CEE 690/ECE 590: Solutions of Homework 2

Wei Wen
Duke University
wei.wen@duke.edu

Abstract

This includes solutions for homework 2.

1 Problem 1

(a)

Please refer to Appendix B in [1] (page 83).

(b)

Refer to slide 30 and 31 in lecture 5,

$$\mathbb{E}\left[F(\boldsymbol{w}_{k+1})\right] - F\left(\boldsymbol{w}_k\right) \leq -\alpha_k \left[\nabla F\left(\boldsymbol{w}_k\right)\right]^T \mathbb{E}\left[\tilde{\boldsymbol{g}}_k\right] + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[||\tilde{\boldsymbol{g}}_k||_2^2\right]. \tag{1}$$

The goal is to maximize the decrease of loss function per iteration, *i.e.*, minimizing the left part of Inequality 1. Therefore, we want to minimize the the upper bound which is the right side of Inequality 1.

We define the right side of Inequality 1 as

$$D(\alpha_k) \triangleq -\alpha_k \left[\nabla F\left(\boldsymbol{w}_k\right)\right]^T \mathbb{E}\left[\tilde{\boldsymbol{g}}_k\right] + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[||\tilde{\boldsymbol{g}}_k||_2^2\right]. \tag{2}$$

In general, given loss function and dataset distribution, the optimal α_k satisfies

$$\frac{\partial D(\alpha_k)}{\partial \alpha_k} = -\left[\nabla F\left(\boldsymbol{w}_k\right)\right]^T \mathbb{E}\left[\tilde{\boldsymbol{g}}_k\right] + \alpha_k L \cdot \mathbb{E}\left[||\tilde{\boldsymbol{g}}_k||_2^2\right] = 0, \tag{3}$$

that is,

$$\alpha_k = \frac{\left[\nabla F\left(\boldsymbol{w}_k\right)\right]^T \mathbb{E}\left[\tilde{\boldsymbol{g}}_k\right]}{L \cdot \mathbb{E}\left[||\tilde{\boldsymbol{g}}_k||_2^2\right]} \tag{4}$$

Substituting $\alpha_k = \frac{1}{cL}$ to Equation (4), we have

$$\mathbb{E}\left[||\tilde{\boldsymbol{g}}_k||_2^2\right] = c\left[\nabla F\left(\boldsymbol{w}_k\right)\right]^T \mathbb{E}\left[\tilde{\boldsymbol{g}}_k\right], \tag{5}$$

which is the condition/regime where the optimal learning rate is $\alpha_k = \frac{1}{cL}$. Suppose we have unbiased gradient, *i.e.*,

$$\mathbb{E}\left[\tilde{\boldsymbol{g}}_k\right] = \nabla F\left(\boldsymbol{w}_k\right) \triangleq \boldsymbol{g}_k, \tag{6}$$

then,

$$\mathbb{E}\left[||\tilde{\boldsymbol{g}}_k||_2^2\right] = Var\left(||\tilde{\boldsymbol{g}}_k||_2\right) + ||\boldsymbol{g}_k||_2^2 = c||\boldsymbol{g}_k||_2^2, \tag{7}$$

that is,

$$Var\left(||\tilde{\boldsymbol{g}}_k||_2\right) = (c-1) \cdot ||\boldsymbol{g}_k||_2^2. \tag{8}$$

where $c \geq 1$.

In case that variance is zero, *i.e.*, $Var\left(||\tilde{\boldsymbol{g}}_k||_2\right) = 0$, then $\alpha_k = \frac{1}{L}$. Otherwise, when c is larger, *i.e.*, a larger variance in the gradient estimation, the optimal learning rate α_k will be smaller.

When the variance is bounded as $Var\left(||\tilde{\boldsymbol{g}}_k||_2\right) \leq M$,

$$(c-1) \cdot ||\boldsymbol{g}_k||_2^2 \leq M, \tag{9}$$

so, the optimal learning rate is

$$\alpha_k = \frac{1}{cL} \geq \frac{||\boldsymbol{g}_k||_2^2}{L\left(M + ||\boldsymbol{g}_k||_2^2\right)}. \tag{10}$$

2 Problem 2

Code and analyses will be pushed here.

References

[1] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.