Homework 2

Problem 1 We have seen in class that for strongly convex and Lipschitz smooth functions, the convergence of the gradient descent algorithm is exponentially fast; whereas for convex and Lipschitz continuous functions, the convergence rate is $\widetilde{O}(1/\sqrt{K})$. You must have wondered which factor is more important in contributing to the faster convergence in the first case, the strong convexity or the smoothness. Now let's explore this problem by considering an objective function that is convex (but not strongly convex) and β-Lipschitz smooth.

1. First prove what we left off in class that if a function $\mathcal{L}$ is convex and β-Lipschitz smooth (i.e., $\mathcal{L}(x) - \mathcal{L}(y) \leq \langle \nabla \mathcal{L}(y), x - y \rangle + \frac{\beta}{2}\|x - y\|^2$), then

$$\mathcal{L}(x) - \mathcal{L}(y) \geq \langle \nabla \mathcal{L}(y), x - y \rangle + \frac{1}{2\beta}\|\nabla \mathcal{L}(x) - \mathcal{L}(y)\|^2.$$

2. What's the convergence rate of gradient descent over such a function $\mathcal{L}$ that is convex (but not strongly convex) and β-Lipschitz smooth? Choose a set of step sizes h_k and prove the result using the technique we learned.

Problem 2 Consider the following stochastic gradient descent algorithm.

Algorithm 1: Stochastic gradient descent algorithm

Input: θ_0
for $k = 0, \ldots, K-1$ **do**
  Sample an i.i.d. set S_k uniformly at random from $\{1, \ldots, N\}$
  Compute stochastic gradient $\widetilde{\mathcal{L}}(\theta_k|S_k) = \frac{1}{|S_k|}\sum_{i \in S_k} \nabla \text{loss}(\theta_k|z_i)$
  Update $\theta_{k+1} = \theta_k - h_k \nabla \widetilde{\mathcal{L}}(\theta_k|S_k)$
end
Return: θ_K

Assume that the loss functions $\text{loss}(\theta|z)$ are convex and L-Lipschitz continuous ($\|\nabla \text{loss}(\theta|z)\| \leq L$) for all data points z (No need to make assumptions about the stochastic gradient variance in this case). What's the convergence rate of the stochastic gradient descent algorithm over risk $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \nabla \text{loss}(\theta|z_i)$? Choose a set of step sizes h_k and prove the result using the technique we learned.

Using your derivation, what's the optimal choice of the batch size $n = |S_k|$? Is the stochastic gradient descent algorithm more efficient than gradient descent in this case?

Date : / /
Page No.

Machine Learning
Assignment - 02

Problem 1

1. To prove :-

$$L(x) - L(y) \geq \langle \nabla L(y), x-y \rangle + \frac{1}{2\beta} \|\nabla L(x) - L(y)\|^2$$

Now, we know that if a function is convex :-

$$L(z) - L(y) \geq \nabla L(y)^T (z-y)$$

Condition for Lipschitz smoothness :-

$$L(x) - L(z) \geq \nabla L(x)^T (x-z) - \frac{\beta}{2}\|x-z\|^2$$

$$(1) \Rightarrow L(x) - L(z) + L(z) - L(y) \geq \nabla L(y)^T (z-y) + \nabla L(x)^T (x-z) - \frac{\beta}{2}\|x-z\|^2$$

To choose z, Differentiating the right side w.r.t z and equating to 0

$$0 = \nabla L(y) - \nabla L(x) - \beta(x-z)$$

$$z = x + \frac{1}{\beta}(\nabla L(x) - \nabla L(y))$$

Substituting z in (1)

$$L.H.S \geq \nabla L(y)^T \left(x + \frac{1}{\beta}(\nabla L(x) - \nabla L(y)) - y\right) +$$

$$\nabla L(x)^T \left(\cancel{x} - \cancel{x} - \frac{1}{\beta}(\nabla L(x) - \nabla L(y))\right) - \frac{\beta}{2}\|x-z\|^2$$

Date : / /

Page No.

