CS331-HW3-Lukang-Sun

September 22, 2021

p1. Let's write the Lagrangian for this optimization problem: $L(p_1, \cdots, p_n, \lambda) = \sum_{i=1}^n \frac{a_i}{p_i} + \lambda(\sum_{i=1}^n p_i - 1.)$ Next, we should equate to zero the derivative with respect to each p_i, for $i = 1, \cdots, n$,

$$\nabla_{p_i} L(p_1, \cdots, p_n, \lambda) = -\frac{a_i}{p_i^2} + \lambda = 0 \Rightarrow p_i = \sqrt{\frac{a_i}{\lambda}}, \tag{1}$$

we also need $\sum_{i=1}^n p_i = 1$, combine this with (1), we reach $\sqrt{\lambda} = \sum_{i=1}^n \sqrt{a_i}$, finally use (1) again, we have $p_i = \frac{\sqrt{a_i}}{\sum_j \sqrt{a_j}}$.

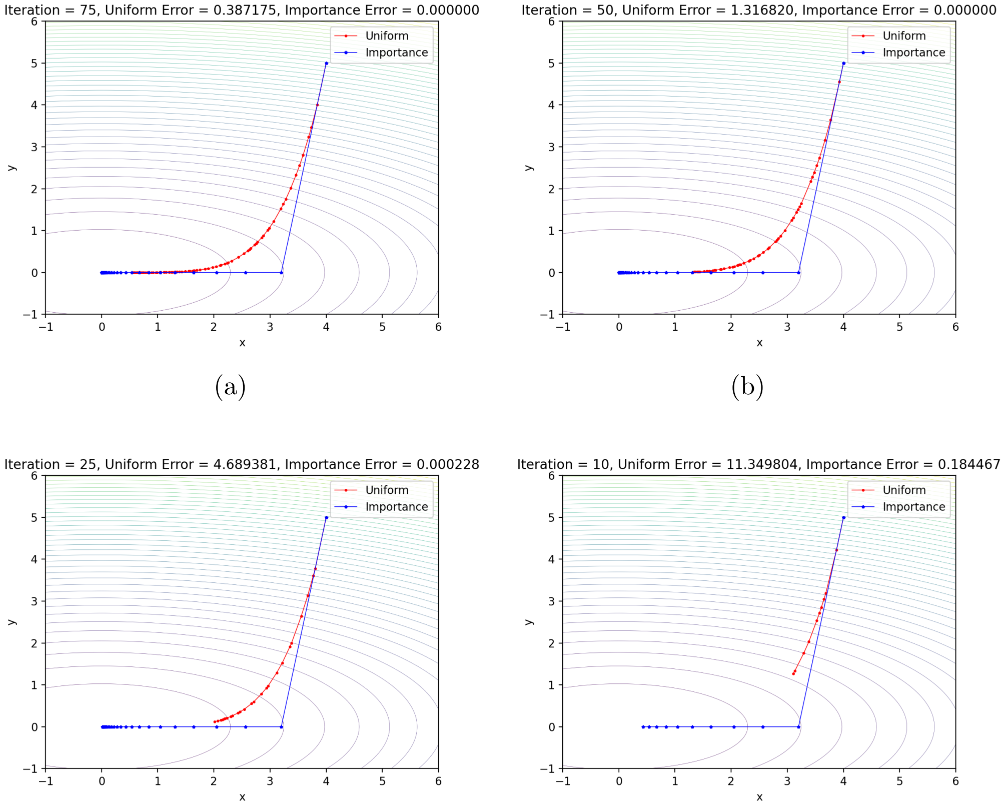
p2. In my experiments(see Figure 1.), I set $g(x, y) = 0.02x^2 + y^2$, $L = [2, 4, 6, 3, 2, 9, 6, 7, 11, 45]$, $f_i(x, y) = 0.5L_i g(x, y)$, $f(x, y) = 0.1 \sum_{i=0}^9 f_i(x, y)$, step size for SGD-US is $\frac{1}{\max_i L_i} = \frac{1}{45}$, for SGD-IS is $\frac{1}{10 \sum_{i=0}^9 L_i} = \frac{1}{9.5}$. We can see that $\frac{\max_i L_i}{\sum_{i=0}^9 L_i} \approx 4.74$, so SGD-IS should approximately converge to the minimum point with much fewer steps than SGD-US should do, actually, this is demonstrated by my simulations. Due to my design of optimization function, you can not observe randomness in the trajectories of SGD-IS, while in the trajectories of SGD-US there is randomness. I run 100 times to estimate the mean error that is Error $= E\left[||x_{iteration}||^2\right] \approx \frac{1}{100} \sum_{i=1}^{100} ||x_{iteration}^i||^2$, $x_{iteration}^i$ is the i-th sample of the last iteration point.

