Potential Problem for Problem Set 3 - Answer / References

Xinyu Luo

October 11, 2021

1 Reference

This question is about the Hogwild! algorithm. The test of time reward on NIPS conference 2020. The story part of this question is coming from this interview. The narrator told a really interesting story behind this algorithm. And that story was the motivation of this question.

Materials reference:

- Lecture 14 from Prof. Dimitris Papailiopoulos: http://papail.io/teaching/901/scribe_14.pdf;
- Original paper: HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent;
- A nice analysis for Hogwild!: Perturbed Iterate Analysis for Asynchronous Stochastic Optimization.

2 Question Answers

2.1 Question (a)

Let the starting radius $\|x^* - x_0\|_2 \leq R$ and m-strong convexity of $\|\nabla f_{s_k}(x)\|_2 \leq M$. From lecture notes we have the classic SGD will convergence after $T = R^2M^2/\epsilon^2$, and get $\mathbb{E}\|x^* - \hat{x}\|_2 \leq \epsilon$. And in each iteration, only one processor is working and need to communicate with all other processors, thus, the total time is $(1 + (p/2)^2)T$, parallel SGD costs more time than classic SGD.

2.2 Question (b)

For noisy SGD, and $\hat{x}_k$ represent x_k with noise, we have:

$$x_{k+1} = x_k - \gamma \nabla f_{s_k}(\hat{x}_k).$$

Then,

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \gamma \nabla f_{s_k}(\hat{x}_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - 2\gamma\langle x_k - x^*, \nabla f_{s_k}(\hat{x}_k)\rangle + \gamma^2 \|\nabla f_{s_k}(\hat{x}_k)\|^2 \\
&= \|x_k - x^*\|^2 - 2\gamma\langle \hat{x}_k - x^*, \nabla f_{s_k}(\hat{x}_k)\rangle + \gamma^2 \|\nabla f_{s_k}(\hat{x}_k)\|^2 + 2\gamma\langle \hat{x}_k - x_k, \nabla f_{s_k}(\hat{x}_k)\rangle
\end{aligned}$$

Since f is m-strongly convex, we know that

$$\langle \hat{x}_k - x^*, \nabla f_{s_k}(\hat{x}_k)\rangle \geq m\|\hat{x}_k - x^*\|^2 \geq \frac{m}{2} - m\|\hat{x}_k - x_k\|^2,$$

by triangle inequality.
Then, we have

$$\mathbb{E}\|x_{k+1} - x^*\|^2 \leq (1 - \gamma m)\mathbb{E}\|x_k - x^*\|^2 + \gamma^2 \mathbb{E}\|\nabla f_{s_k}(\hat{x}_k)\|^2 + 2\gamma m \mathbb{E}\|\hat{x}_k - x_k\|^2 + 2\gamma \mathbb{E}\langle \hat{x}_k - x_k, \nabla f_{s_k}(\hat{x}_k)\rangle.$$

We already have $\|\nabla f_{s_k}(\hat{x}_k)\| \leq M^2$. If $2\gamma m\mathbb{E}\|\hat{x}_k - x_k\|^2$ and $\gamma\mathbb{E}\langle \hat{x}_k - x_k, \nabla f_{s_k}(\hat{x}_k)\rangle$ are both less than or equal to $\mathcal{O}(\gamma^2 M^2)$, by telescoping sum and triangle inequality, we have $\mathbb{E}\|x_T - x^*\|^2 \leq c\epsilon$. Thus, we claimed noisy SGD gets same convergence rate as the classic SGD up to multiple constant. (This claim should include more details... will fix later.)

2.3 Question (c) - (Need to revise based on the 2017 paper)

The error for each overlap sample i.e. s_k, there are only 3 states, either the error is count as positive or negative or doesn't count. Thus, there must be a diagnal matrix S_i^k with entries $\{-1,1,0\}$ satisfies:

$$\hat{x}_k - x_k = \sum_{i=k-\tau, i\neq k}^{k+\tau} \gamma S_i^k \nabla f_{s_i}(\hat{x}_i)$$

.

Prove $2\gamma m \mathbb{E}\|\hat{x}_k - x_k\|^2, 2\gamma \mathbb{E}\langle \hat{x}_k - x_k, \nabla f_{s_k}(\hat{x}_k)\rangle \leq \mathcal{O}(\gamma^2 M^2)$**.**

$$\begin{aligned}
\gamma^2 \langle \sum_{i=k-\tau, i\neq k}^{k+\tau} \gamma S_i^k \nabla f_{s_i}(\hat{x}_i), \nabla f_{s_k}(\hat{x}_k)\rangle &\leq \gamma^2 |\langle \sum_{i=k-\tau, i\neq k}^{k+\tau} \gamma S_i^k \nabla f_{s_i}(\hat{x}_i), \nabla f_{s_k}(\hat{x}_k)\rangle| \\
&\leq \gamma^2 \sum_{i=k-\tau, i\neq k}^{k+\tau} \|\nabla f_{s_i}(\hat{x}_i)\| * \|\nabla f_{s_j}(\hat{x}_j)\| * I_{s_i \cap s_k \neq 0} \\
&\leq \gamma^2 \sum_{i=k-\tau, i\neq k}^{k+\tau} \frac{1}{2}(\|\nabla f_{s_i}(\hat{x}_i)\| + \|\nabla f_{s_j}(\hat{x}_j)\|) * I_{s_i \cap s_k \neq 0} \\
&\leq \gamma^2 \sum_{i=k-\tau, i\neq k}^{k+\tau} M^2 * I_{s_i \cap s_k \neq 0}
\end{aligned}$$

The second "$\leq$" is due to Cauchy-Schwarz inequality.

$$\begin{aligned}
\gamma^2 \mathbb{E}\langle \sum_{i=k-\tau, i\neq k}^{k+\tau} \gamma S_i^k \nabla f_{s_i}(\hat{x}_i), \nabla f_{s_k}(\hat{x}_k)\rangle &\leq \gamma^2 \mathbb{E}\{ \sum_{i=k-\tau, i\neq k}^{k+\tau} M^2 * I_{s_i \cap s_k \neq 0}\} \\
&\leq \gamma^2 * 2\tau * M^2 * \mathbb{E}\{I_{s_i \cap s_k \neq 0}\} \\
&= \gamma^2 * 2\tau * M^2 * Pr\{s_i \cap s_k \neq 0\} \\
&= \gamma^2 * 2\tau * M^2 * \frac{\Delta}{n}
\end{aligned}$$

Thus, let $\tau \leq \frac{n}{2\Delta}$, we have both terms less than or equal to $\mathcal{O}(\gamma^2 M^2)$.

2.4 Question (d) - Need to explain more details

With step size $\gamma = \frac{\epsilon m}{2M^2}$, reaches an accuracy of $\mathbb{E}\|x_k - x^*\|^2 \leq \epsilon$, after

$$T \geq \mathcal{O}(1) \frac{M^2 \log(R/\epsilon)}{\epsilon m^2}.$$

I am not very sure about the explaination of the log part.