Introductory Lectures on Optimization
Homework (3)

Student Yang Yichou
Student ID 3230105697

December 31, 2025

Excercise 1. Convergence of Stochastic Gradient Descent for Convex Function

Consider an optimization problem

$$\underset{\mathbf{x}}{min} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{1}$$

where the objective function F is continuously differentiable and strongly convex with convexity parameter $\mu > 0$. Suppose that the gradient of F, i.e, ∇F, is Lipschitz continuous with Lipschitz constant L, and F can attain its minimum F^* at $\mathbf{x}^*$. We use the stochastic gradient descent (SGD) algorithm to solve the problem (1). Let the solution sequence generated by SGD be $\{\mathbf{x_k}\}$.

1. Please show that $\forall \mathbf{x} \in \mathbf{dom}F$, the following inequality

$$F(\mathbf{x}) - \mathbf{F}^* \leq \frac{1}{2\mu}\|\nabla F(\mathbf{x})\|^2 \tag{2}$$

holds, and interpret the role of strong convexity based on this

2. In practice, for the same problem, SGD enjoys less time cost but more iteration steps than gradient descent methods and may suffer from non-convergence. As a trade-off, consider using mini-batch samples to estimate the full gradient. Taking k^{th} iteration as an example, instead of picking a simple sample, we randomly select a subset $\mathbf{S}_k$ of the sample indices to compute the update direction

$$\mathbf{g}_k(\xi_k) = \frac{1}{|\mathbf{S}_k|}\sum_{i\in\mathbf{S}_k} \nabla f_i(\mathbf{x}_k) \tag{3}$$

where ξ_k is the selected samples. For simplicity, suppose that the mini-batches in all iterations are of constant size,i.e., $|\mathbf{S}_k| = n_m$, and the stepsize α is fixed. Please show that for mini-batch SGD, there holds

$$\mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{x}_k) - F^*] \leq \frac{LM}{2\mu n_m}\alpha + (1-\mu\alpha)^k(F(\mathbf{x}_0) - F^* - \frac{LM}{2\mu n_m}\alpha) \xrightarrow{\text{linear}} \frac{LM}{2\mu n_m}\alpha. \tag{4}$$

Moreover, point out the advantage of mini-batch SGD compared to SGD in terms of the number of the iteration steps.

Proof of Excercise 1: Write down your solution step by step here.
1.F is μ-strongly convex,so we have $\forall x \in \mathbf{dom}F$

$$F^* \geq F(x) + < \nabla F(x), x^* - x > + \frac{1}{2}\mu\|x^* - x\|^2 \tag{5}$$

F is continuously differentiable and strongly convex,so we have

$$\nabla F(x^*) = 0 \tag{6}$$

Therefore,using Cauchy-Schwartz inequation,we can get

$$F(x) - F^* \geq < \nabla F(x), x - x^* > - \frac{1}{2}\mu\|x - x^*\| \overset{C-S}{\leq} \|\nabla F(x)\| \cdot \|x - x^*\| - \frac{\mu}{2}\|x - x^*\|^2 \tag{7}$$

Let $g(t) = \frac{\mu}{2}t^2 + at$,where $a = \|\nabla F(x)\|, t = \|x - x^*\|$,

$$g(t) = -\frac{\mu}{2}(t - \frac{a}{\mu})^2 + \frac{a^2}{2\mu} \leq \frac{\|\nabla F(x)\|^2}{2\mu} \tag{8}$$

Then apply (8) on (7),we have

$$F(x) - F^* \leq \frac{1}{2\mu}\|\nabla F(x)\|^2 \tag{9}$$

2.The following is referenced from
Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. Siam Review, 60(2):223–311, 2018.
Assumption 1(Lipschitz-continuous objective gradients). F is continuously differentiable and ∇F is Lipschitz-continuous with Lipschitz constant $L > 0$,i.e.,

$$\|\nabla F(x) - \nabla F(\overline{x})\|_2 \leq L\|x - \overline{x}\|_2, \forall x, \overline{x} \tag{10}$$

Definition 1(Variance of $g(x_k, \xi_k)$). We define

$$V_{\xi_k}[g(x_k, \xi_k)] := E_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] - \|E_{\xi_k}[g(x_k, \xi_k)]\|_2^2 \tag{11}$$

is the variance of $g(x_k, \xi_k)$.
Assumption 2(First and second moment limits). The objective function F and SG satisfies:

