CS331-HW6-Lukang-Sun

October 16, 2021

Algorithm 1 DCGD-SHIFT

Parameters: shift h_i,learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, compression operators $\mathcal{C}_1 \in \mathbb{B}^d(\omega_1), \ldots, \mathcal{C}_n \in \mathbb{B}^d(\omega_n), \mathcal{C} = Id$

for $k = 0, 1, 2, \ldots$ **do**
  for all workers $i \in \{1, 2, \ldots, n\}$ in parallel **do**
    Compute local gradient $\nabla f_i\left(x^k\right)$
    Compress local gradient $g_i^k = C_i^k\left(\nabla f_i\left(x^k\right) - h_i\right)$
    Send g_i^k to master
  end for
  Master computes the aggregate $\hat{g}^k = \frac{1}{n}\sum_{i=1}^n g_i^k + \frac{1}{n}\sum_{i=1}^n h_i$
  Master broadcasts the compressed aggregate $g^k = \mathcal{C}(\hat{g}^k)$ to all workers
  for all workers $i \in \{1, 2, \ldots, n\}$ in parallel **do**
    Compute the next iterate $x^{k+1} = \operatorname{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
  end for
end for

p1.

Theorem. *Assume f_i is convex and L_i-smooth for all i, and let f be L-smooth. Let the gradient estimator* g *be defined as in Algorithm 1, where*

$$\mathcal{C}_1 \in \mathbb{B}^d(\omega_1), \quad \mathcal{C}_2 \in \mathbb{B}^d(\omega_2), \quad \ldots, \quad \mathcal{C}_n \in \mathbb{B}^d(\omega_n), \quad \mathcal{C} = Id$$

are independent compression operators,$h_i = \nabla f_i(y)$. Then

$$G(x, y) \leq 2AD_f(x, y)$$

where

$$A = L + \max_i\left(L_i \frac{\omega_i}{n}\right).$$

Let the step size $0 \le \gamma \le \frac{1}{A}$*, then*

$$\mathrm{E}\left[\left\|x^k - x^\star\right\|^2\right] \le (1-\gamma\mu)^k \left\|x^0 - x^\star\right\|^2. \tag{1}$$

Proof. $\mathcal{C}_i \in \mathbb{B}^d(\omega_i)$ are independent and $\mathcal{C}(x) = x$. We get the estimate

$$\begin{aligned}\mathrm{E}\left[\|g(x) - \nabla f(y)\|^2\right] &= \mathrm{E}\left[\|\hat{g}(x) - \nabla f(x)\|^2\right] + \mathrm{E}\left[\|\mathrm{E}\left[\hat{g}(x)\right] - \nabla f(y)\|^2\right] \\ &\le \mathrm{E}\left[\|\hat{g}(x) - \nabla f(x)\|^2\right] + 2LD_f(x,y)\end{aligned}$$

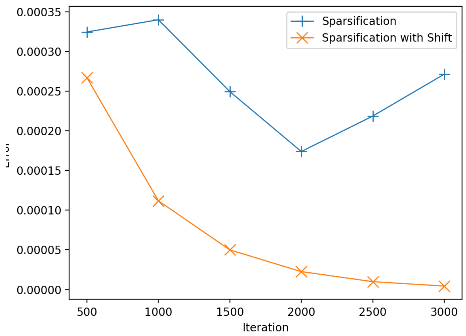
We now bound each of the above terms individually. Let $h_i = \nabla f_i(y)$, $a_i \overset{\text{def}}{=} g_i(x) + h_i - \nabla f_i(x)$ and note that $E\left[a_i\right] = 0$. The second term can be bounded as

$$\begin{aligned}\mathrm{E}\left[\|\hat{g}(x) - \nabla f(x)\|^2\right] &= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (g_i(x) + h - \nabla f_i(x))\right\|^2\right] \\ &= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right] \\ &= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^n \|a_i\|^2 + \sum_{i\neq j}\langle a_i, a_j\rangle\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{E}\left[\|a_i\|^2\right] + \sum_{i\neq j}\mathrm{E}\left[\langle a_i, a_j\rangle\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{E}\left[\|a_i\|^2\right] + \sum_{i\neq j}\langle\underbrace{E\left[a_i\right]}_{=0}, \underbrace{E\left[a_j\right]}_{0}\rangle \\ &\le \frac{1}{n^2}\sum_{i=1}^n \omega_i \left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 \\ &\le \frac{1}{n^2}\sum_{i=1}^n 2\omega_i L_i D_{f_i}(x,y) \\ &\le \frac{2\max_i \omega_i L_i}{n}\frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y) = \frac{2\max_i \omega_i L_i}{n} D_f(x,y),\end{aligned}$$

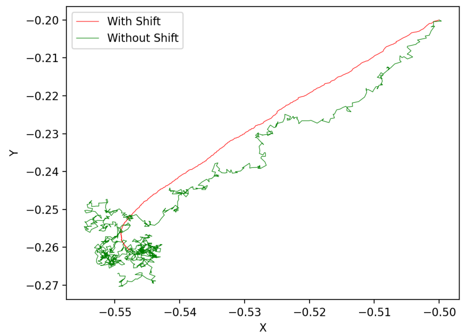
so finally

$$\mathrm{E}\left[\|g(x) - \nabla f(y)\|^2\right] \le 2(L + \frac{\max_i \omega_i L_i}{n})D_f(x,y)$$

