1 Outline

In this lecture, we study

- convergence of stochastic gradient descent,
- optimality conditions for general convex functions,
- Proximal gradient descent.

2 Convergence of stochastic gradient descent

Recall that stochastic gradient descent (SGD) proceeds as the following.

Algorithm 1 Stochastic gradient descent (SGD)

Initialize $x_1 \in C$.
for $t = 1, \dots, T$ **do**
  Obtain an estimator $\hat{g}_{x_t}$ of some $g_t \in \partial f(x_t)$.
  Update $x_{t+1} = \mathrm{Proj}_C \{x_t - \eta_t \hat{g}_{x_t}\}$ for a step size $\eta_t > 0$.
end for
Return $(1/T) \sum_{t=1}^{T} x_t$.

In this section, we analyze the convergence of SGD under the following assumption.

Assumption 1. Assume that $\hat{g}_x$ satisfies

$$\mathbb{E}[\hat{g}_x] = g_x \text{ for some } g_x \in \partial f(x), \quad \mathbb{E}\left[\|\hat{g}_x\|^2\right] \le L^2.$$

This assumption is analogous to Lipschitz continuity. Under the assumption, let us analyze the performance of stochastic gradient descent given by Algorithm 1.

Theorem 15.1. *Algorithm 1 with step sizes $\eta_t = R/(L\sqrt{t})$ satisfies*

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) \le \frac{3LR}{2\sqrt{T}}$$

where the expectation is taken over the randomness in gradient estimation and $x^* \in \mathrm{argmin}_{x \in C} f(x)$.

2.1 Proof via online regret minimization

Suppose that $\mathbb{E}[\hat{g}_{x_t}] = g_t \in \partial f(x_t)$ for $t \geq 1$. First, let us observe the following.

$$\begin{aligned}
\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(x_t)\right] - f(x^*) \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(f(x_t) - f(x^*))\right] \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} g_t^\top (x_t - x^*)\right] \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\hat{g}_{x_t}|x_t\right]^\top (x_t - x^*)\right] \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \hat{g}_{x_t}^\top (x_t - x^*)\right]
\end{aligned}$$

where the inequalities are due to the convexity of f and the last equality is due to the tower rule. Now let us consider functions $f_1, \ldots, f_T$ given by

$$f_t(x) = \hat{g}_{x_t}^\top x.$$

Then

$$\begin{aligned}
\sum_{t=1}^{T} \hat{g}_{x_t}^\top (x_t - x^*) &= \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \\
&\leq \sum_{t=1}^{T} f_t(x_t) - \min_{x \in C} \sum_{t=1}^{T} f_t(x) \\
&\leq \frac{3}{2} LR\sqrt{T}
\end{aligned}$$

where the last inequality is from the convergence result of online gradient descent. Note that this upper bound holds regardless of any realization of $\hat{g}_{x_t}$'s. Therefore, the result follows.

2.2 Proof from the analysis of the subgradient method

Note that

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 \,|x_t\right] &= \mathbb{E}\left[\|\mathrm{Proj}_C(x_t - \eta_t \hat{g}_{x_t}) - x^*\|_2^2 \,|x_t\right] \\
&\leq \mathbb{E}\left[\|x_t - \eta_t \hat{g}_{x_t} - x^*\|_2^2 \,|x_t\right] \\
&= \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] - 2\eta_t \mathbb{E}\left[\hat{g}_{x_t}|x_t\right]^\top (x_t - x^*) \\
&= \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] - 2\eta_t g_t^\top (x_t - x^*) \\
&\leq \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] - 2\eta_t (f(x_t) - f(x^*)).
\end{aligned}$$

Then, based on the tower rule,

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2\right] &\le \mathbb{E}\left[\|x_t-x^*\|_2^2\right]+\eta_t^2\mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2\right]-2\eta_t(\mathbb{E}\left[f(x_t)\right]-f(x^*)) \\ &\le \mathbb{E}\left[\|x_t-x^*\|_2^2\right]+\eta_t^2L^2-2\eta_t(\mathbb{E}\left[f(x_t)\right]-f(x^*)).\end{aligned}$$