1. $\{x_k\}$ is bounded in an open set and F is bounded by F_{inf}
2. $\exists c_G \geq c > 0, s.t. \forall k \in \mathbb{N}$,

$$\nabla F(x_k)^T E_{\xi_k}[g(x_k, \xi_k)] \geq c\|\nabla F(x_k)\|_2^2 \tag{12}$$

$$\|E_{\xi_k}[g(x_k, \xi_k)]\| \leq c_G\|\nabla F(x_k)\|_2 \tag{13}$$

3. $\exists M \geq 0$ and $M_V \geq 0, s.t. \forall k \in \mathbb{N}$,

$$V_{\xi_k}[g(x_k, \xi_k)] \leq M + M_V\|\nabla F(x_k)\|_2^2 \tag{14}$$

Assumption 3(Strongly Convex Function). F is continuously differentiable and strongly convex,then $\forall x, \overline{x} \in \mathbf{dom}F$

$$F(\overline{x}) \geq F(x) + < \nabla F(x), \overline{x} - x > + \frac{1}{2}\mu\|\overline{x} - x\|^2 \tag{15}$$

Stochastic Gradient Descent. For sequence $\{x_k\}$ and stepsize α_k,update direction $g(x_k, \xi_k)$,the update of x_k follows:

$$x_{k+1} - x_k = \alpha_k g(x_k, \xi_k) \tag{16}$$

Lemma 1. Under Assumption 1 and SG,the following holds:

$$E_{\xi_k}[F(x_{k+1})] - F(x_k) \leq -\alpha_k \nabla F(x_k)^T E_{\xi_k}[g(x_k, \xi_k)] + \frac{1}{2}\alpha_k^2 L E_{\xi_k}[\|g(x_k, \xi_k)_2^2] \tag{17}$$

Proof:

$$\begin{aligned} F(x_{k+1}) - F(x_k) &\overset{(10)}{\leq} \nabla F(x_k)^T (x_{k+1} - x_k) + \frac{1}{2}L\|x_{k+1} - x_k\|_2^2 \\ &\overset{(16)}{\leq} -\alpha_k \nabla F(x_k)^T g(x_k, \xi_k) + \frac{1}{2}\alpha_k^2 L \|g(x_k, \xi_k)\|_2^2 \end{aligned}$$

Taking expectation on both sides,we can derive the equation easily.
Lemma 2. The bound of the expectation of $g(x_k, \xi_k)$ is:

$$E_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] \leq M + M_G \|\nabla F(x_k)\|_2^2, M_G := M_V + c_G^2 \tag{18}$$

Proof:

$$\begin{aligned} E_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] &\overset{(11)}{=} V_{\xi_k}[g(x_k, \xi_k)] + \|E_{\xi_k}[g(x_k, \xi_k)]\|_2^2 \\ &\overset{(14)}{\leq} \|E_{\xi_k}[g(x_k, \xi_k)]\|_2^2 + M + M_V \|\nabla F(x_k)\|_2^2 \\ &\overset{(13)}{\leq} M + (M_V + c_G^2)\|\nabla F(x_k)\|_2^2 \\ &= M + M_G \|\nabla F(x_k)\|_2^2, M_G := M_V + c_G^2 \end{aligned}$$

Lemma 3. Further derivation of $E_{\xi_k}[F(x_{k+1})] - F(x_k)$:

$$\begin{aligned} E_{\xi_k}[F(x_{k+1})] - F(x_k) &\leq -c\alpha_k \|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L E_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] & (19) \\ &\leq -(c - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M & (20) \end{aligned}$$

Proof:

$$\begin{aligned} E_{\xi_k}[F(x_{k+1})] - F(x_k) &\overset{(17)}{\leq} -\alpha_k \nabla F(x_k)^T E_{\xi_k}[g(x_k, \xi_k)] + \frac{1}{2}\alpha_k^2 L E_{\xi_k}[\|g(x_k, \xi_k)_2^2] \\ &\overset{(12)}{\leq} -c\alpha_k \|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L E_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] \\ &\overset{(18)}{\leq} -(c - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M \end{aligned}$$

