Optimization for Data Science, FS22 (Bernd Gärtner and Niao He)
Graded Assignment 2

Student Name, Legi Number: Deepak Narayanan, 21-959-622

Coordinate Descent

Assignment 1

From Definition 5.4 in lecture notes, we have a function $f : \mathbb{R}^d \to \mathbb{R}$ is coordinate-wise smooth with parameter $\mathcal{L} = (L_1, L_2, \dots L_d) \in \mathbb{R}^d_+$ if for each coordinate $i \in [d]$, we have

$$f(\mathbf{x} + \lambda \mathbf{e}_i) \leq f(\mathbf{x}) + \lambda \nabla_i f(\mathbf{x}) + \frac{L_i}{2}\lambda^2$$
$$\forall \mathbf{x} \in \mathbb{R}^d, \lambda \in \mathbb{R}$$

In step 2 of the algorithm given, we have

$$\mathbf{y}_{k,j} = \mathbf{y}_{k,j-1} - \frac{1}{L_j}\nabla_j f(\mathbf{y}_{k,j-1})\mathbf{e}_j$$

Here we can view $-\frac{1}{L_j}\nabla_j f(\mathbf{y}_{k,j-1})\mathbf{e}_j$ as $\lambda \mathbf{e}_j$, where $\lambda = -\frac{1}{L_j}\nabla_j f(\mathbf{y}_{k,j-1})$.
Therefore we have using definition 5.4,

$$\begin{aligned} f(\mathbf{y}_{k,j}) = f(\mathbf{y}_{k,j-1} + \lambda \mathbf{e}_j) = f(\mathbf{y}_{k,j-1} - \frac{1}{L_j}\nabla_j f(\mathbf{y}_{k,j-1})\mathbf{e}_j) \\ \leq f(\mathbf{y}_{k,j-1}) + -\frac{1}{L_j}\nabla_j f(\mathbf{y}_{k,j-1})\nabla_j f(\mathbf{y}_{k,j-1}) + \frac{L_j}{2}\frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{L_j^2} \\ = f(\mathbf{y}_{k,j-1}) - \frac{1}{L_j}\nabla_j f(\mathbf{y}_{k,j-1})^2 + \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2L_j} \\ = f(\mathbf{y}_{k,j-1}) - \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2L_j} \end{aligned}$$

Therefore we have

$$f(\mathbf{y}_{k,j}) \leq f(\mathbf{y}_{k,j-1}) - \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2L_j}$$

This implies

$$f(\mathbf{y}_{k,j-1}) - f(\mathbf{y}_{k,j}) \geq \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2L_j} \tag{1}$$

Now take equation 1, and sum on both sides for $j = 1, 2, \dots d$. We then have

$$\sum_{j=1}^{d} f(\mathbf{y}_{k,j-1}) - f(\mathbf{y}_{k,j}) \geq \sum_{j=1}^{d} \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2L_j}$$

Simplifying,

$$f(\mathbf{y}_{k,0}) - f(\mathbf{y}_{k,d}) \geq \sum_{j=1}^{d} \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2L_j} \geq \sum_{j=1}^{d} \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2\bar{L}}$$
$$\because \bar{L} \geq L_j \forall j \in [d]$$

Using the fact that $f(\mathbf{y}_{k,0}) = \mathbf{x}_k$ and $f(\mathbf{y}_{k,d}) = \mathbf{x}_{k+1}$, we therefore have

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \sum_{j=1}^{d} \frac{\nabla_j f(\mathbf{y}_{k,j-1})^2}{2\bar{L}}$$

Assignment 2

We know that f is $L-$smooth. Therefore from Lemma 3.5, we have

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\| \; \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

Firstly note that the by definition of the $\|.\|_2$, $\|\mathbf{x}\|_2 \geq |x_j| \; \forall j \in [d]$ where x_j is the j^{th} coordinate of $\mathbf{x}$. Therefore we have

$$|\left(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\right)_j| = |\nabla f(\mathbf{y})_j - \nabla f(\mathbf{x})_j| \leq \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|$$

This implies

$$|\nabla f(\mathbf{y})_j - \nabla f(\mathbf{x})_j| \leq L\|\mathbf{y} - \mathbf{x}\| \tag{2}$$

Using triangle inequality, we know that,

$$|A - B| + |B| \geq |A|$$

This implies

$$|\nabla f(\mathbf{y})_j| - |\nabla f(\mathbf{x})_j| \leq |\nabla f(\mathbf{y})_j - \nabla f(\mathbf{x})_j| \tag{3}$$

Putting equations 2 and 3 together, we have,

$$|\nabla f(\mathbf{y})_j| - |\nabla f(\mathbf{x})_j| \leq L\|\mathbf{y} - \mathbf{x}\|$$

Now set $\mathbf{x} = \mathbf{y}_{k,j-1}$, and $\mathbf{y} = \mathbf{x}_k$. We then have,

$$|\nabla f(\mathbf{x}_k)_j| \leq L\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\| + |\nabla f(\mathbf{y}_{k,j-1})_j|$$

Now squaring on both sides, we have

$$\nabla f(\mathbf{x}_k)_j^2 \leq \left(L\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\| + |\nabla f(\mathbf{y}_{k,j-1})_j|\right)^2$$

Note that the j^{th} coordinate of $\nabla f(\mathbf{x}_k)$, denoted by $\nabla f(\mathbf{x}_k)_j$ is the same as the partial derivative of f for the j^{th} dimension, evaluated at $\mathbf{x}_k$. That is

$$\nabla f(\mathbf{x}_k)_j = \nabla_j f(\mathbf{x}_k)$$

Similarly,

$$\nabla f(\mathbf{y}_{k,j-1})_j = \nabla_j f(\mathbf{y}_{k,j-1})$$

Therefore, we have

$$\nabla f(\mathbf{x}_k)_j^2 = \nabla_j f(\mathbf{x}_k)^2 \leq \left(L\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\| + |\nabla f(\mathbf{y}_{k,j-1})_j|\right)^2 = \left(L\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\| + |\nabla_j f(\mathbf{y}_{k,j-1})|\right)^2$$

We also know that

$$(a+b)^2 \leq 2a^2 + 2b^2. \ (\because (a-b)^2 \geq 0, 2ab \leq a^2 + b^2)$$

This implies

$$\nabla_j f(\mathbf{x}_k)^2 \leq 2L^2\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\|^2 + 2\nabla_j f(\mathbf{y}_{k,j-1})^2 \tag{4}$$

Now let us look at the quantity

$$\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\|^2$$

From the update rule, we know the following:

$$\mathbf{y}_{k,i} = \mathbf{y}_{k,i-1} - \frac{1}{L_i}\nabla_i f(\mathbf{y}_{k,i-1})\mathbf{e}_i$$

Applying the update rule repeatedly from $i = 1, 2, \ldots j-1$, we have,

$$\mathbf{y}_{k,j-1} = \mathbf{y}_{k,0} - \sum_{i=1}^{j-1} \frac{1}{L_i}\nabla_i f(\mathbf{y}_{k,i-1})\mathbf{e}_i \implies \mathbf{y}_{k,j-1} = \mathbf{x}_k - \sum_{i=1}^{j-1} \frac{1}{L_i}\nabla_i f(\mathbf{y}_{k,i-1})\mathbf{e}_i$$

