Optimization Methods in Machine Learning HW3

Mmesomachi Nwachukwu

November 2024

1 Theoretical Tasks

Problem 2

Given $\eta \in \mathbb{R}^d$ a random vector and $c \in \mathbb{R}^d$ a nonrandom/deterministic vector, resolve

$$\min_{c \in \mathbb{R}^d} \mathbb{E}[\|\eta - c\|^2]$$

Solution: From theorem 29, in lecture notes, we have that

$$\mathbb{E}[\|\eta - c\|^2] = \mathbb{E}[\|\eta - \mathbb{E}[\eta]\|^2] + \|\mathbb{E}[\eta] - c\|^2$$

Now, $\|\mathbb{E}[\eta] - c\|^2 \geq 0$ (all squares of real numbers are nonnegative) with equality holding when $c = \mathbb{E}[\eta]$. Thus, setting $c = \mathbb{E}[\eta]$, we have that,

$$\mathbb{E}[\|\eta - c\|^2] = \mathbb{E}[\|\eta - \mathbb{E}\|\eta\|^2] + \|\mathbb{E}[\eta] - c\|^2 \geq \mathbb{E}[\|\eta - \mathbb{E}\|\eta\|^2]$$

Since, η, is not deterministic and random, and we have shown that there's a c, satisfying the case of equality, thus we have,

$$\min_{c \in \mathbb{R}^d} \mathbb{E}[\|\eta - c\|^2] = \mathbb{E}[\|\eta - \mathbb{E}\|\eta\|^2]$$

Problem 3

Prove that

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j=1}^{B} \nabla f(x; \xi_{k,j}) - \nabla f(x)\right\|^2\right] \leq \frac{\sigma^2}{B}$$

Solution: From the identity,

$$\mathbb{E}[\|\eta - c\|^2] = \mathbb{E}[\|\eta - \mathbb{E}[\eta]\|^2] + \|\mathbb{E}[\eta] - c\|^2$$

set $\eta = \frac{1}{B}\sum_{j=1}^{B} \nabla f(x; \xi_{k,j})$ and $c = \nabla f(x)$, thus from

$$\mathbb{E}\left[\frac{1}{B}\sum_{j=1}^{B} \nabla f(x; \xi_{k,j})\right] = \nabla f(x)$$

we have that,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j=1}^{B}\nabla f(x;\xi_{k,j}) - \nabla f(x)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{j=1}^{B}\nabla f(x;\xi_{k,j}) - \mathbb{E}\left[\frac{1}{B}\sum_{j=1}^{B}\nabla f(x;\xi_{k,j})\right]\right\|^2\right] \\
&= \frac{1}{B^2}\mathbb{E}\left[\left\|\sum_{j=1}^{B}(\nabla f(x;\xi_{k,j}) - \mathbb{E}[\nabla f(x;\xi_{k,j})])\right\|^2\right] \\
&\text{Due to convexity of } \|\cdot\|^2 \text{ we can apply Jensen's Inequality} \\
&\le \frac{1}{B^2}\mathbb{E}\left[\sum_{j=1}^{B}\|\nabla f(x;\xi_{k,j}) - \mathbb{E}[\nabla f(x;\xi_{k,j})]\|^2\right] \\
&= \frac{1}{B^2}\left[\sum_{j=1}^{B}\mathbb{E}\|\nabla f(x;\xi_{k,j}) - \mathbb{E}[\nabla f(x;\xi_{k,j})]\|^2\right] \\
&\le \frac{1}{B^2}\left[\sum_{j=1}^{B}\mathbb{E}\|\nabla f(x;\xi_{k,j}) - \mathbb{E}[\nabla f(x;\xi_{k,j})]\|^2\right] \\
&= \frac{1}{B^2}\left[\sum_{j=1}^{B}\sigma^2\right] \\
&= \frac{\sigma^2 B}{B^2} \\
&= \frac{\sigma^2}{B}
\end{aligned}$$

Thus, we have that,

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j=1}^{B}\nabla f(x;\xi_{k,j}) - \nabla f(x)\right\|^2\right] \le \frac{\sigma^2}{B}$$

Problem 4 Adapt,

$$\frac{1}{T}\sum_{k=0}^{T-1}\left[\mathbb{E}\left\|\nabla f(x^k)\right\|^2\right] \le \frac{2\Delta}{\gamma T} + L\gamma\sigma^2$$

subject to,

$$\mathbb{E}\left[\|\nabla f(x;\xi) - \nabla f(x)\|^2\right] \le \sigma^2 + B\|\nabla f(x)\|^2$$

Solution: First, we follow verbatim the proof from lecture note until we have to apply this new inequality.

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&\text{From SGD update rule} \\
&= f(x^k) - \gamma \langle \nabla(x^k;\xi_k), \nabla f(x^k)\rangle + \frac{L\gamma^2}{2}\|\nabla f(x^k;\xi_k)\|^2 \\
\Rightarrow \mathbb{E}_k[f(x^{k+1}) &\leq f(x^k) - \gamma\|\nabla f(x^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k\|\nabla f(x^k;\xi_k)\|^2 \\
&\text{From variance decomposition equality} \\
&= f(x^k) - \gamma\|\nabla f(x^k)\|^2 + \frac{L\gamma^2}{2}\|\nabla f(x^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k\|\nabla f(x^k;\xi_k) - \nabla f(x^k)\|^2 \\
&\text{Taking the full expectation and applying new inequality} \\
\Rightarrow \mathbb{E}[f(x^{k+1})] &\leq \mathbb{E}[f(x^k)] + (-\gamma + \frac{L\gamma^2}{2} + \frac{BL\gamma^2}{2})\mathbb{E}[\|\nabla f(x^k)\|^2] + \frac{L\gamma^2\sigma^2}{2} \\
&\text{Summing from } k=0 \text{ to } k = T-1 \\
\Rightarrow \mathbb{E}[f(x^T)] &\leq \mathbb{E}[f(x^0)] + (-\gamma + \frac{L\gamma^2}{2} + \frac{BL\gamma^2}{2})\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla f(x^k)\|^2] + \frac{L\gamma^2\sigma^2 T}{2} \\
\Rightarrow \sum_{k=0}^{T-1}\mathbb{E}[\|\nabla f(x^k)\|^2] &\leq \frac{f(x^0) - \mathbb{E}[f(x^T)]}{(\gamma - \frac{L\gamma^2}{2} - \frac{BL\gamma^2}{2})} + \frac{L\gamma^2\sigma^2 T}{2(\gamma - \frac{L\gamma^2}{2} - \frac{BL\gamma^2}{2})} \leq \frac{\Delta}{(\gamma - \frac{L\gamma^2}{2} - \frac{BL\gamma^2}{2})} + \frac{L\gamma^2\sigma^2 T}{2(\gamma - \frac{L\gamma^2}{2} - \frac{BL\gamma^2}{2})} \\
\Rightarrow \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla f(x^k)\|^2] &\leq \frac{\Delta}{T(\gamma - \frac{L\gamma^2}{2} - \frac{BL\gamma^2}{2})} + \frac{L\gamma\sigma^2}{2(1 - \frac{L\gamma}{2} - \frac{BL\gamma}{2})}
\end{aligned}
$$

Take $\gamma \leq \frac{1}{L(B+1)} < \frac{1}{L}$, $\Rightarrow (1 - \frac{L\gamma}{2} - \frac{BL\gamma}{2}) \geq \frac{1}{2}$,

$$\Rightarrow \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2\Delta}{T\gamma^2} + L\gamma\sigma^2$$