$$\geq \nabla L(y)^T(x-y) + \frac{1}{\beta}\nabla L(y)^T(\nabla L(x) - \nabla L(y))$$

$$+ \frac{1}{\beta}\nabla L(x)^T(\nabla L(x) - \nabla L(y))$$

$$- \frac{\beta}{2}\left\| \cancel{x} - \cancel{x} - \frac{1}{\beta}(\nabla L(x) - L(y))\right\|^2$$

$$\geq \nabla L(y)^T(x-y) + \frac{1}{\beta}(\nabla L(x) - \nabla L(y))(\nabla L(y)^T - \nabla L^T(x))$$

$$- \frac{1}{2\beta}\|\nabla L(x) - L(y)\|^2$$

$$\geq \nabla L(y)^T(x-y) + \frac{1}{\beta}\|\nabla L(x) - L(y)\|^2 -$$

$$\frac{1}{2\beta}\|\nabla L(x) - L(y)\|^2$$

$$\geq \nabla L(y)^T(x-y) + \frac{1}{2\beta}\|\nabla L(x) - L(y)\|^2$$

Hence Proved

Problem 1

from β - Lipschitz Smooth eq^n :-

2. $$L(x) - L(y) \leq \langle \nabla L(y), (x-y) \rangle + \frac{\beta}{2} \|x-y\|^2$$

$$\Rightarrow L(x_{t+1}) - L(x_t) \leq \langle \nabla L(x_t), (x_{t+1} - x_t) \rangle + \frac{\beta}{2} \|x_{t+1} - x_t\|^2 \quad (1)$$

Differentiating R.H.S w.r.t x_{t+1} & equate to 0

$$\Rightarrow 0 = \nabla L(x_t) + \beta \|x_{t+1} - x_t\|$$

G.D eqn :-

$$\Rightarrow x_{t+1} = x_t - \frac{1}{\beta} \nabla L(x_t) \Rightarrow x_{t+1} - x_t = -\frac{1}{\beta} \nabla L(x_t) \quad (2)$$

So we will make most progress when

$$h_t = \frac{1}{\beta}$$

$$\Rightarrow L(x_{t+1}) - L(x_t) \leq \langle \nabla L(x_t), (x_{t+1} - x_t) \rangle + \frac{\beta}{2} \|x_{t+1} - x_t\|^2 \quad (3)$$

Now, substituting equation (2) in (3)

$$L(x_t) - L(x_{t+1}) \geq \frac{1}{2\beta} \|\nabla L(x_t)\|^2 \quad - (4)$$

$$(L(x_t) - L(x^*))^2 \leq (\nabla L(x_t)^T (x_t - x^*))^2 \quad \text{[squaring convexity eqn]}$$

$$(L(x_t) - L(x^*))^2 \leq \|\nabla L(x_t)\|^2 \|x_t - x^*\|^2$$

Date : / /
Page No.

$$\frac{(L(x_t) - L(x^*))^2}{\|x_t - x^*\|^2} \le \|\nabla L(x_t)\|^2$$

Substituting (4) in above eqn

$$\frac{(L(x_t) - L(x^*))^2}{\|x_t - x^*\|^2} \le 2\beta\,[L(x_t) - L(x_{t+1})]$$

Because $\|x_t - x^*\| \le \|x_0 - x^*\|$

$$\frac{(L(x_t) - L(x^*))^2}{\|x_0 - x^*\|^2} \le 2\beta\,[L(x_t) - L(x_{t+1})]$$

Put $a_t = L(x_t) - L(x^*)$

$a_{t+1} = L(x_{t+1}) - L(x^*)$

$$\frac{(a_t)^2}{\|x_0 - x^*\|^2} \le 2\beta\,(a_t - a_{t+1})$$

Divide the whole inequality by $a_t\, a_{t+1}$

$$\frac{a_t}{a_{t+1}\,\|x_0 - x^*\|^2} \le 2\beta\left(\frac{1}{a_{t+1}} - \frac{1}{a_t}\right)$$

$$\frac{1}{2\beta}\,\frac{a_t}{a_{t+1}\,\|x_0 - x^*\|^2} \le \frac{1}{a_{t+1}} - \frac{1}{a_t}$$

Because $\dfrac{a_t}{a_{t+1}} = \dfrac{L(x_t) - L(x^*)}{L(x_{t+1}) - L(x^*)} > 1$

Date: / /
Page No.

Multiply R.H.S by $\frac{a_{k+1}}{a_k} < 1$ - inequality remains same

$$\sum_{t=0}^{t=T-1} \frac{1}{a_{t+1}} - \frac{1}{a_t} \geq \sum_{t=0}^{T-1} \frac{1}{2\beta} \frac{1}{\|x_0 - x^*\|^2}$$

$$\frac{1}{a_T} - \frac{1}{a_0} \geq (T) \frac{1}{2\beta} \cdot \frac{1}{\|x_0 - x^*\|^2}$$

Ignoring $\frac{1}{a_0}$ on L.H.S

$$\frac{1}{a_T} \geq \frac{T}{2\beta \|x_0 - x^*\|^2}$$

$$\boxed{L(x_T) - L(x^*) \leq \frac{2\beta \|x_0 - x^*\|^2}{T}}$$

It converges $O\left(\frac{1}{T}\right)$ with step size $\frac{1}{\beta}$

Date: / /
Page No.