Therefore we have,

$$\mathbf{y}_{k,j-1} = \mathbf{x}_k - \sum_{i=1}^{j-1} \frac{1}{L_i}\nabla_i f(\mathbf{y}_{k,i-1})\mathbf{e}_i \implies \mathbf{x}_k - \mathbf{y}_{k,j-1} = \sum_{i=1}^{j-1} \frac{1}{L_i}\nabla_i f(\mathbf{y}_{k,i-1})\mathbf{e}_i$$

This implies

$$\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\|^2 = \left\|\sum_{i=1}^{j-1} \frac{1}{L_i}\nabla_i f(\mathbf{y}_{k,i-1})\mathbf{e}_i\right\|^2 = \sum_{p=1}^{j-1}\sum_{q=1}^{j-1} \frac{1}{L_p.L_q}\nabla_p f(\mathbf{y}_{k,p-1})\nabla_q f(\mathbf{y}_{k,q-1})\mathbf{e}_p^T\mathbf{e}_q$$

In the above double summation if $p \neq q, \mathbf{e}_p^T \mathbf{e}_q = 0$, else if $p = q, \mathbf{e}_p^T \mathbf{e}_q = 1$. Therefore we have

$$\|\mathbf{x}_k - \mathbf{y}_{k,j-1}\|^2 = \sum_{p=1}^{j-1} \sum_{q=1}^{j-1} \frac{1}{L_p.L_q} \nabla_p f(\mathbf{y}_{k,p-1}) \nabla_q f(\mathbf{y}_{k,q-1}) \mathbf{e}_p^T \mathbf{e}_q = \sum_{i=1}^{j-1} \frac{1}{L_i^2} \nabla_i f(\mathbf{y}_{k,i-1})^2 \tag{5}$$

Plugging equation 5 into equation 4, we have

$$\nabla_j f(\mathbf{x}_k)^2 \leq 2L^2 \sum_{i=1}^{j-1} \frac{1}{L_i^2} \nabla_i f(\mathbf{y}_{k,i-1})^2 + 2\nabla_j f(\mathbf{y}_{k,j-1})^2$$

We also know that $\underline{L} \leq L_i$ for $i \in [d]$. Using this we have,

$$\nabla_j f(\mathbf{x}_k)^2 \leq 2L^2 \sum_{i=1}^{j-1} \frac{1}{L_i^2} \nabla_i f(\mathbf{y}_{k,i-1})^2 + 2\nabla_j f(\mathbf{y}_{k,j-1})^2 \leq 2\frac{L^2}{\underline{L}^2} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2 + 2\nabla_j f(\mathbf{y}_{k,j-1})^2$$

Therefore we have,

$$\nabla_j f(\mathbf{x}_k)^2 \leq 2\frac{L^2}{\underline{L}^2} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2 + 2\nabla_j f(\mathbf{y}_{k,j-1})^2 = 2\nabla_j f(\mathbf{y}_{k,j-1})^2 + 2\frac{L^2}{\underline{L}^2} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2$$

Therefore,

$$\nabla_j f(\mathbf{x}_k)^2 \leq 2\nabla_j f(\mathbf{y}_{k,j-1})^2 + \frac{2L^2}{\underline{L}^2} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2 \tag{6}$$

Now, let us first consider, $\nabla_j f(\mathbf{y}_{k,j-1})^2$. From Assignment 1, we know that

$$\nabla_j f(\mathbf{y}_{k,j-1})^2 \leq 2L_j(f(\mathbf{y}_{k,j-1}) - f(\mathbf{y}_{k,j}))$$

Taking summation over $j = 1, 2, 3 \ldots d$, we have,

$$\sum_{j=1}^{d} \nabla_j f(\mathbf{y}_{k,j-1})^2 \leq \sum_{j=1}^{d} 2L_j((f(\mathbf{y}_{k,j-1}) - f(\mathbf{y}_{k,j})) \leq 2\bar{L} \sum_{j=1}^{d}((f(\mathbf{y}_{k,j-1}) - f(\mathbf{y}_{k,j})) = 2\bar{L}(f(\mathbf{y}_{k,0}) - f(\mathbf{y}_{k,d})) \tag{7}$$

Now consider equation 6 and take summation over $j = 1, 2, 3 \ldots d$ on both sides. We then have,

$$\sum_{j=1}^{d} \nabla_j f(\mathbf{x}_k)^2 \leq \sum_{j=1}^{d} 2\nabla_j f(\mathbf{y}_{k,j-1})^2 + \sum_{j=1}^{d} \frac{2L^2}{\underline{L}^2} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2$$

Using Equation 7 above, we further have,

$$\sum_{j=1}^{d} \nabla_j f(\mathbf{x}_k)^2 \leq 2.2\bar{L}(f(\mathbf{y}_{k,0}) - f(\mathbf{y}_{k,d})) + \frac{2L^2}{\underline{L}^2} \sum_{j=1}^{d} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2$$

This implies

$$\sum_{j=1}^{d} \nabla_j f(\mathbf{x}_k)^2 \leq 4\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) + \frac{2L^2}{\underline{L}^2} \sum_{j=1}^{d} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2 \tag{8}$$

Now note that the LHS is just $\|\nabla f(\mathbf{x}_k)\|^2$. Further, observe that

$$\sum_{j=1}^{d} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2 \leq \sum_{j=1}^{d} \sum_{i=1}^{d} \nabla_i f(\mathbf{y}_{k,i-1})^2 = d \sum_{i=1}^{d} \nabla_i f(\mathbf{y}_{k,i-1})^2$$

Now yet again using the result from equation 7, we have

$$d \sum_{i=1}^{d} \nabla_i f(\mathbf{y}_{k,i-1})^2 \leq d2\bar{L}(f(\mathbf{y}_{k,0}) - f(\mathbf{y}_{k,d})) = 2d\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}))$$

Plugging this into equation 8, along with the fact about the LHS, we have

$$\begin{aligned} \|\nabla f(\mathbf{x}_k)\|^2 \leq 4\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) + \frac{2L^2}{\underline{L}^2} \sum_{j=1}^{d} \sum_{i=1}^{j-1} \nabla_i f(\mathbf{y}_{k,i-1})^2 \leq \\ 4\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) + \frac{2L^2}{\underline{L}^2} .2d\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) = \\ 4\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) + 4\bar{L}\frac{L^2 d}{\underline{L}^2}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) = \\ 4\bar{L}\left(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})\right)(1 + \frac{L^2 d}{\underline{L}^2}) \end{aligned}$$

This implies,

$$\|\nabla f(\mathbf{x}_k)\|^2 \leq 4\bar{L}\left(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})\right)(1 + \frac{L^2 d}{\underline{L}^2})$$

This implies,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \|\nabla f(\mathbf{x}_k)\|^2 \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right)}$$

Assignment 3

We know that f is convex. Therefore by using first order characterization of convexity, we have,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}).$$

Consider $\mathbf{y} = \mathbf{x}^*$. Here $\mathbf{x}^*$ belongs to set of optimal points of f. We then have,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}^* - \mathbf{x}).$$