Then it follows that

$$\mathbb{E}\left[f(x_t)\right]-f(x^*) \le \frac{1}{2\eta_t}\left(\mathbb{E}\left[\|x_t-x^*\|_2^2\right]-\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2\right]\right)+\frac{\eta_t}{2}L^2.$$

Summing up this for $t=1,\dots,T$ and dividing each side by T, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(x_t)\right]-f(x^*) &\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|x_t-x^*\|_2^2\right]\left(\frac{1}{2\eta_t}-\frac{1}{2\eta_{t-1}}\right)+\frac{L^2}{2T}\sum_{t=1}^{T}\eta_t \\ &\le \frac{R^2}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta_t}-\frac{1}{2\eta_{t-1}}\right)+\frac{L^2}{2T}\sum_{t=1}^{T}\eta_t \\ &\le \frac{LR}{2\sqrt{T}}+\frac{LR}{\sqrt{T}}.\end{aligned}$$

By convexity,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(x_t)\right] \ge \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)\right],$$

and therefore, the result follows.

2.3 Strongly convex functions

For strongly convex functions, we have the following convergence result.

Theorem 15.2. *Assume the same conditions on $\hat{g}_x$ and that f is α-strongly convex with respect to the ℓ_2 norm for some $\alpha>0$. Algorithm 1 with step sizes $\eta_t=2/(\alpha(t+1))$ satisfies*

$$\mathbb{E}\left[f\left(\sum_{t=1}^{T}\frac{2t}{T(T+1)}x_t\right)\right]-f(x^*) \le \frac{2L^2}{\alpha(T+1)}$$

where the expectation is taken over the randomness in gradient estimation and $x^* \in \mathrm{argmin}_{x\in C} f(x)$.

Therefore, for Lipschitz continuous functions and functions that are stronngly convex and Lipschitz, we recover the same convergence rate as the subgradient method.

2.4 No self-tuning property due to variance

For gradient descent, smoothness does make a difference due to the self-tuning property. For smooth functions, the convergence rate is $O(1/T)$ (we also saw the accelerated method achieving $O(1/T^2)$ rate). For smooth and strongly convex functions, we obtained $O(\gamma^T)$ rate for some $0<\gamma<1$. Is it the case for SGD as well? The answer is no.

The crucial property of smooth functions which we relied on in the convergence analysis was the self-tuning property. For a smooth function f, as we get close to an optimal solution $x^* \in$

$\operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$, the size of the gradient $\|\nabla f(x)\|_2$ gets smaller. However, even if f is smooth and x goes to x^*, $\mathbb{E}\left[\|\hat{g}_x\|_2^2\right]$ does not converge to 0.

Let us consider the mean squared error minimization problem given by

$$\min_{\beta} \quad f(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(y_i - \beta^\top x_i)^2.$$

Here, f is smooth because

$$\begin{aligned}
\|\nabla f(\beta_1) - \nabla f(\beta_2)\|_2 &= \left\|\frac{1}{n}\sum_{i=1}^{n}(\beta_1-\beta_2)^\top x_i x_i\right\|_2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left|(\beta_1-\beta_2)^\top x_i\right|\|x_i\|_2 \\
&\le \|\beta_1-\beta_2\|_2\left(\frac{1}{n}\sum_{i=1}^{n}\|x_i\|_2^2\right) \\
&\le M^2\|\beta_1-\beta_2\|_2
\end{aligned}$$

where $\max_{i\in[n]}\|x_i\| = M$.

Next take the optimal solution $\beta^* \in \operatorname{argmin}_\beta f(\beta)$ which satisfies $\nabla f(\beta^*) = 0$. Then sample a data point (x_i, y_i) to obtain an unbiased estimator

$$\hat{g}_{\beta^*} = (y_i - x_i^\top \beta^*)(-x_i).$$

Here, if the data point (x_i, y_i) is not on the line $y = \beta^\top x$ and x_i is nonzero, then $\hat{g}_{\beta^*} \neq 0$.

