)\right\|^2 + \underbrace{\frac{LC\gamma^2}{2} + M\tilde{C}\gamma^2}_{=c} \\
\overset{\text{PL-condition}}{\leq} & (a - 2\mu b)\left[f\left(x^k\right) - f^{\inf} + \frac{\frac{LB_1}{2} + M\tilde{B}_1}{a - 2\mu b}\gamma^2\sigma^k\right] + c
\end{aligned}
\tag{9}
$$

where

$$
\begin{aligned}
a &\overset{\text{def}}{=} 1 + LA\gamma^2 + 2M\tilde{A}\gamma^2 \\
b &\overset{\text{def}}{=} \quad \gamma - \tfrac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2 \\
c &\overset{\text{def}}{=} \tfrac{LC\gamma^2}{2} + M\tilde{C}\gamma^2,
\end{aligned}
$$

we also require $b \geq 0$, that is $\gamma \leq \frac{1}{\frac{LB_2}{2} + M\tilde{B}_2}$. In order to turn (314) into a recursion which has Δ^k on the right hand side, we need to make sure that

$$
\frac{\frac{LB_1}{2} + M\tilde{B}_1}{a - 2\mu b} \leq M
$$

We can choose $M = \frac{LB_1}{2(a - 2\mu b - \tilde{B}_1)}$. With this choice of M, we can continue from (10) and obtain the recursion

$$
\mathrm{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right] \leq (a - 2\mu b)\Delta^k + c \tag{10}
$$

By applying expectation to both sides of this, and using the tower property of expectation, we get the recursion

$$
\begin{aligned}
\mathrm{E}\left[\Delta^{k+1}\right] &= \mathrm{E}\left[\mathrm{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right]\right] \\
&(\leq (a - 2\mu b)\mathrm{E}\left[\Delta^k\right] + c
\end{aligned}
$$

Finally,

$$
\mathrm{E}\left[\Delta^k\right] \leq (a - 2\mu b)^k\Delta^0 + c\sum_{i=0}^{k}(a - 2\mu b)^i \tag{11}
$$

by choosing $\gamma \leq \frac{\mu}{LA + 2MA + \mu LB_2 + 2\mu M\tilde{B}_2}$, we will have $a - 2\mu b \leq 1 - \mu\gamma$ so from (11), we have

$$
\mathrm{E}\left[\Delta^k\right] \leq (1 - \mu\gamma)^k\Delta^0 + \frac{c}{\gamma\mu} \tag{12}
$$

□