We know from Cauchy-Schwartz,

$$-\|\nabla f(\mathbf{x})\|\|\mathbf{x}^* - \mathbf{x}\| \leq \nabla f(\mathbf{x})^T(\mathbf{x}^* - \mathbf{x}) \leq \|\nabla f(\mathbf{x})\|\|\mathbf{x}^* - \mathbf{x}\|$$

Using this, we have,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}^* - \mathbf{x}) \geq f(\mathbf{x}) - \|\nabla f(\mathbf{x})\|\|\mathbf{x}^* - \mathbf{x}\|$$

This implies,

$$f(\mathbf{x}^*) - f(\mathbf{x}) \geq -\|\nabla f(\mathbf{x})\|\|\mathbf{x}^* - \mathbf{x}\| \implies f(\mathbf{x}) - f(\mathbf{x}^*) \leq \|\nabla f(\mathbf{x})\|\|\mathbf{x}^* - \mathbf{x}\|$$

This implies

$$\|\nabla f(\mathbf{x})\| \geq \frac{f(\mathbf{x}) - f(\mathbf{x}^*)}{\|\mathbf{x}^* - \mathbf{x}\|}$$

Now note that we start with $\mathbf{x}_0$. We proved in Assignment 2 that $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \|\nabla f(\mathbf{x}_k)\|^2 \frac{1}{4\bar{L}\left(1+\frac{dL^2}{\underline{L}^2}\right)} \geq 0$. This implies $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \forall k$. This implies starting at $\mathbf{x}_0$, the iterates $\mathbf{x}_i, i = 1, 2, \ldots$, always belong to the sublevel set $f^{\leq \mathbf{x}_0}$. In particular, $\mathbf{x}_k$ belongs to the sublevel. Therefore, plugging in $\mathbf{x} = \mathbf{x}_k$, we have,

$$\|\nabla f(\mathbf{x}_k)\| \geq \frac{f(\mathbf{x}_k) - f(\mathbf{x}^*)}{\|\mathbf{x}^* - \mathbf{x}_k\|}$$

We just showed that we always remain in the sublevel. Using this fact and the fact that maximum distance from any point in the sublevel to a point in the optimal set, we have,

$$\|\nabla f(\mathbf{x}_k)\| \geq \frac{f(\mathbf{x}_k) - f(\mathbf{x}^*)}{\|\mathbf{x}^* - \mathbf{x}_k\|} \geq \frac{f(\mathbf{x}_k) - f(\mathbf{x}^*)}{R}$$

This implies,

$$\|\nabla f(\mathbf{x}_k)\|^2 \geq \left(\frac{f(\mathbf{x}_k) - f(\mathbf{x}^*)}{R}\right)^2$$

Combining this with the result of Assignment 2, we have,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \|\nabla f(\mathbf{x}_k)\|^2 \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right)} \geq \left(\frac{f(\mathbf{x}_k) - f(\mathbf{x}^*)}{R}\right)^2 \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right)}$$

This implies

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \left(\frac{f(\mathbf{x}_k) - f(\mathbf{x}^*)}{R}\right)^2 \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right)}$$

Since $\mathbf{x}^*$ belongs to set of optimal points of f, we have $f(\mathbf{x}^*) = f^*$. Therefore, we have,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right) R^2} (f(\mathbf{x}_k) - f^*)^2.$$

Assignment 4

We firstly know that f is $L-$smooth. Using definition of smoothness, we have,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Consider $\mathbf{x} = \mathbf{x}^*$, where $\mathbf{x}^*$ belongs to set of optimal points for f (that is $f(\mathbf{x}^*) = f^*$), and $\mathbf{y} = \mathbf{x}_0$. This implies,

$$f(\mathbf{x}_0) - f(\mathbf{x}^*) = f(\mathbf{x}_0) - f^* \leq \nabla f(\mathbf{x}^*)^T (\mathbf{x}_0 - \mathbf{x}^*) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Since $\mathbf{x}^*$ is a minimizer (since f is convex, it is a global minimizer indeed), and we consider whole of $\mathbb{R}^d$ as domain, $\nabla f(\mathbf{x}^*) = \mathbf{0}$ by Lemma 2.22. This implies,

$$f(\mathbf{x}_0) - f^* \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \frac{L}{2}R^2$$

where we make use of the fact from Assignment 3 about sublevel sets. Now define, $a_k = f(\mathbf{x}_k) - f^*$. Also observe that

$$a_k - a_{k+1} = f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right) R^2} (f(\mathbf{x}_k) - f^*)^2 = \gamma a_k^2$$

where we used results from Assignment 3 and set $\gamma = \frac{1}{4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right) R^2}$. We need $a_0 \leq \frac{1}{m\gamma}$, for $m > 0$. Observe that the LHS in the smoothness equation is a_0, we have,

$$a_0 \leq \frac{LR^2}{2}$$

A stronger requirement would be to require (this is stronger as we may have m, such that $a_0 \leq \frac{1}{m\gamma} < \frac{LR^2}{2}$),

$$a_0 \leq \frac{LR^2}{2} \leq \frac{1}{m\gamma} \iff m \leq \frac{2}{\gamma LR^2}$$

In particular, we need to show that $m = \frac{8}{d}$ satisfies the above inequality for our question, That would imply, we will have $a_0 \leq \frac{1}{m\gamma}$, $a_k - a_{k+1} \geq \gamma a_k^2$ (from Assignment 3), and thus can conclude that

$$a_k = f(\mathbf{x}_k) - f^* \leq 4\bar{L}(1 + \frac{dL^2}{\underline{L}^2})R^2 \left(\frac{1}{k + \frac{8}{d}}\right) = \frac{1}{\gamma(k+m)}$$

Claim: $m = \frac{8}{d}$ satisfies the above inequality.
Proof: Observe that since $d \geq 1$

$$\frac{1}{d} \leq d + \frac{\underline{L}}{L} = \frac{Ld + \underline{L}}{L} = \frac{L\underline{L}d + \underline{L}^2}{L\underline{L}} \leq \frac{L^2 d + \underline{L}^2}{L\underline{L}} = \frac{\underline{L}}{L\underline{L}^2}\left(dL^2 + \underline{L}^2\right) \leq$$
$$\frac{\bar{L}}{L\underline{L}^2}\left(dL^2 + \underline{L}^2\right) = \frac{\bar{L}}{L}\left(1 + \frac{dL^2}{\underline{L}}\right)$$

Above we use the fact the since f is $L-$smooth, it should also be coordinatewise L smooth which implies $\underline{L} \leq L, \bar{L} \leq L$. Thus we have,

$$\frac{1}{d} \leq \frac{\bar{L}}{L}\left(1 + \frac{dL^2}{\underline{L}}\right) \implies \frac{8}{d} \leq \frac{8\bar{L}}{L}\left(1 + \frac{dL^2}{\underline{L}}\right) = \frac{8\bar{L}R^2}{LR^2}\left(1 + \frac{dL^2}{\underline{L}}\right) =$$
$$\frac{2}{LR^2}\left(4\bar{L}\left(1 + \frac{dL^2}{\underline{L}^2}\right)R^2\right) = \frac{2}{LR^2\gamma}$$