3 Optimality conditions for non-differentiable convex functions

Now we consider the convex minimization problem with a general convex objective function that is not necessarily differentiable.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array} \quad = \quad \begin{array}{ll} \text{minimize} & f(x) + I_C(x) \\ \text{subject to} & x \in \mathbb{R}^d. \end{array}$$

The first formulation is the constrained version, and the second formulation shows its unconstrained version with the indicator function. We discussed optimality conditions for convex minimization problems with a differentiable objective. In this section, we state and prove optimality conditions for the general case, in which the objective can be non-differentiable.

Remember that when a convex function f is differentiable and C is a convex domain, $x^* \in C$ is an optimal solution to $\min_{x\in C} f(x)$ if and only if

$$\nabla f(x)^\top (x - x^*) \ge 0 \quad \text{for all } x \in C.$$

When f is not differentiable, subgradients generalize the gradient even for the optimality condition.

Theorem 15.3. *For a convex optimization problem* $\min_{x\in C} f(x)$, $x^* \in C$ *is an optimal solution if and only if there exists* $s \in \partial f(x^*)$ *such that*

$$s^\top (x - x^*) \ge 0 \quad \textit{for all } x \in C.$$

An immediate corollary of Theorem 15.3 is the following optimality condition for unconstrained problems.

Corollary 15.4. *For a convex optimization problem* $\min_{x\in\mathbb{R}^d} f(x)$, $x^* \in \mathbb{R}^d$ *is an optimal solution if and only if* $0 \in \partial f(x^*)$.

Corollary 15.4 can be applied to the unconstrained formulation of constrained convex minimization. Remember that when a convex function f is differentiable and C is a convex domain, $x^* \in C$ satisfies

$$\nabla f(x)^\top (x - x^*) \geq 0 \quad \text{for all } x \in C$$

if and only if

$$0 \in \nabla f(x^*) + N_C(x^*)$$

because

$$\begin{aligned} N_C(x^*) &= \left\{ g \in \mathbb{R}^d : \ g^\top (y - x^*) \leq 0 \quad \forall y \in C \right\} \\ &= \left\{ g \in \mathbb{R}^d : \ I_C(y) \geq g^\top (y - x^*) + I_C(x^*) \quad \forall y \in \mathrm{dom}(I_C) \right\}. \end{aligned}$$

Corollary 15.5. *For a convex optimization problem* $\min_{x\in C} f(x)$, $x^* \in C$ *is an optimal solution if and only if*

$$0 \in \partial f(x^*) + N_C(x^*).$$

Likewise, we have the following condition for general convex functions.

Proof. By Corollary 15.4, it follows that $x^* \in C$ is an optimal solution to $\min\left(f(x) + I_C(x)\right)$ if and only if

$$0 \in \partial\left(f(x^*) + I_C(x^*)\right) = \partial f(x^*) + \partial I_C(x^*).$$

Recall that

$$\begin{aligned} \partial I_C(x^*) &= \left\{ g \in \mathbb{R}^d : \ I_C(y) \geq g^\top (y - x^*) + I_C(x^*) \quad y \in \mathrm{dom}(I_C) \right\} \\ &= \left\{ g \in \mathbb{R}^d : \ g^\top (y - x^*) \leq 0 \quad \forall y \in C \right\} \\ &= N_C(x^*). \end{aligned}$$

Therefore, $0 \in \partial f(x^*) + \partial I_C(x^*)$ holds if and only if

$$0 \in \nabla f(x^*) + N_C(x^*)$$

holds, as required. □

In this section, we will prove Theorem 15.3 which states the optimality condition for convex minimization. A tool that we need is the separating hyperplane theorem, which is an important result in convex analysis on its own. We state the separating hyperplane theorem without proof.

