CS331-HW14-Lukang-Sun

December 10, 2021

p1. Recall the definition of the PAGE gradient estimator:

$$g^{k+1}=\begin{cases}\nabla f\left(x^{k+1}\right) & \text{with probability } p\\ g^k+\frac{1}{\tau}\sum_{i\in S^k}\left(\nabla f_i\left(x^{k+1}\right)-\nabla f_i\left(x^k\right)\right) & \text{with probability } 1-p,\end{cases}$$

in the $p-$ branch and $1-p-$branch, we need to compute the gradient, if τ is big, this will be costly. So the problem is to design new gradient estimator in the p and $1-p$ branch which is easy to compute. This is more of emprical nature.

p2.

$$g^{k+1}=\begin{cases}\nabla f\left(x^{k+1}\right) & \text{with probability} \quad p\\ g^k+H_{\xi^k}(x^{k+1},x^k) & \text{with probability} \quad 1-p,\end{cases}.$$

and make assumption $\mathrm{E}\left[||H_{\xi^k}(x^{k+1},x^k)||^2\right]\le C^2||x-y||^2$, for any $x,y\in\mathbb{R}^d$.

Theorem 1. *Assume f is L-smooth, lower bounded by f^{inf} and suppose that Assumption* **??** *holds. Choose probability $p\in(0,1]$, and stepsize*

$$0<\gamma\le\max_{\tau\in(0,\frac{p}{1-p})}\left\{\gamma_{p,\tau}\overset{def}{=}\frac{1}{L+\sqrt{\frac{2(1-p)(1+\tau)(L^2+C^2)}{\tau(p-\tau+p\tau)}}}\right\},$$

where the restriction $\tau\le\frac{p}{1-p}$ is designed to ensure that $p-\tau+p\tau>0$. Fix $K\ge 1$ and let $\hat{x}^K$ be chosen from the iterates $x^0,x^1,\ldots,x^{K-1}$ of PAGE *uniformly at random. Then*

$$\mathbb{E}\left[\left\|\nabla f\left(\hat{x}^K\right)\right\|^2\right]\le\frac{2\left(f\left(x^0\right)-f^{\inf}\right)}{\gamma K}.$$

Proof. Letting $G^k=\mathbb{E}\left[\left\|g^k-\nabla f\left(x^k\right)\right\|^2\right]$, $F^k=\mathbb{E}\left[f\left(x^k\right)-f^{\inf}\right]$ and $D^{k+1}=\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right]$, note that

$$F^{k+1}\le F^k-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)D^{k+1}+\frac{\gamma}{2}G^k. \tag{1}$$

We shall now bound G^{k+1} in terms of G^k and D^{k+1} . Let $\mathcal{Z}^k\overset{\text{def}}{=}[x^{k+1},x^k,g^k]$. By tower property,

$$\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\right] \quad = \quad \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\mid\mathcal{Z}^k,\xi^k\right]\mid\mathcal{Z}^k\right]\right]. \tag{2}$$

First, let us compute the inner-most expectation:

$$\begin{aligned}\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\mid\mathcal{Z}^k,\xi^k\right] &= (1-p)\left\|g^k+H_{\xi^k}\left(x^{k+1},x^k\right)-\nabla f\left(x^{k+1}\right)\right\|^2\\ &= (1-p)\left\|g^k-\nabla f\left(x^k\right)+H_{\xi^k}\left(x^{k+1},x^k\right)-\left(\nabla f\left(x^{k+1}\right)-\nabla f\left(x^k\right)\right)\right\|^2.\end{aligned}$$

Using Young's inequality $\|a+b\|^2\le(1+\tau)\|a\|^2+\left(1+\tau^{-1}\right)\|b\|^2$, which holds for any $a,b\in\mathbb{R}^d$ and all $\tau>0$, we get

$$\begin{aligned}\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\mid\mathcal{Z}^k,\xi^k\right] &\le (1-p)(1+\tau)\left\|g^k-\nabla f\left(x^k\right)\right\|^2\\ &\quad+(1-p)\left(1+\frac{1}{\tau}\right)\left\|H_{\xi^k}\left(x^{k+1},x^k\right)-\left(\nabla f\left(x^{k+1}\right)-\nabla f\left(x^k\right)\right)\right\|^2.\end{aligned}$$

