CS331-HW4-Lukang-Sun

October 7, 2021

p1.

Theorem. *Assume f is convex and L-smooth and $\mu-$convex. Define the gradient estimator*

$$g(x) \overset{def}{=} \mathcal{C}(\nabla f(x))$$

where $\mathcal{C}: \mathbb{R}^d \to \mathbb{R}^d$ is a compression operator defined in problem 1. Then(AC-condition)

$$G(x,y) \leq 2\underbrace{(2\omega+1)L}_{A} D_f(x,y) + \underbrace{2\omega\|\nabla f(y)\|^2 + \delta}_{C(y)}, \tag{1}$$

and the iteration of CGD(Algorithm 11 in the lecture) satisfy

$$\mathrm{E}\left[||x^k - x^\star||^2\right] \leq (1-\gamma\mu)^k ||x^0 - x^\star||^2 + \frac{\gamma(2\omega||\nabla f(x^\star)||^2 + \delta)}{\mu}, \tag{2}$$

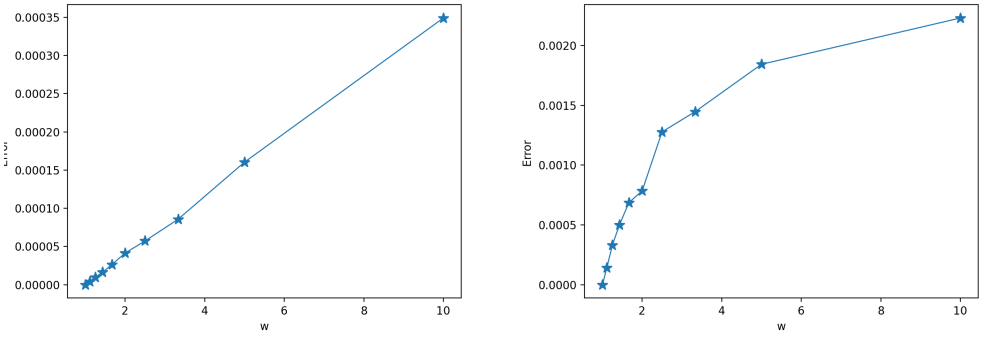
where $0 < \gamma \leq \frac{1}{A}$

Proof.

$$\begin{aligned}
G(x,y) &\overset{\text{ref}}{=} \mathrm{E}\left[||g(x) - \nabla f(y)||^2\right] \\
&= \mathrm{E}\left[||g(x) - \nabla f(x)||^2\right] + ||\nabla f(x) - \nabla f(y))||^2 \\
&= \mathrm{E}\left[||\mathcal{C}(\nabla f(x)) - \nabla f(x)||^2\right] + ||\nabla f(x) - \nabla f(y)||^2 \\
&= w||\nabla f(x)||^2 + \delta + ||\nabla f(x) - \nabla f(y)||^2 \\
&= w||\nabla f(x) - \nabla f(y) + \nabla f(y)||^2 + ||\nabla f(x) - \nabla f(y)||^2 + \delta \\
&= 2w||\nabla f(x) = \nabla f(y)||^2 + 2w||\nabla f(y)||^2 + ||\nabla f(x) - \nabla f(y)||^2 + \delta \\
&= (2w+1)||\nabla f(x) - \nabla f(y)||^2 + 2w||\nabla f(y)||^2 + \delta \\
&= 2(2w+1)LD_f(x,y) + 2w||\nabla f(y)||^2 + \delta,
\end{aligned}$$

where in the last step we have used convexity and $L-$smoothness of f, this proves (1). (2) is a direct consequence of the AC-condition, where $A = (2\omega+1)L, C = 2\omega||\nabla f(x^\star)||^2 + \delta$.

