CS331-HW11-Lukang-Sun

November 18, 2021

p1.

Proof. if $p = 1$, then $g^k = \nabla f(x^k)$, which is eactly the GD. if $\tau = n$, I will use induction to prove this: for $k = 1$, $g^k = \nabla f(x^k)$. if for $k < K$, this is true, we will prove for $k = K$, $g^K = \nabla f(x^K)$, then by induction, for all k, we have $g^k = \nabla f(x^k)$, which is eactly the GD method. for g^K it could be $\nabla f(x^K)$ or $g^{K-1} + \nabla f(x^K) - \nabla f(x^{K-1})$, since by assumption $g^{K-1} = \nabla f(x^{K-1})$, so $g^{K-1} + \nabla f(x^K) - \nabla f(x^{K-1}) = \nabla f(x^K)$. $\square$

p2.

Theorem. *Assume f is L-smooth, lower bounded by f^{inf} and suppose that Assumption 12 holds [10]. Assume $n > 1$, choose minibatch size $\tau \in \{1, 2, \ldots, n\}$, probability $p \in (0, 1]$ and stepsize*

$$0 < \gamma \le \min\{\frac{p}{2\mu}, \frac{1}{L + L_{\text{avg}}\sqrt{\frac{2(1-p)(n-\tau)}{p(n-1)\tau}}}\} \overset{def}{=} \gamma_{p,\tau}$$

Fix $K \ge 1$,then we have

$$E\left[f\left(x^K\right) - f^{inf} + m\|g^K - \nabla f(x^K)\|^2\right] \le (1 - \gamma\mu)^K(f(x^0) - f^{inf}),$$

where $m = \frac{\gamma}{2(p-\gamma\mu)}$.

Proof. A direct calculation now reveals that

$$G \overset{\text{def}}{=} \mathrm{E}\left[\|g^{k+1} - \nabla f\left(x^{k+1}\right)\|^2 \mid x^{k+1}, x^k, g^k, s^k\right]$$

$$\overset{(365)}{=} p\underbrace{\|\nabla f\left(x^{k+1}\right) - \nabla f\left(x^{k+1}\right)\|^2}_{=0} + (1-p)\left\|g^k + \frac{1}{\tau}\sum_{i\in S^k}\left(\nabla f_i\left(x^{k+1}\right) - \nabla f_i\left(x^k\right)\right) - \nabla f\left(x^{k+1}\right)\right\|^2$$

$$= (1-p)\|\underbrace{g^k - \nabla f\left(x^k\right)}_{X} + \frac{1}{\tau}\sum_{i\in S^k}\underbrace{\left(\nabla f_i\left(x^{k+1}\right) - \nabla f_i\left(x^k\right)\right)}_{a_i} - \underbrace{\left(\nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right)\right)}_{\bar{a} = \frac{1}{n}\sum_i a_i}\|^2$$

$$= (1-p)\|X\|^2 + 2(1-p)\left\langle X, \frac{1}{\tau}\sum_{i\in S^k} a_i - \bar{a}\right\rangle + (1-p)\left\|\frac{1}{\tau}\sum_{i\in S^k} a_i - \bar{a}\right\|^2.$$

Take full expectation, we have

$$\mathrm{E}\left[\|g^{k+1} - \nabla f(x^{k+1})\|^2\right] \leq = (1-p)\mathrm{E}\left[\|g^k - \nabla f\left(x^k\right)\|^2\right] + (1-p)\frac{n-\tau}{(n-1)\tau}L_{\text{avg}}^2\mathrm{E}\left[\|x^{k+1} - x^k\|^2\right]$$

$$(1)$$

Then by lemma 127, we have

$$E\left[f\left(x^{k+1}\right) - f^{\inf} + m\|g^{k+1} - \nabla f(x^{k+1})\|^2\right]$$

$$\leq \mathrm{E}\left[f\left(x^k\right) - f^{\inf}\right] - \frac{\gamma}{2}\mathrm{E}\left[\|\nabla f\left(x^k\right)\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathrm{E}\left[\|x^{k+1} - x^k\|^2\right] + \frac{\gamma}{2}\mathrm{E}\left[\|g^k - \nabla f\left(x^k\right)\|^2\right]$$