Moving on to the middle-level expectation, we deduce

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^{2} \mid \mathcal{Z}^{k}, \xi^{k}\right] \mid \mathcal{Z}^{k}\right] &\leq (1-p)(1+\tau)\left\|g^{k}-\nabla f\left(x^{k}\right)\right\|^{2} \\
&\quad +(1-p)\left(1+\frac{1}{\tau}\right) \mathbb{E}\left[\left\|H_{\xi^{k}}\left(x^{k+1}, x^{k}\right)-\left(\nabla f\left(x^{k+1}\right)-\nabla f\left(x^{k}\right)\right)\right\|^{2} \mid \mathcal{Z}^{k}\right] \\
&\leq (1-p)(1+\tau)\left\|g^{k}-\nabla f\left(x^{k}\right)\right\|^{2} \\
&\quad +(1-p)\left(1+\frac{1}{\tau}\right)\left(2 C^{2}\left\|x^{k+1}-x^{k}\right\|^{2}+2 L^{2}\left\|x^{k+1}-x^{k}\right\|^{2}\right).
\end{aligned}$$

Applying the tower property we finaly arrive at the inequality

$$\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^{2}\right] \leq (1-p)(1+\tau) \mathbb{E}\left[\left\|g^{k}-\nabla f\left(x^{k}\right)\right\|^{2}\right]+2(1-p)\left(1+\frac{1}{\tau}\right)\left(L^{2}+C^{2}\right) \mathbb{E}\left[\left\|x^{k+1}-x^{k}\right\|^{2}\right],$$

which can be written in the more compact form

$$G^{k+1} \leq (1-p)(1+\tau) G^{k}+2(1-p)\left(1+\frac{1}{\tau}\right)\left(L^{2}+C^{2}\right) D^{k+1}. \tag{3}$$

Adding inequality (1) with an m-multiple of inequality (3), where $m>0$, we get

$$\begin{aligned}
F^{k+1}+m G^{k+1} &\leq F^{k}-\frac{\gamma}{2} \mathbb{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^{2}\right]-\left(\frac{1}{2 \gamma}-\frac{L}{2}\right) D^{k+1}+\frac{\gamma}{2} G^{k}+m G^{k+1} \\
&\leq F^{k}-\frac{\gamma}{2} \mathbb{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^{2}\right]-\left(\frac{1}{2 \gamma}-\frac{L}{2}\right) D^{k+1}+\frac{\gamma}{2} G^{k} \\
&\quad +m\left((1-p)(1+\tau) G^{k}+2(1-p)\left(1+\frac{1}{\tau}\right)\left(L^{2}+C^{2}\right) D^{k+1}\right) \\
&= F^{k}+m G^{k}-\frac{\gamma}{2}\left\|\nabla f\left(x^{k}\right)\right\|^{2}-\underbrace{\left(\frac{1}{2 \gamma}-\frac{L}{2}-2 m(1-p)\left(1+\frac{1}{\tau}\right)\left(L^{2}+C^{2}\right)\right)}_{A} D^{k+1} \\
&\leq F^{k}+m G^{k}-\frac{\gamma}{2} \mathbb{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^{2}\right],
\end{aligned}$$

where we choose $m=\frac{\gamma}{2(p-\tau+p\tau)}$, and the last inequality is due to $\gamma \leq \frac{1}{L+\sqrt{\frac{2(1-p)(1+\tau)\left(L^{2}+C^{2}\right)}{\tau(p-\tau+p\tau)}}}$.

Multiplying both sides by $\frac{2}{\gamma K}$ and using the fact that $m\left\|g^{0}-\nabla f(x^{0})\right\|^{2}=m\left\|\nabla f(x^{0})-\nabla f(x^{0})\right\|^{2}=0$ (which follows from our choice $g^{0}=\nabla f\left(x^{0}\right)$), after rearranging we get

$$\sum_{k=0}^{K-1} \frac{1}{K} \mathbb{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^{2}\right] \leq \frac{2 F^{0}}{\gamma K}$$

□

p3. (1.) $H_{\xi^k}(x^{k+1}, x^k) = m(x^{k+1} - x^k)$ with m a constant to be chosen. This choice satisfy the assumption.
(2.)$H_{\xi^k}(x^{k+1}, x^k) = A(x^{k+1} - x^k)$ with A a matrix to be chosen,A contain the Hessian information of f. If $||A||_{op}$ is bounded, this choice also satisfy the assumption.
(3.) Third direction is to consider the accelaration for various $H_{\xi^k}(x^{k+1}, x^k)$. How to acclerate is not clear yet.