□



(a) (b)

Figure 1: The horizontal axis is $\omega+1$, vertical axis is the error, you can see the trend that when ω gets bigger, the error will also gets bigger, this quite match the theoretical analysis. In picture (a), I use constant step size $\gamma = \frac{1}{22000}$, while in picture (b), I use step size $\gamma = \frac{1}{A}$, where $A = L + \frac{2w}{n} L_{max}$, where $n = 10$.

p2. In my experiments(see Figure1.), I set $d = 1, n = 10, \epsilon = 10^{-4}, f(x) = \frac{1}{10}\sum_{i=1}^{10} f_i(x), f_i(x) = \frac{1}{2}\left\|a_i x - b_i\right\|_2^2, a = [2, 4, 6, 3, 2, 9, 6, 7, 11, 45], b = [1, 2, 3, 4, 5, 6, 7, 8, 9, 11]$, based on these information, we can get that $x_\star = 0.3343133137337253, \sigma_\star^2 = 3993.739347850591, \max_i L_i = 2025, L = \mu = 238.1$.I use Bernoulli compressor(I.I.D.) with $p = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$, so $\omega + 1 = [10, 5, 3.3, 2.5, 2, 1.67, 1.43, 1.25, 1, 11, 1]$, I sample 500 points in the last step to estimate $\mathrm{E}\left[||x_{\text{last step}} - x_\star||^2\right]$. From the experiment results, it quite match Lemma 51 in the lecture.

p3. Let $\delta(\omega_1)$ uniformly chosen from $[-\epsilon, +\epsilon]$, random matrix $A(\omega_1)$ and its entry $A(\omega_1)_{ij}$equals1, if $i = j$, equals $\delta(\omega_1)$if $i = j + 1$, equals $-\delta(\omega_1)$ if $i = j - 1$. Let $C(\omega_2)$ be drawn which independtly from $A(\omega_1)$ from random sparsification operator ensemble, then we construct the linear compression operator $K(\omega_1, \omega_2) = C(\omega_2)A(\omega_1)$. We can see random matrix$K$ is a small perturbation of the sparsifier. We can prove that $\mathrm{E}\left[K(\omega_1, \omega_2)\right] = \boldsymbol{I}_d$, and $\mathrm{E}\left[K^\top K\right]$ exists(at least exists for small enough ϵ). It is meaningful, since it acts just like sparsifier except add some noise to the non zero entry after sparsification.

p4. for symmetry,

$$\mathrm{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right]^\top = \mathrm{E}\left[(\boldsymbol{C}^\top \boldsymbol{C})^\top\right] = \mathrm{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right],$$

for positive definiteness, choose any vector $x \in \mathbb{R}^d$, we have

$$x^\top \mathrm{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right] x = \mathrm{E}\left[x^\top \boldsymbol{C}^\top \boldsymbol{C} x\right] \geq 0,$$

since $x^\top \boldsymbol{C}^\top \boldsymbol{C} x = ||\boldsymbol{C}x||_2^2 \geq 0$.

p5.

$$\begin{aligned}
\lambda_{\max}(\boldsymbol{P}^{1/2}\boldsymbol{L}\boldsymbol{P}^{1/2}) &= \max_x \frac{x^\top \boldsymbol{P}^{1/2}\boldsymbol{L}\boldsymbol{P}^{1/2}x}{x^\top \boldsymbol{P}x}\frac{x^\top \boldsymbol{P}x}{||x||^2} \\
&\leq \max_{\boldsymbol{P}^{1/2}x} \frac{x^\top \boldsymbol{P}^{1/2}\boldsymbol{L}\boldsymbol{P}^{1/2}x}{x^\top \boldsymbol{P}x} \max_x \frac{x^\top \boldsymbol{P}x}{||x||^2} \\
&\leq \lambda_{\max}(\boldsymbol{L}) \max_x \frac{x^\top \boldsymbol{P}x}{||x||^2} \\
&\leq \lambda_{\max}(\boldsymbol{L})\lambda_{\max}(\boldsymbol{P}).
\end{aligned}$$