Theorem(Strongly Convex Objective, Fixed Stepsize) SG runs with a fixed stepsize $\alpha_k = \alpha$,satisfying

$$0 < \alpha \leq \frac{c}{LM_G}. \tag{21}$$

Then,the expected optimality gap satisfies:

$$\begin{aligned} E[F(x_k) - F^*] \leq &\frac{\alpha L M}{2c\mu} + (1 - \alpha c\mu)^{k-1}(F(x_1) - F^* - \frac{\alpha L M}{2c\mu}) \\ &\overset{k\to\infty}{\to} \frac{\alpha L M}{2c\mu} \end{aligned} \tag{22}$$

Proof:

$$\begin{aligned}
E_{\xi_k}[F(x_{k+1})] - F(x_k) &\overset{(20)}{\leq} -(c - \frac{1}{2}\alpha_k LM_G)\alpha_k \|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
&\overset{(21)}{\leq} -\frac{1}{2}\alpha c\|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\
&\overset{(2)}{\leq} -\alpha c\mu(F(x_k) - F^*) + \frac{1}{2}\alpha^2 LM
\end{aligned}$$

Subtracting F^* on both sides,we get

$$E[F(x_{k+1}) - F^*] \leq (1 - \alpha c\mu)E[F(x_k) - F^*] + \frac{1}{2}\alpha^2 LM$$

Subtracting $\frac{\alpha LM}{2c\mu}$ on both sides,we get

$$E[F(x_{k+1}) - F^*] - \frac{\alpha LM}{2c\mu} \leq (1 - \alpha c\mu)(E[F(x_k) - F^*] - \frac{\alpha LM}{2c\mu}) \tag{23}$$

where $0 < \alpha c\mu \leq \frac{\mu c^2}{LM_G} < \frac{\mu c^2}{Lc^2} = \frac{c}{L} \leq 1$
Iteration k:

$$\begin{aligned}
E[F(x_k) - F^*] \leq& \frac{\alpha LM}{2c\mu} + (1 - \alpha c\mu)^{k-1}(F(x_1) - F^* - \frac{\alpha LM}{2c\mu}) \\
&\overset{k\to\infty}{\to} \frac{\alpha LM}{2c\mu}
\end{aligned}$$

Now we apply mini-batch $\mathbf{g}_k(\xi_k) = \frac{1}{|\mathbf{S}_k|}\sum_{i\in\mathbf{S}_k}\nabla f_i(\mathbf{x}_k)$ to the **Theorem**.
The main difference is that mini-batch divides the variance by $|\mathbf{S}_k| = n_m$ times,then (14) is updated to:

$$V_{\xi_k}[g(x_k, \xi_k)] \leq \frac{M}{n_m} + M_V\|\nabla F(x_k)\|_2^2$$

Change the proof of the **Theorem**:

$$\begin{aligned}
E_{\xi_k}[F(x_{k+1})] - F(x_k) &\leq -\frac{1}{2}\alpha c\|\nabla F(x_k)\|_2^2 + \frac{1}{2}\alpha^2 L\frac{M}{n_m} \\
&\overset{(2)}{\leq} -\alpha c\mu(F(x_k) - F^*) + \frac{1}{2}\alpha^2 L\frac{M}{n_m}
\end{aligned}$$

Then follow the same derivation as **Theorem**:

$$\begin{aligned}
E_{\xi_0:\xi_{k-1}}[F(x_k) - F^*] \leq& \frac{LM}{2c\mu n_m}\alpha + (1 - \alpha c\mu)^k(F(x_0) - F^* - \frac{LM}{2c\mu n_m}\alpha) \\
&\overset{k\to\infty}{\to} \frac{LM}{2c\mu n_m}\alpha
\end{aligned}$$

Note that mini-batch sampling is unbiased gradient estimation,which means:

$$E_{\xi_k}[g(x_k, \xi_k)] = \nabla F(x_k) \tag{24}$$

Apply it to **Assumption 2**,we can get

$$c_G = c = 1$$

Then,we obtain the final equation:

$$\begin{aligned}
E_{\xi_0:\xi_{k-1}}[F(x_k) - F^*] \leq& \frac{LM}{2\mu n_m}\alpha + (1 - \alpha\mu)^k(F(x_0) - F^* - \frac{LM}{2\mu n_m}\alpha) \\
&\overset{k\to\infty}{\to} \frac{LM}{2\mu n_m}\alpha
\end{aligned}$$

The main advantage of mini-batch SGD compared to SGD in terms of the number of iteration steps is that:
Although the computation cost for each step is higher,the variance will be reduced by $O(n_m)$ times,leading to a more steady path.The constant term M^2 in the convergence now reduces to M^2/n_m. □