Problem 2

$\mathcal{L}(\theta|z)$ ~ Convex L-Lipschitz continuous for all point z

$\mathbb{E}_{z\sim 0} \|\nabla L(\theta_k|C_k\|^2 \leq b^2$ for some constant b^2 variance bounded — (1)

$$\|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^*\|^2 - 2h_k \langle \nabla \mathcal{L}(\theta_k)\theta_k - \theta^* \rangle$$
$$h_k^2 \|\nabla L(\theta_k)\|^2$$

$$\langle \nabla y(\theta_k), \theta_k - \theta^* \rangle = \langle \frac{1}{|S_k|} \sum_{i \in S_k} \nabla loss(\theta_k | z_i), \theta_k - \theta^* \rangle$$

$$= \frac{1}{|S_k|} \langle \sum_{i \in S_k} \nabla loss(\theta_k | z_i), \theta_k - \theta^* \rangle$$

$$\Rightarrow \langle \sum_{i \in z_i} \nabla loss(\theta_k | z_i), \theta_k - \theta^* \rangle \leq \sum_{i \in z} loss(\theta_k | z_i) - \sum_{i \in z_i} loss(\theta^* | z_i)$$

$$\Rightarrow \frac{1}{S_k} \langle \nabla L(\theta_k), \theta_k - \theta^* \rangle \leq \frac{1}{|S_k|} \sum_{i \in z_i} loss(\theta_k | z_i) - \frac{1}{S_k} \sum_{i \in z_i} loss(\theta^* | z_i)$$

$$\Rightarrow \|\nabla \mathcal{L}(\theta_k)\|^2 = \left\|\frac{1}{S_k} \sum_{i \in S_k} \nabla loss(\theta_k | z_i)\right\|^2 \leq \frac{1}{|S_k|^2} \left(\sum_{i \in S_i} \nabla loss(\theta_k | z_i)\right)^2$$

$\Rightarrow$ Since $loss(\theta(z_i))$ is convex and L-Lipschitz continuous

$$\|\nabla loss(\theta | z)\| \leq L$$

$$\|\nabla \mathcal{L}(\theta_k | z_i)\|^2 \leq \frac{1}{\cancel{|S_k|}^2} \cdot \cancel{|S_k|}^2 \cdot L^2 = L^2$$

Date: / /
Page No.

$$\Rightarrow \|\theta_{k+1} - \theta^*\|^2 \le \|\theta_k - \theta^*\|^2 - 2h_k \,|\, \frac{1}{S_K} \left(\text{loss}(\theta_k|z_i) - \text{loss}(\theta^*|z_i)\right) + (h_t L)^2$$

$$\Rightarrow \|\theta_{k+1} - \theta^*\|^2 - \|\theta_k - \theta^*\|^2 - (h_k L)^2 \le$$

$$- 2h_k \left[\frac{1}{|S_k|} \sum_{i \in Z_i} \text{loss}(\theta_k|z_i) - \frac{1}{|S_k|} \sum_{i \in Z_i} \text{loss}(\theta^*|z_i)\right]$$

$$\Rightarrow 2h_k \left[\frac{1}{|S_k|} \sum_{i \in Z} \text{loss}(\theta_k|z_i) - \frac{1}{|S_k|} \sum_{i \in Z} \text{loss}(\theta^*|z_i)\right]$$

$$\le \|\theta_k - \theta^*\|^2 - \|\theta_{k+1} - \theta^*\|^2 + (h_k L)^2$$

Take expectation

$$\Rightarrow 2h_k \left[\mathbb{E}\left[\frac{1}{S_k} \sum_{i \in Z_i} \text{loss}(\theta_k|z_i)\right] - \mathbb{E}\left[\frac{1}{|S_k|} \sum_{i \in Z_i} \text{loss}(\theta_k|z_i)\right]\right]$$

$$\le \mathbb{E}\|\theta_k - \theta^*\|^2 - \mathbb{E}\|\theta_{k+1} - \theta^*\|^2 + (h_k L)^2$$

$$\Rightarrow 2h_k \left[\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)\right] \le \mathbb{E}\|\theta_k - \theta^*\|^2 - \mathbb{E}\|\theta_{k+1} - \theta^*\|^2 + (h_k L)^2$$