($\boldsymbol{P}$ should be $\boldsymbol{P}^{-1}$.)For the first inequality, I can not prove the general case. The first inequality is equivalent to

$$\max_{||x||_2^2 \leq 1} \sum_{i,j} L_{i,j}\frac{P_{i,j}}{P_iP_j}x_ix_j \leq \max_{||x||_2^2 \leq 1} \sum_{i,j} L_{i,j}\frac{P_i^{1/2}P_j^{1/2}}{P_iP_j}x_ix_j,$$

where $P_{i,j} := \sum_S P(S)1_S(i)1_S(j), P_i := \sum_S P(S)1_S(i)$, we know $P_{i,j} \leq P_i^{1/2}P_j^{1/2}$, for any (i,j). But for matrix $[P_{i,j}]$ whose entry is $P_{i,j}$ and matrix $[P_i^{1/2}P_j^{1/2}]$ whose entry is $P_i^{1/2}P_j^{1/2}$, we don't have $[P_{i,j}] \preccurlyeq [P_i^{1/2}P_j^{1/2}]$(we have $\lambda_{\max}([P_{i,j}]) \leq \lambda_{\max}([P_i^{1/2}P_j^{1/2}])$, the purpose is to prove $\lambda_{\max}([P_{i,j}] \circ L) \leq \lambda_{\max}([P_i^{1/2}P_j^{1/2}] \circ L)$, where L is semi positive definite.)

Case one: L has nonnegative entries(this include the case that L is diagonal.) Choose x such that $\lambda_{\max}(\mathrm{E}[C^\top LC]) = \sum_{i,j} L_{i,j}\frac{P_{i,j}}{P_iP_j}x_ix_j$, then $\sum_{i,j} L_{i,j}\frac{P_{i,j}}{P_iP_j}x_ix_j \leq \sum_{i,j} L_{i,j}\frac{P_{i,j}}{P_iP_j}|x_i||x_j| \leq \sum_{i,j} L_{i,j}\frac{P_i^{1/2}P_j^{1/2}}{P_iP_j}|x_i||x_j| \leq \max_{||x||_2^2\leq 1}\sum_{i,j} L_{i,j}\frac{P_i^{1/2}P_j^{1/2}}{P_iP_j}x_ix_j = \lambda_{\max}\mathrm{E}[P^{-1/2}LP^{-1/2}]$.

Case two: L has rank one. In this case there is vector $\tilde{L}$, such that $L_{i,j} = \tilde{L}_i\tilde{L}_j$. Choose x such that $\lambda_{\max}(\mathrm{E}[C^\top LC]) = \sum_{i,j}\frac{P_{i,j}}{P_iP_j}\tilde{L}_ix_i\tilde{L}_jx_j$, then $\sum_{i,j}\frac{P_{i,j}}{P_iP_j}\tilde{L}_ix_i\tilde{L}_jx_j \leq \sum_{i,j}\frac{P_{i,j}}{P_iP_j}|\tilde{L}_ix_i||\tilde{L}_jx_j| \leq \sum_{i,j}\frac{P_i^{1/2}P_j^{1/2}}{P_iP_j}|\tilde{L}_ix_i||\tilde{L}_jx_j| \overset{\text{change sign of } x_i \text{ such that } \tilde{L}_ix_i \text{ nonne}}{\leq} \max_{||x||_2^2\leq 1}\sum_{i,j}\frac{P_i^{1/2}P_j^{1/2}}{P_iP_j}\tilde{L}_ix_i\tilde{L}_jx_j = \lambda_{\max}\mathrm{E}[P^{-1/2}LP^{-1/2}]$.

If

$$\lambda_{\max}\left(\mathrm{E}\left[C_SLC_S\right]\right) \leq \lambda_{\max}\left(P^{-1/2}LP^{-1/2}\right) \leq \lambda_{\max}(L)\lambda_{\max}(P^{-1}),$$

this means the step size for the third analysis can be at least as big as the second convergence analysis and in the second convergence analysis, the step size can be as big as the the step size in the first convergence analysis. So the convergence rate for the third one is better than the second one, the second one is better than the first one.

$$F := \frac{1}{n}\sum_{i=1}^{n} f_i + g, f_i \text{ strongly smooth nonconvex, } g \text{ convex but nonsmooth}$$