Inequality (1) is the corollary under AC-condition, $A = L + \max_i\left(L_i \frac{\omega_i}{n}\right), C = 0$. □



(a)



(b)

Figure 1: (a) shows the error in terms of iteration,(b) shows the trajectories of SGD-IND with shift and SGD-IND without shift

p2. In my experiments(see Figure1.), I set $d=2, n=10, f(x)=\frac{1}{10} \sum_{i=1}^{10} f_i(x), f_i(x)=\frac{1}{2}\left\|a_i x-b_i\right\|_2^2$, $a=[matrix([[0.94884523, 0.31257516]]), matrix([[0.64695759, 0.79089169]]), matrix([[0.70218109, 0.91473775]]), matrix([[0.03035042, 0.21034799]]), matrix([[0.99278455, 0.2554682]]), matrix([[0.16064759, 0.09062056]]), matrix([[0.41438167, 0.77718962]]), matrix([[0.4953842, 0.93027311]]), matrix([[0.7692516, 0.19772597]]), matrix([[0.12430258, 0.03779965]])]$, $b=[matrix([[0.4231786]]), matrix([[0.524466]]), matrix([[0.17981303]]), matrix([[0.50033184]]), matrix([[0.71116473]]), matrix([[0.0604399]]), matrix([[0.37100656]]), matrix([[0.91263449]]), matrix([[0.66417151]]), matrix([[0.65995097]])]$, based on these information, we can get that $x_\star=matrix([[-0.54051746][-0.26890662]])$, I sample 10 points in the last step to estimate $\mathrm{E}\left[\left\|x_{\text{last step}}-x_\star\right\|^2\right]$. I use SGD-IND (p = 0.5)with and without shift in my experiments, the experiments shows that SGD-IND without shift will converge to a neighborhood of the optimal point, but SGD-IND with shift will converge to the optimal point, these results quite match the theory.

p3. (1.) Since in this case

$$g^k=g\left(x^k\right)-g\left(y^k\right)+\nabla f\left(y^k\right)=\nabla f(x^k),$$

which is exactly the descent direction for GD. (2.)I don't find Corollary 95. Do you mean Corollary 97?

Theorem. *Assume f is μ-convex and L-smooth .The gradient estimator $g=\nabla f+\xi$ is unbiased and satisfies the expected smoothness bound*

$$\sigma(x) \stackrel{def}{=} \mathrm{E}\left[\left\|g(x)-g\left(x^{\star}\right)\right\|^2\right] \leq 2 L D_f\left(x, x^{\star}\right)$$

Then L-SVRG with stepsize $\gamma = \frac{1}{6L}$ *satisfies*

$$\mathrm{E}\left[V^k\right] \leq \left(1 - \min\left\{\frac{\mu}{6L}, \frac{p}{2}\right\}\right)^k V^0$$

where

$$V^k \overset{(183)}{=} \left\|x^k - x^\star\right\|^2 + M\gamma^2\sigma^k, \quad M = \frac{4}{p}, \quad \sigma^k = \sigma\left(y^k\right) \overset{def}{=} \mathrm{E}\left[\left\|g\left(y^k\right) - g\left(x^\star\right)\right\|^2\right]$$

So,

$$k \geq \max\left\{\frac{6L}{\mu}, \frac{2}{p}\right\} \log\frac{1}{\varepsilon} \quad \Rightarrow \quad \mathrm{E}\left[V^k\right] \leq \varepsilon V^0$$

since $\max\left\{\frac{6L}{\mu}, \frac{2}{p}\right\} > \frac{L}{\mu}$, so this kind of analysis is worse than GD. When p is not small, the order of convergence rate is the same between GD and L-SVRG($g = \nabla f + \xi$)(ignore the constant factor), when p is small(like $p = \epsilon$), then the method in this problem is much worse than GD.

p4. The expected smoothness constant is

$$A'' = \frac{n-\tau}{\tau(n-1)} \max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L$$

In each iteration, one has to compute $2|S| = 2\tau$ 1-gradient and p n-gradient,so

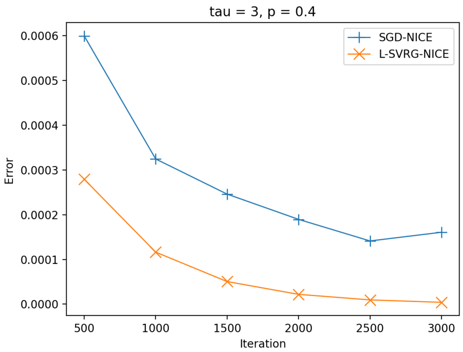
$$\mathbf{COST} = 2\tau + pn,$$

so the total complexity is

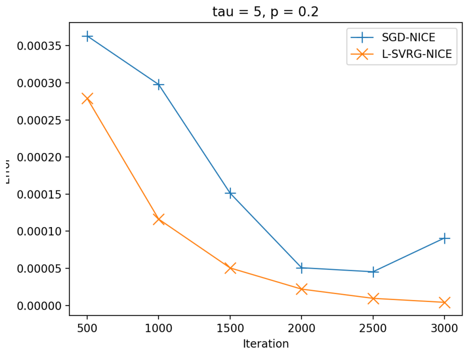
$$\max\left\{\frac{6(\frac{n-\tau}{\tau(n-1)} \max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L)(2\tau + pn)}{\mu}, \frac{2(2\tau + pn)}{p}\right\} \log(\frac{1}{\epsilon}),$$