Take expectation of θ_k

$$\Rightarrow 2h_k \mathop{\mathbb{E}}_{\theta_k} \left[\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)\right] \le \mathop{\mathbb{E}}_{\theta_k} \|\theta_k - \theta^*\|^2 - \mathbb{E}\|\theta_k - \theta^*\|^2 + (h_k L)^2$$

$$\Rightarrow 2 \sum_{k=0}^{k+1} h_k \, \mathbb{E}_{\theta_k} \left[L(\theta_k) - L(\theta^*) \right] \leq \mathbb{E}\|\theta_0 - \theta^*\|^2 + L^2 \sum_{k=0}^{k-1} h_k^2$$

$$\Rightarrow \sum_{k=0}^{k+1} h_k \frac{\mathbb{E}[L(\theta_k)] - \mathbb{E}[L(\theta^*)]}{\sum_{i=0}^{k-1} h_i} \leq \frac{1}{2} \frac{\mathbb{E}\|\theta_0 - \theta^*\|^2}{\sum_{i=0}^{k-1} h_i} - \frac{1}{2} \frac{\mathbb{E}\|\theta_k - \theta^*\|^2}{\sum_{i=0}^{k-1} h_i} + \frac{L^2}{2} \frac{\sum_{k=0}^{k-1} h_k}{\sum_{i=0}^{k-1} h_i}$$

Since $\mathbb{E}\|\theta_k - \theta^*\|^2 \sum_{i=0}^{k-1} h_i > 0$

$$\sum_{k=0}^{k+1} h_k \frac{\mathbb{E}[L(\theta_k)] - \mathbb{E}[L(\theta^*)]}{\sum_{i=0}^{k-1} h_i} \leq \frac{\mathbb{E}\|\theta_0 - \theta^*\|^2}{\sum_{i=0}^{k-1} h_i} + \frac{L^2}{2} \frac{\sum_{k=0}^{k-1} h_k^2}{\sum_{i=0}^{k-1/2} h_i}$$

Let $h_k = \sqrt{\mathbb{E}\|\theta_0 - \theta^*\|^2} \,/\, L\sqrt{k+1}$

$$\Rightarrow \frac{E\|\theta_0-\theta^*\|^2}{2\sum_{i=0}^{k-1} h_i} + \frac{L^2}{2} \frac{\sum_{k=0}^{k-1} h_k}{\sum_{i=0}^{k-1} h_i} \leq \frac{L}{2}\sqrt{E\|\theta_0-\theta^*\|^2} \left(\frac{1+\sum_{k=1}^{k} 1/k}{\sum_{i=1}^{k} \frac{1}{\sqrt{k+1}}} \right)$$

$$\Rightarrow \leq \frac{L}{2}\sqrt{E\|\theta_0-\theta^*\|^2} \left(\frac{1+1+\int_1^k 1/u \, du}{\int_1^k \frac{1}{u} du} \right)$$

$$\leq \frac{L}{2}\sqrt{E\|\theta_0-\theta^*\|^2} \left(\frac{1+1/2 \log(k)}{(\sqrt{k}-1)} \right)$$

Since

$$E[L(\theta)-L(\theta^*)] = E\left[L\left(\sum_{k=0}^{k-1} \frac{h_k}{\sum_{i=0}^{k-1} h_i} \theta_k\right) - L(\theta^*)\right]$$

By convex property →

$$\leq \sum_{k=0}^{k-1} \frac{h_k}{\sum_{i=0}^{k-1} h_i} \left[E[L(\theta_k)] - E[L(\theta^*)] \right]$$

$$\leq \frac{L}{2}\sqrt{E\|\theta_0-\theta^*\|^2} \left(\frac{1+1/2 \log(k)}{\sqrt{k}-1} \right)$$

So convergence rate $\sim \tilde{O}\left(\frac{1}{\sqrt{k}}\right)$

(ii) Since $\mathbb{E}[L(\theta)-L(\theta^*)] \sim \tilde{O}\left(\frac{1}{\sqrt{K}}\right)$, the optimal choice of iid $n = |S_K|$ that optimal choice is $n=1$ with same optimal choice for gradient descent – convergence rate of GD $\sim \tilde{O}\left(\frac{1}{\sqrt{K}}\right)$ but for each iteration, GD have gradient more than SGD $\Rightarrow$ (SGD is more efficient)