p3. Denote $|S|$ as the cardinality of S, $P_k = C_n^k P(S = U)$, where U is any set whose cardinality is k. We define the unbiased estimator as

$$g(x) \overset{\text{def}}{=} \frac{1}{nP_{|S|}|S|} \sum_{i \in S} \nabla f_i(x). \tag{2}$$

Theorem. *The gradient estimator g defined in (2) is unbiased. If we further assume that $n \geq 2$, f_i is convex and L_i−smooth for all i, and f is L−smooth, then*

$$\mathrm{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y) \tag{3}$$



Iteration = 75, Uniform Error = 0.387175, Importance Error = 0.000000

Iteration = 50, Uniform Error = 1.316820, Importance Error = 0.000000

Iteration = 25, Uniform Error = 4.689381, Importance Error = 0.000228

Iteration = 10, Uniform Error = 11.349804, Importance Error = 0.184467

(a)

(b)

(c)

(d)

Figure 1: The trajectories and empirical error of SGD-US and SGD-IS with different iterations, you can see the blue lines converge much faster than the red line.

where

$$A'' = \sum_{k=1}^{n} \frac{1}{n^2 P_k} \left(\frac{n-k}{k(n-1)} \max_i L_i + \frac{n(k-1)}{k(n-1)} L \right). \tag{4}$$

Proof. **Unbiasedness.** Let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}$$

It is easy to show that

$$\mathrm{E}\left[\chi_i \mid |S|\right] = \frac{|S|}{n}$$

Unbiasedness of $g(x)$ now follows via direct computation:

$$\begin{aligned} \mathrm{E}[g(x)] &= \mathrm{E}\left[\mathrm{E}\left[\frac{1}{nP_{|S|}|S|}\sum_{i\in S}\nabla f_i(x) \mid |S|\right]\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\frac{1}{nP_{|S|}|S|}\sum_{i=1}^{n}\chi_i\nabla f_i(x) \mid |S|\right]\right] \\ &= \mathrm{E}\left[\frac{1}{nP_{|S|}|S|}\sum_{i=1}^{n}\mathrm{E}\left[\chi_i \mid |S|\right]\nabla f_i(x)\right] \\ &= \mathrm{E}\left[\frac{|S|}{n}\frac{1}{nP_{|S|}|S|}\sum_{i=1}^{n}\nabla f_i(x)\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} P_k \left(\frac{1}{nP_k}\sum_{i=1}^{n}\nabla f_i(x)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) = \nabla f(x). \end{aligned}$$

Expected smoothness (i.e., computing constant A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y)$$

Let χ_{ij} be the random variable defined by

$$\chi_{ij} = \begin{cases} 1 & i \in S \text{ and } j \in S \\ 0 & \text{otherwise} \end{cases}$$

Note that

$$\chi_{ij} = \chi_i \chi_j$$

Further, it is easy to show that

$$\mathrm{E}\left[\chi_{ij} \mid |S|\right] = \frac{|S|(|S|-1)}{n(n-1)}$$

It is easy to check that for any vectors $b_1, \ldots, b_n \in \mathbb{R}^d$ we have the identity

$$\left\| \sum_{i=1}^n b_i \right\|^2 - \sum_{i=1}^n \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle .$$

We will use this identity twice in what follows:

