1 Outline

In this lecture, we study

- Frank-Wolfe algorithm,
- Applications in LASSO and matrix completion,
- Introduction to online convex optimization.

2 Stochastic gradient descent

Let us consider convex optimization stated as

$$\min_{x \in C} f(x).$$

If we have an access to its gradient or one of its subgradients, then we can apply gradient descent or the subgradient method. However, depending on situations, it may not be realistic to assume that we have an oracle that provides exact gradients. For example, we have just considered the stochastic optimization setting where f is given by $f(x) = \mathbb{E}_{\xi \sim \mathbb{P}}[h(x, \xi)]$, the expectation of a random function. Here, $\nabla f(x) = \mathbb{E}_{\xi \sim \mathbb{P}}[\nabla h(x, \xi)]$, to compute which we need to know the distribution $\mathbb{P}$ in general. Instead of computing the expectation exactly, what we did was to obtain a sample ξ_t so that we may use $\nabla h(x, \xi_t)$ for each iteration t. Here $\nabla h(x, \xi_t)$ is an unbiased estimator of $\nabla f(x)$.

Another example is the mean squared error minimization problem for regression.

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

where $(x_1, y_1), \ldots, (x_n, y_n)$ are the given data. In fact, this setting is also a stochastic optimization problem as we can define $\mathbb{P}$ as the empirical distribution over the n samples. To be more specific,

$$\mathbb{P}\left((x, y) = (x_i, y_i)\right) = \frac{1}{n}.$$

Then the gradient of f at β is given by

$$\nabla f(\beta) = \mathbb{E}_{(x,y) \sim \mathbb{P}}[\nabla h(\beta, (x, y))] = -\frac{1}{n} \sum_{i=1}^{n} (y_i - \beta^\top x_i) x_i.$$

In this example, we know the precise description of the underlying distribution, from which we can compute the exact gradient. Then, what is the problem? Here, to compute the gradient, we have to go through all data points $(x_1, y_1), \ldots, (x_n, y_n)$, which may not be practical especially when the number of data is large. For this scenario, a strategy is to obtain an estimation of the gradient. We sample a data (x_r, y_r) from the data set uniformly at random and obtain

$$g_r = -(y_r - \beta^\top x_r) x_r.$$

Here r is a random variable following the uniform distribution over $\{1, \ldots, n\}$. Note that

$$\mathbb{E}[g_r] = \sum_{i=1}^{n} \mathbb{P}(r = i) \cdot g_i = \sum_{i=1}^{n} \frac{1}{n} \cdot g_i = -\frac{2}{n} \sum_{i=1}^{n} (y_i - \beta^\top x_i) x_i = \nabla f(\beta).$$

Hence, g_r is an unbiased estimator of g_r. What we do next is to use g_r to replace $\nabla f(\beta)$ when running gradient descent. More generally, let $\tilde{g}_x$ be an unbiased estimator of the gradient of f at x or the subgradient for f at x.

Algorithm 1 Stochastic gradient descent (SGD)

Initialize $x_1 \in C$.
for $t = 1, \ldots, T$ **do**
  Obtain an estimator $\hat{g}_{x_t}$ of some $g \in \partial f(x_t)$.
  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 $(1/T) \sum_{t=1}^{T} x_t$.

2.1 Convergence of stochastic gradient descent

Assume that $\tilde{g}_x$ satisfies

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

Under this assumption, let us analyze the performance of stochastic gradient descent given by Algorithm 1.

Theorem 13.1. *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 $x^* \in \mathrm{argmin}_{x \in C} f(x)$.

Proof. Note that

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

Then, based on the tower rule,

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

Then it follows that

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

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

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

By convexity,

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

and therefore, the result follows. □

2.2 Strongly convex functions

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

Theorem 13.2. *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*

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

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

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

2.3 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

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

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 $\hat{g}_{\beta^*} \neq 0$.

2.4 Stochastic optimization

Stochastic optimization (SO) is an optimization problem of the following form.

$$\underset{x \in C}{\text{minimize}} \quad \mathbb{E}_{\xi \sim \mathbb{P}} \left[h(x, \xi) \right]$$

where

- ξ is a random parameter vector whose underlying distribution is given by $\mathbb{P}$,
- $h(x, \xi)$ is convex with respect to x for any fixed ξ,
- C is the feasible set for the decision vector x.

Then

$$f(x) = \mathbb{E}_{\xi \sim \mathbb{P}} \left[h(x, \xi) \right]$$

is convex. For example, for the linear regression problem, we consider

$$h(\beta, (x, y)) = \frac{1}{2}(y - \beta^\top x)^2,$$

and

$$\text{minimize} \quad \mathbb{E}_{(x,y) \sim \mathbb{P}} \left[h(\beta, (x, y)) \right] \quad = \quad \text{minimize} \quad \mathbb{E}_{(x,y) \sim \mathbb{P}} \left[\frac{1}{2}(y - \beta^\top x)^2 \right]$$

where x is the feature vector, y is the response variable, and (x, y) follows distribution $\mathbb{P}$.

Algorithm 2 Stochastic gradient descent for stochastic optimization

Initialize $x_1 \in C$.
for $t = 1, \dots, T$ **do**
  Obtain a random vector $\xi_t \sim \mathbb{P}$ and a subgradient $g(x_t, \xi_t) \in \partial h(x_t, \xi_t)$.
  Obtain $x_{t+1} = \mathrm{Proj}_C \{x_t - \eta_t g(x_t, \xi_t)\}$ for a step size $\eta_t > 0$.
end for

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

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

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

Theorem 13.3. *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*

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

An immediate corollary of Theorem 13.3 is the following optimality condition for unconstrained problems.

Corollary 13.4. *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 13.4 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 13.5. *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 13.4, 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 13.3 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 13.6 (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 13.3 using Theorem 13.6.

Proof of Theorem 13.3. ($\Leftarrow$) 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.

$$\begin{aligned}
C &= \left\{ (x - x^*, t) : \ f(x) - f(x^*) \leq t \right\}, \\
D &= \left\{ (x - x^*, t) : \ x \in C, \ t < 0 \right\}.
\end{aligned}$$

Since $f(x) - f(x^*) \geq 0$ for any $x \in C$, these two sets are disjoint. Then by Theorem 13.6, there exists $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$ such that $(a, b) \neq (0, 0)$ and

$$a^\top (x - x^*) + bt \geq c, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t \tag{13.1}$$

$$a^\top (x - x^*) + bt \leq c, \quad \forall x \in C, \ t < 0. \tag{13.2}$$

In (13.2), t can be arbitrarily small, so $b \geq 0$. Suppose that $b = 0$, in which case (13.1) 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 (13.1), 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 (13.1) and (13.2) become

$$a^\top (x - x^*) + t \geq 0, \quad \forall x \in \mathbb{R}^d,\ f(x) - f(x^*) \leq t \tag{13.3}$$

$$a^\top (x - x^*) + t \leq 0, \quad \forall x \in C,\ t < 0. \tag{13.4}$$

Here, we take $t = f(x) - f(x^*)$ in (13.3). Then (13.3) 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 (13.4). 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. □