Theorem 15.6 (Separating hyperplane theorem)**.** *Let* $C, D \subseteq \mathbb{R}^d$ *be disjoint convex sets, i.e.,* $C \cap D = \emptyset$*, then there exists* $a \in \mathbb{R}^d \setminus \{0\}$ *and* $b \in \mathbb{R}$ *such that*

$$\begin{aligned} a^\top x &\geq b, \quad \textit{for all } x \in C \\ a^\top x &\leq b, \quad \textit{for all } x \in D \end{aligned}$$

Let us prove Theorem 15.3 using Theorem 15.6.

Proof of Theorem 15.3. ($\Leftarrow$) Assume that there exists $s \in \partial f(x^*)$ such that $s^\top (x - x^*) \geq 0$ holds for all $x \in C$. Then it follows from the definition of subgradients that

$$f(x) - f(x^*) \geq s^\top (x - x^*) \geq 0 \quad \text{for all } x \in C.$$

This implies that $f(x) \geq f(x^*)$ for all $x \in C$, so x^* is optimal.

($\Rightarrow$) Let us consider the following two sets.

$$C = \{(x - x^*, t) : \ f(x) - f(x^*) \leq t\},$$
$$D = \{(x - x^*, t) : \ x \in C, \ t < 0\}.$$

Since $f(x) - f(x^*) \geq 0$ for any $x \in C$, these two sets are disjoint. Then by Theorem 15.6, there exists $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$ such that $(a, b) \neq (0, 0)$ and

$$a^\top (x - x^*) + bt \geq c, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t \tag{15.1}$$
$$a^\top (x - x^*) + bt \leq c, \quad \forall x \in C, \ t < 0. \tag{15.2}$$

In (15.2), t can be arbitrarily small, so $b \geq 0$. Suppose that $b = 0$, in which case (15.1) becomes

$$a^\top (x - x^*) 0 \geq c, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t.$$

Here, $x - x^*$ can be $\lambda \cdot a$ where λ is an arbitrarily small number. This implies that $a = 0$. However, this contradicts the condition that $(a, b) \neq (0, 0)$. Therefore, $b > 0$. Then, without loss of generality, we may assume that $b = 1$. Then taking $x = x^*$ and $t = 0$ in (15.1), we obtain $0 \geq c$. Moreover, taking $x = x^*$ and a number that is arbitrarily close to 0 for t, it follows that $0 \leq c$. Hence, $c = 0$. Then (15.1) and (15.2) become

$$a^\top (x - x^*) + t \geq 0, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t \tag{15.3}$$
$$a^\top (x - x^*) + t \leq 0, \quad \forall x \in C, \ t < 0. \tag{15.4}$$

Here, we take $t = f(x) - f(x^*)$ in (15.3). Then (15.3) becomes

$$f(x) \geq f(x^*) - a^\top (x - x^*),$$

which implies that $-a \in \partial f(x^*)$. Moreover, we take a number that is arbitrarily close to 0 for t in (15.4). Then it becomes $a^\top (x - x^*) \leq 0$, which is equivalent to $-a^\top (x - x^*) \geq 0$. Hence, $-a$ is the desired vector. □

4 Proximal gradient descent

Recall the formulation of LASSO, given by

$$\min_\beta \quad \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

Here, the objective function is non-differentiable because of the ℓ_1-regularization term $\lambda\|\beta\|_1$, and therefore, it is non-smooth. On the other hand, the objective is convex, and we have a characterization of the subdifferential of $\|\beta\|_1$, so we can simply apply the subgradient method. To bound the additive error by ϵ, the subgradient method requires $O(1/\epsilon^2)$ iterations.

If you take a closer look at the objective, it consists of two part. One part is smooth, and the other part is something whose subdifferential is well understood. Can we use this structure to obtain a better algorithm? The main subject of this section is developing an algorithm that converges to an ϵ-approximate solution after $O(1/\epsilon)$ iterations.