$$\begin{aligned}
\mathrm{E}\left[\|g(x)-g(y)\|^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\left\|\tfrac{1}{nP_{|S|}|S|}\textstyle\sum_{i\in S}\nabla f_i(x) - \tfrac{1}{nP_{|S|}|S|}\textstyle\sum_{i\in S}\nabla f_i(y)\right\|^2 \mid |S|\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left\|\tfrac{1}{nP_{|S|}|S|}\textstyle\sum_{i=1}^n \chi_i a_i\right\|^2 \mid |S|\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[(\tfrac{1}{nP_{|S|}|S|})^2 \textstyle\sum_{i=1}^n \|\chi_i a_i\|^2 + (\tfrac{1}{nP_{|S|}|S|})^2 \textstyle\sum_{i\neq j} \langle \chi_i a_i, \chi_j a_j\rangle \mid |S|\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[(\tfrac{1}{nP_{|S|}|S|})^2 \textstyle\sum_{i=1}^n \|\chi_i a_i\|^2 + (\tfrac{1}{nP_{|S|}|S|})^2 \textstyle\sum_{i\neq j} \chi_{ij} \langle a_i, a_j\rangle \mid |S|\right]\right] \\
&= \mathrm{E}\left[(\tfrac{1}{nP_{|S|}|S|})^2 \left(\textstyle\sum_{i=1}^n E\left[\chi_i \mid |S|\right] \|a_i\|^2 + \textstyle\sum_{i\neq j} E\left[\chi_{ij} \mid |S|\right] \langle a_i, a_j\rangle\right)\right] \\
&= \mathrm{E}\left[(\tfrac{1}{nP_{|S|}|S|})^2 \left(\textstyle\sum_{i=1}^n \tfrac{|S|}{n} \|a_i\|^2 + \textstyle\sum_{i\neq j} \tfrac{|S|(|S|-1)}{n(n-1)} \langle a_i, a_j\rangle\right)\right] \\
&= \mathrm{E}\left[(\tfrac{1}{nP_{|S|}})^2 \left(\tfrac{1}{|S|n} \textstyle\sum_{i=1} ||a_i||^2 + \tfrac{|S|-1}{|S|n(n-1)} \textstyle\sum_{i\neq j} \langle a_i, a_j\rangle\right)\right] \\
&= \mathrm{E}\left[(\tfrac{1}{nP_{|S|}})^2 \left(\tfrac{1}{|S|n} \textstyle\sum_{i=1}^n \|a_i\|^2 + \tfrac{|S|-1}{|S|n(n-1)} \left(\|\textstyle\sum_{i=1}^n a_i\|^2 - \textstyle\sum_{i=1}^n \|a_i\|^2\right)\right)\right] \\
&= \mathrm{E}\left[(\tfrac{1}{nP_{|S|}})^2 \left(\tfrac{n-|S|}{|S|(n-1)} \tfrac{1}{n} \textstyle\sum_{i=1}^n \|a_i\|^2 + \tfrac{n(|S|-1)}{|S|(n-1)} \left\|\tfrac{1}{n} \textstyle\sum_{i=1}^n a_i\right\|^2\right)\right]
\end{aligned} \tag{5}$$

Since f_i is convex and L_i-smooth, we know that

$$\|a_i\|^2 = \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i D_{f_i}(x,y)$$

Since f is convex and L-smooth, we know that

$$\left\| \frac{1}{n} \sum_{i=1}^n a_i \right\|^2 = \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x,y)$$

It only remains to plug these bounds to (5), apply the bound $L_i \leq \max_i L_i$

and use the identity $D_f(x,y) = \frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y)$:

$$\begin{aligned}
\mathrm{E}\left[\|g(x)-g(y)\|^2\right] &\leq \mathrm{E}\left[(\frac{1}{nP_{|S|}})^2\left(\frac{n-|S|}{|S|(n-1)}\frac{1}{n}\sum_{i=1}^n 2L_i D_{f_i}(x,y) + \frac{n(|S|-1)}{|S|(n-1)}2LD_f(x,y)\right)\right] \\
&\leq 2\mathrm{E}\left[(\frac{1}{nP_{|S|}})^2\left(\frac{n-|S|}{|S|(n-1)}\max_i L_i \frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y) + 2\frac{n(|S|-1)}{|S|(n-1)}LD_f(x,y)\right)\right] \\
&= 2\mathrm{E}\left[(\frac{1}{nP_{|S|}})^2\left(\frac{n-|S|}{|S|(n-1)}\max_i L_i D_f(x,y) + 2\frac{n(|S|-1)}{|S|(n-1)}LD_f(x,y)\right)\right] \\
&= 2\mathrm{E}\left[(\frac{1}{nP_{|S|}})^2\left(\frac{n-|S|}{|S|(n-1)}\max_i L_i + \frac{n(|S|-1)}{|S|(n-1)}L\right)D_f(x,y)\right] \\
&= \left(2\sum_{k=1}^n P_k(\frac{1}{nP_k})^2\left(\frac{n-k}{k(n-1)}\max_i L_i + \frac{n(k-1)}{k(n-1)}L\right)\right)D_f(x,y) \\
&= \left(2\sum_{k=1}^n \frac{1}{n^2P_k}\left(\frac{n-k}{k(n-1)}\max_i L_i + \frac{n(k-1)}{k(n-1)}L\right)\right)D_f(x,y),
\end{aligned}$$

so $A'' = \sum_{k=1}^n \frac{1}{n^2P_k}\left(\frac{n-k}{k(n-1)}\max_i L_i + \frac{n(k-1)}{k(n-1)}L\right)$.