This implies $\frac{8}{d} \leq \frac{2}{LR^2\gamma}$ Therefore, we have $m = \frac{8}{d}, m > 0$ indeed satisfies the requirement for $a_0 \leq \frac{1}{m\gamma}$. And we have already shown that $a_k - a_{k+1} \geq \gamma a_k^2$ from assignment 3. Therefore, indeed we have,

$$a_k = f(\mathbf{x}_k) - f* \leq 4\bar{L}(1 + \frac{dL^2}{\underline{L}^2})R^2 \left(\frac{1}{k + \frac{8}{d}}\right) = \frac{1}{\gamma(k+m)}$$

Projected Subgradient Descent

Assignment 5

We are given the following relation.

$$\forall \mathbf{y} \in \mathbf{X}, \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 \leq \|\mathbf{x}_k - \mathbf{y}\|^2 - 2\alpha_k(f(\mathbf{x}_k) - f(\mathbf{y})) + \alpha_k^2\|\mathbf{s}_k\|^2$$

Plugging in value for $\|\mathbf{x}_k - \mathbf{y}\|$

$$\begin{aligned}\forall \mathbf{y} \in \mathbf{X}, \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 \leq \|\mathbf{x}_{k-1} - \mathbf{y}\|^2 - 2\alpha_k(f(\mathbf{x}_k) - f(\mathbf{y}))\\ -2\alpha_{k-1}(f(\mathbf{x}_{k-1}) - f(\mathbf{y})) + \alpha_k^2\|\mathbf{s}_k\|^2 + \alpha_{k-1}^2\|\mathbf{s}_{k-1}\|^2\end{aligned}$$

Taking Induction, we then have,

$$\|\mathbf{x}_{k+1} - \mathbf{y}\|^2 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2\alpha_i(f(\mathbf{x}_i) - f(\mathbf{y})) + \sum_{i=0}^{k} \alpha_i^2\|\mathbf{s}_i\|^2$$

Since $\|\mathbf{x}_{k+1} - \mathbf{y}\|^2 \geq 0$, we have

$$0 \leq \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2\alpha_i(f(\mathbf{x}_i) - f(\mathbf{y})) + \sum_{i=0}^{k} \alpha_i^2\|\mathbf{s}_i\|^2$$

For contradiction, assume the following.

$$\inf_{k=0,1,2,\ldots\infty} f(\mathbf{x}_k) \geq f^* + \delta + \epsilon \text{ for } \epsilon \geq 0$$

This implies that all of the elements of the sequence $\{\mathbf{x}_k\}$is at least $f^*+\delta+\epsilon$. This implies none of the iterates $\mathbf{x}_k$ is less than $f^* + \delta + \epsilon$. This implies $\forall k \geq 0, f(\mathbf{x}_k) \geq f^* + \delta + \epsilon$.

Further we are given f is continuous on $\mathbf{X}$. This implies, we have $\mathbf{y} \in \mathbf{X}$, such that $f(\mathbf{y}) = f^* + \epsilon$, where f^* is attained by a minimizer $\mathbf{x}^* \in \mathbf{X}$ of f. Or equivalently, we have, $f(\mathbf{x}_k) \geq f(\mathbf{y}) + \delta$. This is a direct consequence of our assumption. Further we have .

$$\alpha_i := \frac{f(\mathbf{x}_i) - \hat{f}_i}{||\mathbf{s}_i||^2}, \hat{f}_i = \min_{0 \leq j \leq i} f(x_j) - \delta$$

Using $f(\mathbf{x}_k) \geq f(\mathbf{y}) + \delta$, we further get, .

$$\min_{0 \leq j \leq k} f(\mathbf{x}_k) - \delta \geq f(\mathbf{y}) + \delta - \delta \implies f(\mathbf{y}) \leq \hat{f}_k \forall k$$

Now going back to our inequality, we have,

$$0 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2\alpha_i(f(\mathbf{x}_i) - f(\mathbf{y})) + \sum_{i=0}^{k} \alpha_i^2\|\mathbf{s}_i\|^2$$

Using the value of $\mathbf{y}$ that we defined above, and using the relation we derived between $f(\mathbf{y}), \hat{f}_k$, we have

$$0 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2\alpha_i (f(\mathbf{x}_i) - f(\mathbf{y})) + \sum_{i=0}^{k} \alpha_i^2 \|\mathbf{s}_i\|^2 \leq$$
$$\|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2\alpha_i (f(\mathbf{x}_i) - \hat{f}_k) + \sum_{i=0}^{k} \alpha_i^2 \|\mathbf{s}_i\|^2$$

Now plugging in the value of α_i, we have,

$$0 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2 \left(\frac{f(\mathbf{x}_i) - \hat{f}_i}{\|\mathbf{s}_i^2\|} \right) (f(\mathbf{x}_i) - \hat{f}_k) + \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^4} \|\mathbf{s}_i\|^2$$

This implies

$$0 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} 2 \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2} + \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2}$$

This implies

$$0 \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 - \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2}$$

This implies

$$\sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2} \leq \|\mathbf{x}_0 - \mathbf{y}\|^2 < \infty$$

as $\mathbf{x}_0$ and $\mathbf{y}$ are finite. Therefore

$$\lim_{k \to \infty} \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2} < \infty$$

Further we have Lipschitzness of f, which gives us bounded subgradients from Lemma 5.6 from Handout. That is $\|\mathbf{s}_i\| \leq B \forall i$ Therefore we have,

$$\sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2} \geq \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{B^2}$$

We also know that

$$f(\mathbf{x}_i) - \hat{f}_i = f(\mathbf{x}_i) - (\min_{0 \leq j \leq i} f(x_j) - \delta) = f(\mathbf{x}_i) - \min_{0 \leq j \leq i} f(x_j) + \delta \geq \delta$$

This implies,

$$\sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2} \geq \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{B^2} \geq \sum_{i=0}^{k} \frac{\delta^2}{B^2} = (k+1)\frac{\delta^2}{B^2}$$

Now taking limit, we have,

$$\lim_{k\to\infty} \sum_{i=0}^{k} \frac{(f(\mathbf{x}_i) - \hat{f}_i)^2}{\|\mathbf{s}_i\|^2} \geq \lim_{k\to\infty} (k+1)\frac{\delta^2}{B^2} = \infty$$

This is a contradiction to the fact that the same quantity is bounded above! Observe that this result was a consequence of our assumption. Therefore, $\inf_{k=0,1,\ldots,\infty} f(\mathbf{x}_k) \leq f^* + \delta$

Stochastic Gradient Descent

Assignment 6

Claim 1: $F_{m,k}(\mathbf{x}_k)$ is a $\mu-$strongly convex function.
Proof: Since each f_i is strongly convex with parameter μ, we have for any $i \in [n]$,

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}||\mathbf{x}-\mathbf{y}||^2.$$

Therefore for all indices $\{i_k^1, i_k^2, \dots i_k^m\}$ chosen in a minibatch we will further have $\forall j$,

$$f_{i_k^j}(\mathbf{y}) \geq f_{i_k^j}(\mathbf{x}) + \nabla f_{i_k^j}(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}||\mathbf{x}-\mathbf{y}||^2.$$