4.1 Projection and proximal operator

We studied the projected gradient descent method, where at each step, we take a projection to the constraint set. When the constraint set is given by C, the projection operator is given by

$$\mathrm{Proj}_C(x) = \operatorname*{argmin}_{u \in C} \frac{1}{2}\|u - x\|_2^2 = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ I_C(u) + \frac{1}{2}\|u - x\|_2^2 \right\}$$

where $I_C(u)$ is the indicator function of C. This definition is proper as there is a unique minimizer for the optimization problem. Hence, the projection operator is defined by the indicator function and the proximity term $(1/2)\|u - x\|_2^2$. The proximal operator is a generalization of the projection operator replacing the indicator function by other general functions.

The proximal operator with respect to a convex function h is defined as follows.

$$\mathrm{Prox}_h(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ h(u) + \frac{1}{2}\|u - x\|_2^2 \right\}.$$

Again the definition is proper because the objective of the optimization problem is strongly convex. Hence, for any $\eta > 0$,

$$\mathrm{Prox}_{\eta h}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ h(u) + \frac{1}{2\eta}\|u - x\|_2^2 \right\}.$$

As projected gradient descent proceeds with the update rule

$$x_{t+1} = \mathrm{Proj}_C \left\{ x_t - \eta \nabla f(x_t) \right\},$$

we can defined the proximal gradient method with the update rule

$$x_{t+1} = \mathrm{Prox}_{\eta h}(x_t - \eta \nabla f(x_t)).$$

In particular, when we take the indicator function I_C for h, the proximal gradient method reduces to the projeced gradient descent method.

Lemma 15.7. *$u = prox_{\eta h}(x)$ if and only if $x - u \in \eta \partial h(u)$.*

Proof. Note that $u = \mathrm{prox}_{\eta h}(x)$ means that u minimizes $h(u) + (1/2\eta)\|u - x\|_2^2$. By the optimality condition, it is equivalent to $0 \in \partial h(u) + \{(1/\eta)(u - x)\}$, and this is equivalent to $x - u \in \eta \partial h(u)$. □

4.2 Example: ℓ_1 regularization

Consider $h(x) = \|x\|_1$. Then

$$\mathrm{prox}_{\eta h}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ \|u\|_1 + \frac{1}{2\eta}\|u - x\|_2^2 \right\}.$$

Let $u = \mathrm{prox}_{\eta h}(x)$. Then, by Lemma 15.7,

$$x - u \in \eta \partial \|u\|_1.$$

Recall that $g \in \partial \|u\|_1$ if and only if

$$g_i = \begin{cases} \mathrm{sign}(u_i), & \text{if } u_i \neq 0, \\ \text{a value in } [-1, 1], & \text{if } u_i = 0. \end{cases}$$

Based on this, we can argue that $x - u \in \eta \partial \|u\|_1$ if and only if

$$u_i = \begin{cases} x_i - \eta, & \text{if } x_i \geq \eta, \\ 0, & \text{if } -\eta \leq x_i \leq \eta. \\ x_i + \eta, & \text{if } x_i \leq -\eta. \end{cases}$$

Moreover, $x - u \in \eta \partial \|u\|_1$ if and only if

$$u_i = \max\{0, |x_i| - \eta\} \cdot \text{sign}(x_i).$$

For example,

$$\text{prox}_h((3, 1, -2)^\top) = (2, 0, -1)^\top.$$

Note that when $h = \|x\|_1$, the corresponding proximal operator "shrinks" the vector. For this reason, the operator is called the self-thresholding operator or the shrinkage operator.

4.3 Example: quadratic function

Consider $h(x) = (1/2)x^\top Ax + b^\top x + c$ where A is positive semidefinite. Then

$$\text{prox}_{\eta h}(x) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2} u^\top Au + b^\top u + c + \frac{1}{2\eta} \|u - x\|_2^2 \right\}.$$

Setting $v = \text{prox}_{\eta h}(x)$, it follows from the optimality condition that

$$0 = Av + b + \frac{1}{\eta}(v - x).$$

Therefore,

$$\text{prox}_{\eta h}(x) = v = (I + \eta A)^{-1}(x - \eta b).$$