Computing the minimum point of $(\frac{n-\tau}{\tau(n-1)} \max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L)(2\tau + pn)$, it is $\tau_\star = \sqrt{\frac{(\max L_i - L)n^2 p}{2(nL - \max L_i)}}$, since optimal τ should be an integer between 1 and n-1, we find optimal $\tau^\star$ between $\lfloor\tau_\star\rfloor$ and $\lceil\tau_\star\rceil$(only consider $\lfloor\tau_\star\rfloor, \lceil\tau_\star\rceil \in [1, n-1]$, if one of them is not in [1,n-1], ignore it), consider the $\hat{\tau}$, such that $\frac{6A''}{\mu} = \frac{1}{p}$, if $\hat{\tau} > \tau^\star$, choose $\tau_{min} = \tau^\star$, else $\tau_{min} = optimal\{\lfloor\hat{\tau}\rfloor, \lceil\hat{\tau}\rceil \in [1, n-1]\}$. τ_{min} depends on p, if p is small, then τ_{min} is also small, if p is big, then τ_{min} will also get bigger.

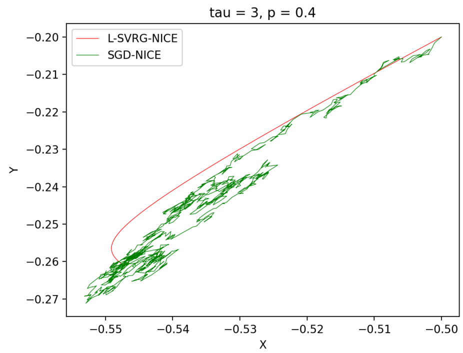
p5. In my experiments(see Figure2.), I set $d = 2, n = 10, f(x) = \frac{1}{10}\sum_{i=1}^{10} f_i(x), f_i(x) = \frac{1}{2}\left\|a_i x - b_i\right\|_2^2$, $a = [matrix([[0.94884523, 0.31257516]]), matrix([[0.64695759, 0.79089169]]), matrix([[0.70218109, 0.91473775]]), matrix([[0.03035042, 0.21034799]]), matrix([[0.99278455, 0.2554682]]), matrix([[0.16064759, 0.09062056]]), matrix([[0.41438167, 0.77718962]]), matrix([[0.4953842, 0.93027311]]), matrix([[0.7692516, 0.19772597]]), matrix([[0.12430258, 0.03779965]])]$, $b = [matrix([[0.4231786]]), matrix([[0.524466]]), matrix([[0.17981303]]), matrix([[0.50033184]]), matrix([[0.71116473]]), matrix([[0.0604399]]), matrix([[0.37100656]]), matrix([[0.91263449]]), matrix([[0.66417151]]), matrix([[0.65995097]])]$, based on these information, we can get that $x_\star = matrix([[-0.54051746][-0.26890662]])$, I sample 10 points in the last step to estimate $\mathrm{E}\left[||x_{\text{last step}} - x_\star||^2\right]$. The theory states that L-SVRG-NICE will converge to the optimal point while SGD-NICE will only converge to the neighborhood of the optimal point, this quite match the experiments results. You can see that the error of L-SVRG-NICE keeps dropping while SGD-NICE not and the trajectory of L-SVRG-NICE converge to the optimal point smoothly but the trajectory of SGD-NICE converges to the neighborhood of the optimal point with fluctuation.



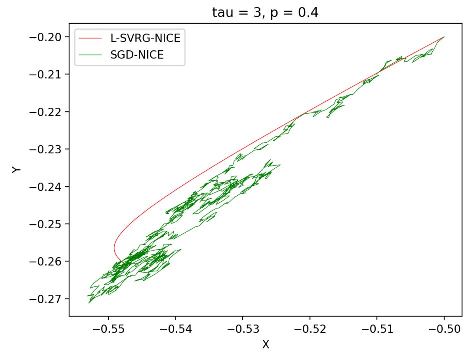
(a)



(b)



(c)



(d)

Figure 2: (a) shows the error in terms of iteration with $\tau = 3, p = 0.4$(b) shows the error in terms of iteration with $\tau = 5, p = 0.2$,(c)shows the trajectories of SGD-NICE and L-SVRG-NICE with $\tau = 3, p = 0.4$,(d) shows the trajectories of SGD-NICE and L-SVRG-NICE with $\tau = 5, p = 0.2$