Taking sum over all j and dividing by m, we have,

$$\frac{1}{m}\sum_{j=1}^{m} f_{i_k^j}(\mathbf{y}) \geq \frac{1}{m}\sum_{j=1}^{m} f_{i_k^j}(\mathbf{x}) + \frac{1}{m}\sum_{j=1}^{m} \nabla f_{i_k^j}(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{1}{m}\sum_{j=1}^{m}\frac{\mu}{2}||\mathbf{x}-\mathbf{y}||^2.$$

This implies,

$$F_{m,k}(\mathbf{y}) \geq F_{m,k}(\mathbf{x}) + \nabla F_{m,k}(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}||\mathbf{x}-\mathbf{y}||^2.$$

Coming to the update rule, we have

$$\delta_{k+1} = \delta_k - \alpha \nabla F_{m,k}(\mathbf{x}_k)$$

Taking norm and squaring, we have,

$$\|\delta_{k+1}\|^2 = \|\delta_k - \alpha\nabla F_{m,k}(\mathbf{x}_k)\|^2 = \|\delta_k\|^2 + \alpha^2\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 - 2\alpha\langle \delta_k, \nabla F_{m,k}(\mathbf{x}_k)\rangle$$

From strong convexity applied to points $\mathbf{x}^*$ and $\mathbf{x}_k$, we have,

$$F_{m,k}(\mathbf{x}^*) \geq F_{m,k}(\mathbf{x}_k) + \nabla F_{m,k}(\mathbf{x}_k)^T(\mathbf{x}^* - \mathbf{x}_k) + \frac{\mu}{2}||\mathbf{x}_k - \mathbf{x}^*||^2.$$

As per question, we have, $f_i(\mathbf{x}^*) = 0 \implies F_{m,k}(\mathbf{x}^*) = 0$. This implies,

$$-\nabla F_{m,k}(\mathbf{x}_k)^T(\mathbf{x}^* - \mathbf{x}_k) \geq \frac{\mu}{2}||\mathbf{x}_k - \mathbf{x}^*||^2 + F_{m,k}(\mathbf{x}_k)$$

This implies,

$$\nabla F_{m,k}(\mathbf{x}_k)^T\delta_k \geq \frac{\mu}{2}||\delta_k||^2 + F_{m,k}(\mathbf{x}_k)$$

This implies

$$-\nabla F_{m,k}(\mathbf{x}_k)^T\delta_k \leq -\frac{\mu}{2}||\delta_k||^2 - F_{m,k}(\mathbf{x}_k)$$

Plugging this into our equation, we have,

$$\begin{aligned}\|\delta_{k+1}\|^2 &= \|\delta_k\|^2 + \alpha^2\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 - 2\alpha\langle\delta_k, \nabla F_{m,k}(\mathbf{x}_k)\rangle \\ &\leq \|\delta_k\|^2 + \alpha^2\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 + 2\alpha(-\frac{\mu}{2}||\delta_k||^2 - F_{m,k}(\mathbf{x}_k))\end{aligned}$$

This implies

$$\begin{aligned}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2 + \alpha^2\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 - 2\alpha F_{m,k}(\mathbf{x}_k) = \\ (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\left(F_{m,k}(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\end{aligned}$$

Now taking expectation with respect to the said variables, we have,

$$\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\|\delta_{k+1}\|^2 \leq \mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\left[(1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\left(F_{m,k}(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\right]$$

By linearity of expectation and by observing that δ_k doesn't depend on the considered random variables, we have,

$$\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\left[\left(F_{m,k}(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\right]$$

$$\begin{aligned}\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}[F_{m,k}(\mathbf{x}_k)] = \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}[f_{i_k^{(j)}}(\mathbf{x}_k)] = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}f_i(\mathbf{x}_k)\frac{1}{n} = \\ \frac{1}{m}\sum_{j=1}^{m}f(\mathbf{x}_k) = f(\mathbf{x}_k)\end{aligned}$$

Here the $\frac{1}{n}$ comes from using a uniform distribution over all the indices. More formally we have,

$$\begin{aligned}\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}[f_{i_k^{(j)}}(\mathbf{x}_k)] = \mathbb{E}_{i_k^{(j)}}[f_{i_k^{(j)}}(\mathbf{x}_k)] = \sum_{p=1}^{n}\mathbb{P}[i_k^{(j)} = p]f_{i_k^{(j)}=p}(\mathbf{x}_k) = \sum_{p=1}^{n}\frac{1}{n}f_p(\mathbf{x}_k) = f(\mathbf{x}_k) \\ \because \mathbb{P}[i_k^{(j)} = p] = \frac{1}{n}\end{aligned}$$

Therefore we have,

$$\begin{aligned}\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\left(f(\mathbf{x}_k) - \frac{\alpha}{2}\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}[\|\nabla F_{m,k}(\mathbf{x}_k)\|^2]\right) = \\ (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\mathbb{E}_{i_k^{(1)},,\ldots,i_k^{(m)}}\left[\left(f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\right]\end{aligned}$$

since expectation of $f(\mathbf{x}_k)$ is itself, since it is a constant with respect to the given random variables. In summary,

$$\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\left[\left(f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\right]$$

Assignment 7

We first consider,

$$\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 = \left\| \sum_{j=1}^{m} \frac{1}{m} \nabla f_{i_k^{(j)}}(\mathbf{x}_k) \right\|^2 = \sum_{p=1}^{m} \sum_{q=1}^{m} \frac{1}{m^2} \nabla f_{i_k^{(p)}}(\mathbf{x}_k)^T \nabla f_{i_k^{(q)}}(\mathbf{x}_k)$$

We can break this double summation into two. First $p = q$, second case, $p \neq q$. Therefore we have

$$\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 = \sum_{p=1}^{m} \frac{1}{m^2} ||\nabla f_{i_k^{(p)}}(\mathbf{x}_k)||^2 + \sum_{p \neq q} \frac{1}{m^2} \nabla f_{i_k^{(p)}}(\mathbf{x}_k)^T \nabla f_{i_k^{(q)}}(\mathbf{x}_k)$$

Now taking expectation with respect to $i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}$, we have, by linearity of expectation,

$$\begin{aligned}\mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}} \|\nabla F_{m,k}(\mathbf{x}_k)\|^2 = \sum_{p=1}^{m} \frac{1}{m^2} \mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}} ||\nabla f_{i_k^{(p)}}(\mathbf{x}_k)||^2 + \\ \sum_{p \neq q} \mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}} \left[\frac{1}{m^2} \nabla f_{i_k^{(p)}}(\mathbf{x}_k)^T \nabla f_{i_k^{(q)}}(\mathbf{x}_k) \right]\end{aligned}$$

We firstly observe that the first term is identical for all the random variables. This is because the we have uniform distribution and irrespective of the index p, we will have the same value. Therefore, we have

$$\sum_{p=1}^{m} \frac{1}{m^2} \mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}} ||\nabla f_{i_k^{(p)}}(\mathbf{x}_k)||^2 = \frac{1}{m^2} \sum_{p=1}^{m} \mathbb{E}_{i_k^{(p)}} ||\nabla f_{i_k^{(p)}}(\mathbf{x}_k)||^2 = \frac{1}{m} \mathbb{E}_{i_k^{(p)}} ||\nabla f_{i_k^{(p)}}(\mathbf{x}_k)||^2$$

Now the last term is precisely $\mathbb{E}_{i_k^{(p)}} ||\nabla f_{i_k^p}(\mathbf{x}_k)||^2 = \mathbb{E}_{i_k^{(1)}} ||\nabla F_{1,k}(\mathbf{x}_k)||^2$ since the expectations will be identical and don't depend on the element picked as the underlying distribution is uniform. Also $F_{m,k}$ with $m = 1$ is standard SGD.

