ECE590-16 Intro to Deep Learning

HW2 Yifan Li yl506

Problem 1.

(a) Fundamental theorem of calculus:

$$F(\vec{y}) = F(\vec{x}) + \int_0^1 (\vec{y}-\vec{x})^T \nabla F(\vec{x}+t(\vec{y}-\vec{x}))\,dt$$

Lipschitz continuous:

$$\|\nabla F(\vec{y}) - \nabla F(\vec{x})\|_2 \le L\|\vec{y}-\vec{x}\|_2$$

$$\begin{aligned} F(\vec{y}) &= F(\vec{x}) + \int_0^1 (\vec{y}-\vec{x})^T \nabla F(\vec{x}+t(\vec{y}-\vec{x}))\,dt \\ &= F(\vec{x}) + \int_0^1 \nabla F(\vec{x}+t(\vec{y}-\vec{x}))^T (\vec{y}-\vec{x})\,dt \\ &= F(\vec{x}) + \nabla F(\vec{x})^T(\vec{y}-\vec{x}) + \int_0^1 [\nabla F(\vec{x}+t(\vec{y}-\vec{x})) - \nabla F(\vec{x})]^T (\vec{y}-\vec{x})\,dt \end{aligned}$$

By the defintion of lipschitz continuous, we have

$$\nabla F(\vec{x}+t(\vec{y}-\vec{x})) - \nabla F(\vec{x}) \le L\|t(\vec{y}-\vec{x})\|_2$$

Then

$$F(\vec{y}) \le F(\vec{x}) + \nabla F(\vec{x})^T(\vec{y}-\vec{x}) + \int_0^1 L\|t(\vec{y}-\vec{x})\|_2 \|\vec{y}-\vec{x}\|_2\,dt$$

$$F(\vec{y}) \le F(\vec{x}) + \nabla F(\vec{x})^T(\vec{y}-\vec{x}) + \tfrac{1}{2}L\|\vec{y}-\vec{x}\|_2^2$$

(b) Lemma (sequence for stochastic gradient): $\vec{W_{k+1}} = \vec{W_k} - \alpha_k \tilde{g}_k$ satisfies in expectation ~~$E[F(\vec{W_{k+1}})] - F(\vec{W_k}) \le -\alpha_k[\nabla F(\vec{W_k})]^T$~~

$$E[F(\vec{W_{k+1}})] - F(\vec{W_k}) \le -\alpha_k[\nabla F(\vec{W_k})]^T E[\tilde{g}_k] + \frac{\alpha_k^2 L}{2} E[\|\tilde{g}_k\|_2^2]$$

Want to minimize R.H.S

let $f(\alpha_k) = -\alpha_k[\nabla F(\vec{W_k})]^T E[\tilde{g}_k] + \frac{\alpha_k^2 L}{2} E[\|\tilde{g}_k\|_2^2]$

$$\frac{df(\alpha_k)}{d\alpha_k} = 0 = -\alpha_k g_k^T E[\tilde{g}_k] + \alpha_k L\, E[\|\tilde{g}_k\|_2^2]$$

By defintion of variance, $Var[\|\tilde{g}_k\|_2] = E[\|\tilde{g}_k\|_2^2] - E[\|\tilde{g}_k\|_2]^2$

so we have

$$\alpha_k = \frac{g_k^T E[\tilde{g}_k]}{L\,(Var[\|\tilde{g}_k\|_2] + E[\|\tilde{g}_k\|]^2)} = \frac{\|g_k\|_2^2}{L(M + E[\|\tilde{g}_k\|_2]^2)}, \quad var(\|\tilde{g}_k\|_2) \le M$$

if we consider an unbiased gradient estimator, and assume that the variance is bounded, we have $E[\tilde{g}_k] = g_k$, $var(\|\tilde{g}_k\|_2^2) \le M$

If ① $\alpha_k = \frac{1}{L} = \frac{\|g_k\|_2^2}{L(M+E[\|\tilde{g}_k\|_2]^2)}$, we know that M (the upper bound of the variance) must be 0.

② $\alpha_k = \frac{1}{2L} = \frac{\|g_k\|_2^2}{L(M+E[\|\tilde{g}_k\|_2]^2)}$, we have $M = \|g_k\|_2^2$ for α_k to be optimal

③ $\alpha_k = \frac{1}{10L} = \frac{\|g_k\|_2^2}{L(M+E[\|\tilde{g}_k\|_2]^2)}$, we have $M = 9\|g_k\|_2^2$ for α_k to be optimal

In general, if $\alpha_k = \frac{1}{cL}$, the variance is bounded to be $\le (c-1)\|g_k\|_2^2$ for α_k to be optimal.