$$+ m\mathrm{E}\left[\|g^{k+1} - \nabla f(x^{k+1})\|^2\right]$$

$$\leq \mathrm{E}\left[f\left(x^k\right) - f^{\inf}\right] - \frac{\gamma}{2}\mathrm{E}\left[\|\nabla f\left(x^k\right)\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathrm{E}\left[\|x^{k+1} - x^k\|^2\right] + \frac{\gamma}{2}\mathrm{E}\left[\|g^k - \nabla f\left(x^k\right)\|^2\right]$$

$$+ m\left((1-p)\|g^k - \nabla f\left(x^k\right)\|^2 + (1-p)\frac{n-\tau}{(n-1)\tau}L_{\text{avg}}^2\|x^{k+1} - x^k\|^2\right)$$

$$= (1-\gamma\mu)E\left[(f\left(x^k\right) - f^{\inf}) + m\|g^k - \nabla f(x^k)\|^2\right]$$

$$- \underbrace{\left(\frac{1}{2\gamma} - \frac{L}{2} - m(1-p)\left(\frac{n-\tau}{(n-1)\tau}L_{\text{avg}}^2\right)\right)}_{A}\mathrm{E}\left[\|x^{k+1} - x^k\|^2\right]$$

$$\leq (1-\mu\gamma)E\left[f\left(x^k\right) - f^{\inf} + m\|g^k - \nabla f(x^k)\|^2\right]$$

$$(2)$$

where $m = \frac{\gamma}{2(p-\gamma\mu)}$, we choose $\gamma \leq \frac{p}{2\mu}$, then $p - \gamma\mu \geq \frac{p}{2}$, the last inequality is due to $\gamma \leq \frac{1}{L+L_{\text{avg}}\sqrt{\frac{21-p)(n-\tau)}{p(n-1)\tau}}}$ (due to lemma 128 and the fact $\frac{(1-p)(n-\tau)L_{\text{avg}}^2}{(p-\gamma\mu)(n-1)\tau} \leq \frac{2(1-p)(n-\tau)L_{\text{avg}}^2}{(n-1)\tau p}$ when $\gamma \leq \frac{p}{2\mu}$, we have $A \geq 0$.) Use (2) for K times, then we have

$$E\left[f\left(x^K\right) - f^{\inf} + m\|g^K - \nabla f(x^K)\|^2\right] \leq (1-\gamma\mu)^K(f(x^0) - f^{\inf})$$
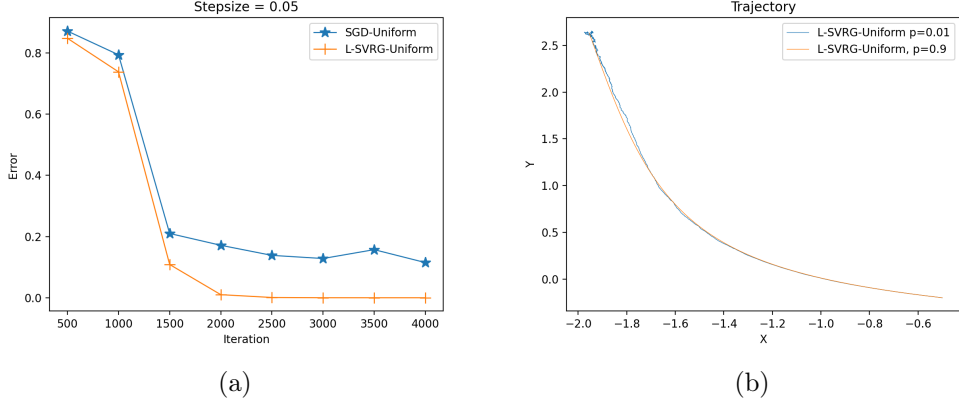
Figure 1: (a) shows $\mathrm{E}\left[||\nabla f||^2\right]$ changes in terms of iteration number K when set step size $\gamma = 0.05$,(b) shows the trajectories of L-SVRG-US with different p.

☐

p3. (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.