Therefore we have,

$$\mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}} \|\nabla F_{m,k}(\mathbf{x}_k)\|^2 = \frac{1}{m} \mathbb{E}_{i_k^{(1)}} ||\nabla F_{1,k}(\mathbf{x}_k)||^2 + \sum_{p \neq q} \mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}} \left[\frac{1}{m^2} \nabla f_{i_k^{(p)}}(\mathbf{x}_k)^T \nabla f_{i_k^{(q)}}(\mathbf{x}_k) \right]$$

Now since $i_k^{(p)}$ and $i_k^{(q)}$ are independently chosen, the second term expectation can be split into product of expectations, both of which have identical expectations albeit with a transpose. This implies

$$\sum_{p\neq q}\mathbb{E}\left[\frac{1}{m^2}\nabla f_{i_k^{(p)}}(\mathbf{x}_k)^T\nabla f_{i_k^{(q)}}(\mathbf{x}_k)\right] = \frac{1}{m^2}\sum_{p\neq q}\mathbb{E}[\nabla f_{i_k^{(p)}}(\mathbf{x}_k)^T]\mathbb{E}[\nabla f_{i_k^{(q)}}(\mathbf{x}_k)] =$$

$$\frac{1}{m^2}\sum_{p\neq q}\sum_{k=1}^{n}\frac{1}{n}\nabla f_k(\mathbf{x}_k)^T\sum_{l=1}^{n}\frac{1}{n}\nabla f_l(\mathbf{x}_k) =$$

$$\frac{1}{m^2}\sum_{p\neq q}\|\nabla f(\mathbf{x}_k)\|^2 = \frac{m.(m-1)}{m^2}\|\nabla f(\mathbf{x}_k)\|^2 = \frac{m-1}{m}\|\nabla f(\mathbf{x}_k)\|^2$$

In the above, we use,

$$\mathbb{E}[\nabla f_{i_k^{(q)}}(\mathbf{x}_k)] = \sum_{j=1}^{n}\mathbb{P}[i_k^{(q)} = j]\nabla f_{i_k^{(q)}=j}(\mathbf{x}_k) = \sum_{j=1}^{n}\frac{1}{n}\nabla f_j(\mathbf{x}_k) = \nabla f(\mathbf{x})$$

Above, we used independence to get the product of expectations, and then applied definition of expectation as above. Then we observe there are $m(m-1)$ many such terms and thus obtain the result. In the above, $\mathbb{E} := \mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}$.

Therefore, we finally have,

$$\mathbb{E}_{i_k^{(1)},i_k^{(2)},\ldots,i_k^{(m)}}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2 = \frac{1}{m}\mathbb{E}_{i_k^{(1)}}||\nabla F_{1,k}(\mathbf{x}_k)||^2 + \frac{m-1}{m}\|\nabla f(\mathbf{x}_k)\|^2$$

Next, we will consider sufficient decrease for f and f_i. For f, consider the following:

$$\mathbf{x}_t = \mathbf{x}_k - \frac{1}{\lambda}\nabla f(\mathbf{x}_k)$$

By smoothness of f, we have,

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}_t - \mathbf{x}_k) + \frac{\lambda}{2}||\mathbf{x}_t - \mathbf{x}_k||^2$$

Since each $f_i \geq 0$, $f(\mathbf{x}) \geq 0\forall\mathbf{x}$. This implies $f(\mathbf{x}_t) \geq 0$. Using this fact and plugging in the value for $\mathbf{x}_t - \mathbf{x}_k$ from the update rule, we have

$$0 \leq f(\mathbf{x}_t) \leq f(\mathbf{x}_k) + -\frac{1}{\lambda}\|\nabla f(\mathbf{x}_k)\|^2 + \frac{\lambda}{2\lambda^2}\|\nabla f(\mathbf{x}_k)\|^2 = f(\mathbf{x}_k) - \frac{\|\nabla f(\mathbf{x}_k)\|^2}{2\lambda}$$

This implies,

$$f(\mathbf{x}_k) \geq \frac{\|\nabla f(\mathbf{x}_k)\|^2}{2\lambda} - \implies -\|\nabla f(\mathbf{x}_k)\|^2 \geq -2\lambda f(\mathbf{x}_k)$$

Similarly, we have, by smoothness of f_i,

$$-\|\nabla f_i(\mathbf{x}_k)\|^2 \geq -2Lf_i(\mathbf{x}_k)$$

Taking summation on both sides for $i \in [n]$, and dividing by n, we have

$$-\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}_k)\|^2 \geq -\frac{1}{n}\sum_{i=1}^{n} 2Lf_i(\mathbf{x}_k) = -2Lf(\mathbf{x}_k)$$

Now consider:

$$f(\mathbf{x}_k) - \frac{\alpha}{2}\mathbb{E}\nabla\|F_{m,k}(\mathbf{x}_k)\|^2 = f(\mathbf{x}_k) - \frac{\alpha}{2}\left[\frac{1}{m}\mathbb{E}_{i_k^{(1)}}||\nabla F_{1,k}(\mathbf{x}_k)||^2 + \frac{m-1}{m}\|\nabla f(\mathbf{x}_k)\|^2\right]$$

By using the fact that we have a uniform distribution, taking expectation, and using smoothness for f_i, we have,

$$-\mathbb{E}_{i_k^{(1)}}||\nabla F_{1,k}(\mathbf{x}_k)||^2 = -\sum_{p=1}^{n}\mathbb{P}[i_k^{(1)} = p]||\nabla f_{i_k^{(1)}=p}(\mathbf{x}_k)||^2 = -\sum_{p=1}^{n}\frac{1}{n}\|\nabla f_p(\mathbf{x}_k)\|^2 \geq -2Lf(\mathbf{x}_k)$$

Similarly replacing $\|\nabla f(\mathbf{x}_k)\|^2$, we have,

$$f(\mathbf{x}_k) - \frac{\alpha}{2}\left[\frac{1}{m}\mathbb{E}_{i_k^{(1)}}||\nabla F_{1,k}(\mathbf{x}_k)||^2 + \frac{m-1}{m}\|\nabla f(\mathbf{x}_k)\|^2\right] \geq f(\mathbf{x}_k) - \frac{\alpha}{2m}2Lf(\mathbf{x}_k) - \frac{\alpha.(m-1)}{2m}2\lambda f(\mathbf{x}_k)$$

This implies

$$f(\mathbf{x}_k) - \frac{\alpha}{2}\mathbb{E}\nabla\|F_{m,k}(\mathbf{x}_k)\|^2 \geq f(\mathbf{x}_k)\left[1 - \frac{\alpha L}{m} - \frac{\alpha\lambda(m-1)}{m}\right]$$

Since we have, $f(\mathbf{x}_k) \geq 0 \forall k (\because f_i(\mathbf{x}_k) \geq 0 \forall i)$, for LHS to be nonnegative, we need,

$$\left[1 - \frac{\alpha L}{m} - \frac{\alpha\lambda(m-1)}{m}\right] \geq 0 \iff \frac{\alpha L}{m} + \frac{\alpha\lambda(m-1)}{m} \leq 1$$