□

p4. In my experiments(see Figure2.), I set $d = 1, n = 10, \epsilon = 10^{-4}, f(x) = \frac{1}{10}\sum_{i=1}^{10} f_i(x), f_i(x) = \frac{1}{2}\|a_ix - b_i\|_2^2$, $a = [2, 4, 6, 3, 2, 9, 6, 7, 11, 45], b = [1, 2, 3, 4, 5, 6, 7, 8, 9, 11]$, based on these information, we can get that $x_\star = 0.3343133137337253, \sigma_\star^2 = 3993.739347850591, \max_i L_i = 2025, L = \mu = 238.1$, then insert this information into the following formula

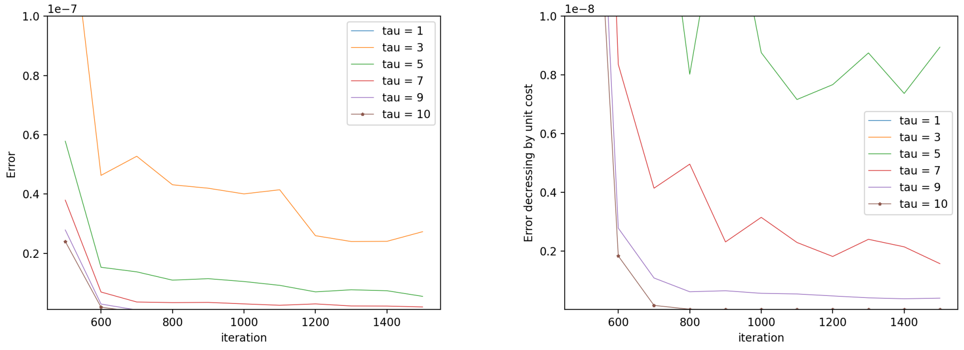
$$\tau^\star = \frac{n\left(\theta + L - \max_i L_i\right)}{\theta + nL - \max_i L_i}, \text{ where } \theta = \frac{2\sigma_\star^2}{\varepsilon\mu},$$

we get $\tau_\star = 9.936$, then we calculate the value of

$$\mathcal{C}(\tau) = \frac{2}{\mu(n-1)}\max\{(n-\tau)\max_i L_i + n(\tau-1)L, (n-\tau)\frac{2\sigma_\star^2}{\varepsilon\mu}\}, \text{ for } \tau = 9, 10,$$

we get $\mathcal{C}(9) = 335467.3958715322 > \mathcal{C}(10) = 21429$, so $\tau_\star = 10$. I sample 100 points in the last step to estimate $\mathrm{E}\left[||x_{\text{last step}} - x_\star||^2\right]$

p5.



(a)

(b)

Figure 2: Note some line are not in the plot, since their values are bigger than the upper limit of y-axis. In picture (a), the y-axis is the mean square distance from the minimum point sampled from the last step, in picture (b), I divide the error by (step times tau), which means the error decreasing by unit cost, the lower the value, means the better of the choice of tau. In both plots, you can see the line of tau equals 10 achieves the lowest value.

Proof. As before, let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}$$

It is easy to show that

$$\mathrm{E}\left[\chi_i\right] = \mathrm{Prob}(i \in S) = \mathrm{Prob}\left(i \in S_i\right) = p_i$$

$$\mathrm{E}\left[\chi_i \chi_j\right] = \mathrm{E}\left[\chi_i\right] \mathrm{E}\left[\chi_j\right] = p_i p_j, \text{ for } i \neq j,$$

denote $a_i(x) = \nabla f_i(x)$, $\mathrm{E}\left[||g(x)||^2\right]$ now follows via direct computation:

$$\begin{aligned}
\mathrm{E}[||g(x)||^2] &= \mathrm{E}\left[||\sum_{i\in S}\frac{1}{np_i}\nabla f_i(x)||^2\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[||\sum_{i=1}^{n}\chi_i\frac{1}{p_i}a_i(x)||^2\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{n}\frac{1}{p_i^2}||\chi_i a_i(x)||^2 + \sum_{i\neq j}\frac{1}{p_ip_j}\langle\chi_i a_i(x), \chi_j a_j(x)\rangle\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\frac{1}{p_i^2}||a_i(x)||^2\mathrm{E}\left[\chi_i\right] + \sum_{i\neq j}\frac{1}{p_ip_j}\mathrm{E}\left[\chi_i\chi_j\right]\langle a_i(x), a_j(x)\rangle\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\frac{1}{p_i}||a_i(x)||^2 + \sum_{i\neq j}\langle a_i(x), a_j(x)\rangle\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\frac{1}{p_i}||a_i(x)||^2 + ||\sum_{i=1}^{n}a_i(x)||^2 - \sum_{i=1}^{n}||a_i(x)||^2\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\frac{1}{p_i}-1\right)||a_i(x)||^2 + ||\sum_{i=1}^{n}a_i(x)||^2\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\frac{1}{p_i}-1\right)||\nabla f_i(x)||^2 + ||\sum_{i=1}^{n}\nabla f_i(x)||^2\right) \\
&= ||\nabla f(x)||^2 + \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{p_i}-1\right)||\nabla f_i(x)||^2
\end{aligned}$$

□