Now α is defined as:

$$\min\left\{\frac{pm}{L}, \frac{1-p}{\lambda}\frac{m}{m-1}\right\}$$

Suppose $\alpha = \frac{pm}{L}$. We then have

$$\frac{pm}{L} < \frac{(1-p)m}{\lambda(m-1)} \iff (m-1)p < \frac{(1-p)L}{\lambda}$$

Now the term we are concerned with,

$$\frac{\alpha L}{m} + \frac{\alpha\lambda(m-1)}{m} = \frac{p.m.L}{L.m} + \frac{(m-1)pm\lambda}{mL} = p + \frac{\lambda.(m-1)p}{L} < p + 1 - p = 1$$

This implies

$$f(\mathbf{x}_k) - \frac{\alpha}{2}\mathbb{E}\nabla\|F_{m,k}(\mathbf{x}_k)\|^2 \geq f(\mathbf{x}_k)\left[1 - \frac{\alpha L}{m} - \frac{\alpha\lambda(m-1)}{m}\right] \geq 0$$

Hence we are done for this case. Now suppose $\alpha = \frac{1-p}{\lambda}\frac{m}{m-1}$. We then have,

$$\frac{pm}{L} \geq \frac{(1-p)m}{\lambda(m-1)} \iff \frac{p}{L} \geq \frac{(1-p)}{\lambda(m-1)}$$

Once again the term we are concerned with, Now the term we are concerned with,

$$\frac{\alpha L}{m} + \frac{\alpha\lambda(m-1)}{m} = \frac{(1-p)m}{L}\lambda(m-1)m + \frac{(m-1).(1-p).m.\lambda}{m.\lambda.(m-1)} =$$
$$\frac{(1-p)L}{\lambda(m-1)} + (1-p) \leq \frac{pL}{L} + (1-p) = 1$$

Once more we have therefore have,

$$f(\mathbf{x}_k)\left[1 - \frac{\alpha L}{m} - \frac{\alpha\lambda(m-1)}{m}\right] \geq 0$$

Therefore for the given values of α, we indeed have,

$$\mathbb{E}\left[f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right] \geq 0.$$

In the above $\mathbb{E} := \mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}}$.

Assignment 8

In the below $\mathbb{E} := \mathbb{E}_{i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(m)}}$. From Assignment 6, we have,

$$\mathbb{E}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\mathbb{E}\left[\left(f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\right]$$

From assignment 7, we have for the choice of learning rate,

$$\mathbb{E}\left[f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right] \geq 0. \implies -\mathbb{E}\left[f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right] \leq 0.$$

Combining these two facts, we have, for the choice of α defined in Assignment 7 where we also make use of the fact that $\alpha > 0$

$$\mathbb{E}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2 - 2\alpha\mathbb{E}\left[\left(f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla F_{m,k}(\mathbf{x}_k)\|^2\right)\right] \leq (1-\alpha\mu)\|\delta_k\|^2$$

$$\mathbb{E}\|\delta_{k+1}\|^2 \leq (1-\alpha\mu)\|\delta_k\|^2$$

To find optimal step size, we need to minimize $(1-\alpha\mu)$ wrt p. This means we need to maximize $\alpha\mu$, which is the same as maximizing α with respect to p. Yet again, α is defined as:

$$\min\left\{\frac{pm}{L}, \frac{1-p}{\lambda}\frac{m}{m-1}\right\}$$

Note that we are interested in maximizing the above quantity as a function of p. Note that both terms are linear functions of p, with one term increasing as p increases while the other decreases. Therefore, the maximum is attained exactly when both the values are equal, since otherwise as we take the minimum, we will always have a smaller learning rate. This implies, we can get the p we are interested in by equating the two quantities in the expression. Therefore,

$$\frac{pm}{L} = \frac{(1-p)m}{\lambda(m-1)} \implies p(\lambda)(m-1) = L - pL$$
$$\implies p(\lambda(m-1)+L) = L \implies p* = \frac{L}{\lambda(m-1)+L}$$

Therefore the optimal learning rate is $\alpha^* = \frac{pm}{L} = \frac{m}{\lambda(m-1)+L}$

Assignment 9

f is $\lambda-$smooth. Using smoothness on $\mathbf{x}_k$ and $\mathbf{x}^*$, we have,

$$f(\mathbf{x}_k) \leq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T(\mathbf{x}_k - \mathbf{x}^*) + \frac{\lambda}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2$$

At the optimal point $f(\mathbf{x}^*) = 0$, $\nabla f(\mathbf{x}^*) = 0$. This follows from Lemma 2.22. Therefore, we have,

$$f(\mathbf{x}_k) \leq \frac{\lambda}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 = \frac{\lambda}{2}\|\delta_k\|^2$$

Therefore,

$$f(\mathbf{x}_k) \leq \frac{\lambda}{2}\|\delta_k\|^2$$

We are interested in $\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)]$. The randomness in $\mathbf{x}_k$ comes from the randomness of all the minibatches used for generating the iterate $\mathbf{x}_k$. We use the law of total expectation below.

$$\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)] = \mathbb{E}_{\mathbf{x}_{k-1}}\mathbb{E}_{\mathbf{x}_k|\mathbf{x}_{k-1}}[f(\mathbf{x}_k)]$$

Observe that the randomness in $\mathbf{x}_k|\mathbf{x}_{k-1}$ comes from $i_{k-1}^{(1)}, i_{k-1}^{(2)} \dots i_{k-1}^{(m)}$. Therefore we have,

$$\mathbb{E}_{\mathbf{x}_{k-1}}\mathbb{E}_{\mathbf{x}_k|\mathbf{x}_{k-1}}[f(\mathbf{x}_k)] = \mathbb{E}_{\mathbf{x}_{k-1}}\mathbb{E}_{i_{k-1}^{(1)}, i_{k-1}^{(2)}, \dots, i_{k-1}^{(m)}}[f(\mathbf{x}_k)]$$

Combining this with smoothness result we obtained, and result of assignment 8 with optimal learning rate we have,

$$\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)] = \mathbb{E}_{\mathbf{x}_{k-1}}\mathbb{E}_{i_{k-1}^{(1)}, \dots, i_{k-1}^{(m)}}[f(\mathbf{x}_k)] \leq \frac{\lambda}{2}\mathbb{E}_{\mathbf{x}_{k-1}}\mathbb{E}_{i_{k-1}^{(1)}, \dots, i_{k-1}^{(m)}}[||\delta_k||^2] \leq \frac{\lambda}{2}\mathbb{E}_{\mathbf{x}_{k-1}}[(1-\alpha^*\mu)\|\delta_{k-1}\|^2]$$

Inductively applying the above law of total expectations and also inductively using the relation from Assignment 8, we therefore have,

$$\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)] \le \frac{\lambda}{2}(1-\alpha^*\mu)^k \|\delta_0\|^2 = \frac{\lambda}{2}(1-\alpha^*\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Therefore we have,

$$\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)] \le \frac{\lambda}{2}(1-\alpha^*\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Assignment 10

We are interested in the number of iterations it takes to achieve expected error $\epsilon > 0$. That is, we are interested in the number of iterations it takes $\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] = \epsilon$. Since $f(\mathbf{x}^*) = 0$ ($\because f_i(\mathbf{x}^*) = 0 \implies \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}^*) = 0$), we essentially just require the result of Assignment 9 to compute the number of iterations we need. Indeed, we have,

$$\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)] \le \frac{\lambda}{2}(1-\alpha^*\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \le \frac{\lambda}{2}(1-\alpha^*\mu)^k R$$

Where the final inequality comes from the fact given in the question regarding the distance between starting point and optimum. Further we know 1 - $x \le e^{-x}$, Using this, we have

$$\mathbb{E}_{\mathbf{x}_k}[f(\mathbf{x}_k)] \le \frac{\lambda}{2}(1-\alpha^*\mu)^k R \le \frac{\lambda}{2}e^{-\alpha^*\mu k} R$$

Now we need this error to be ϵ. That is, we need,

$$\frac{\lambda}{2}e^{-\alpha^*\mu k} R = \epsilon \implies e^{-\alpha^*\mu k} = \frac{2\epsilon}{R\lambda} \implies -\alpha^*\mu k = \log\frac{2\epsilon}{R\lambda} \implies \alpha^*\mu k = \log\frac{R\lambda}{2\epsilon}$$

This implies

$$k = \frac{1}{\mu\alpha^*}\log\frac{R\lambda}{2\epsilon}$$

Plugging in the value of $\alpha^* = \frac{m}{L+\lambda(m-1)}$ from Assignment 8, we have

$$k = \frac{L+\lambda(m-1)}{m\mu}\log\frac{R\lambda}{2\epsilon}$$

Therefore

$$k = \mathcal{O}\left(\frac{L+\lambda(m-1)}{m\mu}\log\frac{R\lambda}{2\epsilon}\right)$$

iterations.

Variance Reduction

Assignment 11

We are given the following update rule.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*))$$

Which is the same as,

$$\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \gamma(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*))$$

Now taking norm and squaring, we get,

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_k - \mathbf{x}^* - \gamma(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*))\|^2 \\ = \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma^2\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 &- 2\gamma(\mathbf{x}_k - \mathbf{x}^*)^T(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*))\end{aligned}$$

Using Law of Total Expectation, yet again, we consider,

$$\mathbb{E}_{\mathbf{x}_{k+1}} = \mathbb{E}_{\mathbf{x}_k}\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}$$

Here $\mathbb{E}_{\mathbf{x}_k}$ considers all the randomness in iterate $\mathbf{x}_k$ Now applying this on both sides, we have,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_{k+1}}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] = \mathbb{E}_{\mathbf{x}_k}\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}[\|\mathbf{x}_k - \mathbf{x}^*\|^2 &+ \gamma^2\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 \\ &-2\gamma(\mathbf{x}_k - \mathbf{x}^*)^T(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*))]\end{aligned}$$

By linearity of expectation and first applying inner expectation, the RHS becomes,

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}[\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma^2\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 - 2\gamma(\mathbf{x}_k - \mathbf{x}^*)^T(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*))]$$

$$= \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma^2\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 - 2\gamma((\mathbf{x}_k - \mathbf{x}^*)^T\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)))$$

Now consider,

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)))$$

Here the randomness in $\mathbf{x}_{k+1}$ comes from picking the index i_k. That is,

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}} = \mathbb{E}_{i_k}$$

Moreover,

$$\mathbb{E}_{i_k}[\nabla f_{i_k}(\mathbf{x})] = \sum_{i=1}^{n}\mathbb{P}[i_k = i]\nabla f_{i_k=i}(\mathbf{x}) = \sum_{i=1}^{n}\frac{1}{n}\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$$

Therefore we have, by linearity of expectation, and by the above result,

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}(\nabla f_{i_k}(\mathbf{x}_k)) - \mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}(\nabla f_{i_k}(\mathbf{x}^*)) = \sum_{i=1}^{n} \frac{1}{n}\nabla f_i(\mathbf{x}_k) - \sum_{i=1}^{n} \frac{1}{n}\nabla f_i(\mathbf{x}^*) = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*) = \nabla f(\mathbf{x}_k)$$

since $\mathbf{x}^*$ is the minimizer of f, and f has a unique minimizer by Lemma 3.12, and Lemma 2.22, $\nabla f(\mathbf{x}^*) = 0$ Now consider

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2$$

The randomness comes from picking the indices. Using this, we again have,

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 = \mathbb{E}_{i_k}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 =$$

$$\sum_{i=1}^{n} \mathbb{P}(i_k = i)\|\nabla f_{i_k=i}(\mathbf{x}_k) - \nabla f_{i_k=i}(\mathbf{x}^*)\|^2 = \sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}^*)\|^2$$

Now we have Lemma 7.2 from Handout, which upper bounds precisely the above quantity. Using Lemma 7.2 we therefore have,

$$\mathbb{E}_{\mathbf{x}_{k+1|\mathbf{x}_k}}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^*)\|^2 = \sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}^*)\|^2 \leq 2\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}^*))$$

Plugging these values back into our analysis equation, we have,

$$\mathbb{E}_{\mathbf{x}_{k+1}}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}_{\mathbf{x}_k}[\|\mathbf{x}_k - \mathbf{x}^*\|^2 + 2\gamma^2\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) - 2\gamma(\mathbf{x}_k - \mathbf{x}^*)^T\nabla f(\mathbf{x}_k)]$$

Using strong convexity of f, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}^* - \mathbf{x}_k) + \frac{\mu}{2}||\mathbf{x}^* - \mathbf{x}_k||^2$$

This implies,

$$-\nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{x}^*) \leq f(\mathbf{x}^*) - f(\mathbf{x}_k) - \frac{\mu}{2}||\mathbf{x}^* - \mathbf{x}_k||^2$$

This implies, plugging the values back,

$$\mathbb{E}_{\mathbf{x}_{k+1}}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}_{\mathbf{x}_k}[\|\mathbf{x}_k - \mathbf{x}^*\|^2 + 2\gamma^2\bar{L}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + 2\gamma(f(\mathbf{x}^*) - f(\mathbf{x}_k) - \frac{\mu}{2}||\mathbf{x}^* - \mathbf{x}_k||^2)]$$

This implies

$$\mathbb{E}_{\mathbf{x}_{k+1}}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}_{\mathbf{x}_k}[(1 - \gamma\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2 + (2\gamma^2\bar{L} - 2\gamma)(f(\mathbf{x}_k) - f(\mathbf{x}^*))]$$

We know that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \geq 0$. However, for $0 < \gamma \leq \frac{1}{\bar{L}}$, we have,

$$2\gamma^2\bar{L} - 2\gamma \leq 2(1/\bar{L}^2)\bar{L} - 2(1/\bar{L}) = 0$$

This implies,

$$(2\gamma^2\bar{L} - 2\gamma)(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq 0$$

Therefore we have,

$$\mathbb{E}_{\mathbf{x}_{k+1}}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}_{\mathbf{x}_k}[(1 - \gamma\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2] = (1 - \gamma\mu)\mathbb{E}_{\mathbf{x}_k}[\|\mathbf{x}_k - \mathbf{x}^*\|^2]$$

which is the required result.