[CS3958: Lecture 7] Convex Optimization

Instructor: Chihao Zhang, Scribed by Yulin Wang

November 25, 2022

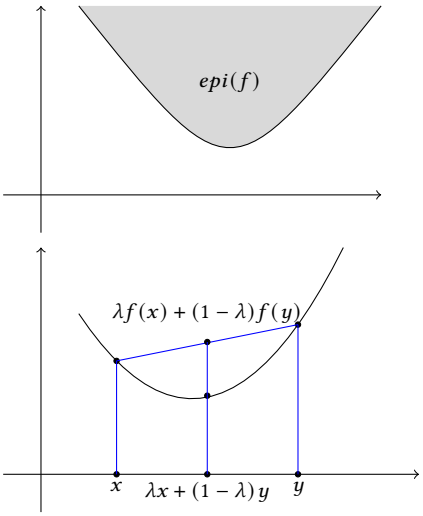
1 Convex Optimization

In the following, we will review some background of convex optimization and examine the classic gradient descent algorithm. In convex optimization, one is asked to solve the following problem

$$\min f(x) \text{ s.t. } x \in V \subseteq \mathbb{R}^n$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a *convex function* and V is a *convex set*.

Recall that we say $V \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in V$ and $\lambda \in [0,1]$, it holds that $\lambda x + (1-\lambda)y \in V$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its *epigraph*, namely the set $\text{epi}(f) = \left\{(x, y) \in \mathbb{R}^{n+1} \,\middle|\, y \geq f(x)\right\}$, is a convex set.



We will make use of the following useful properties of a convex function f.

- (The Jensen's inequality) For any $x, y \in \text{dom}(f)$ and $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$. In particular $f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$.
- (Taylor expansion with Lagrange remainder) If $f \in C^2$, then

  $$f(y) = f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(\xi)(y-x),$$

  for some ξ on the line connecting x and y. Moreover, f is convex is equivalent to the fact that $\nabla^2 f(z) \geq 0$ for any $z \in \text{dom} f$.
- (The first order optimality condition)

  $$x^* = \arg\min_x f(x) \iff \nabla f(x^*) = 0;$$
  $$x^* = \arg\min_{x \in V} f(x) \iff \forall y \in V, \nabla f(x^*)^\top (y - x^*) \geq 0.$$

1.1 The Gradient Descent Algorithm

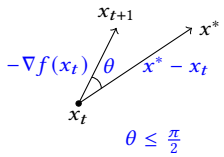
We first assume the optimization problem is unconstrained, namely $V = \mathbb{R}^n$. We also assume the existence of a *first order oracle* for the function f, that is, given any $x \in \mathbb{R}^n$, we can get the value of $\nabla f(x)$. Then the *gradient descent* algorithm is simply the following updating rule:

$$x_{t+1} = x_t - \eta \nabla f(x_t),$$

where $\eta > 0$ is the step size.

Let $x^* = \arg\min_x f(x)$. Why $-\nabla f(x_t)$ is a good direction towards x^*. It follows from the convexity of f that

$$-\nabla f(x_t)(x^* - x_t) \geq f(x_t) - f(x^*) \geq 0$$



as long as x_t is not at x^*. Therefore, moving along $-\nabla f(x_t)$ should make progress.

We now quantitatively calculate the progress. Let $\phi(x) := \frac{1}{2}\|x - x^*\|^2$ be a potential function to measure the distance between x and x^*. Then

$$\begin{aligned}
\phi(x_{t+1}) - \phi(x_t) &= \frac{1}{2}\left(\langle x_{t+1} - x^*, x_{t+1} - x^*\rangle - \langle x_t - x^*, x_t - x^*\rangle\right) \\
&= \frac{1}{2}\left(\langle x_t - \eta\nabla f(x_t) - x^*, x_t - \eta\nabla f(x_t) - x^*\rangle - \langle x_t - x^*, x_t - x^*\rangle\right) \\
&= -\eta\langle \nabla f(x_t), x_t - x^*\rangle + \frac{1}{2}\eta^2\|\nabla f(x_t)\|^2 \\
&\le \eta(f(x^*) - f(x_t)) + \frac{1}{2}\eta^2\|\nabla f(x_t)\|^2,
\end{aligned}$$

where the last inequality follows from the convexity of f.

Summing up above for $t = 0, 1, \ldots, T-1$, we obtain

$$\phi(x_T) - \phi(x_0) \le \frac{1}{2}\eta^2 \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 - \eta \sum_{t=0}^{T-1} (f(x_t) - f(x^*)).$$

Rearranging yields

$$\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \le \frac{\phi(x_0) - \phi(x_T)}{\eta} + \frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2. \tag{1}$$

The bound (1) is important and informative. We further explain and discuss some of its extension below.

Averaging

If we divide T on both sides of (1), it becomes to

$$\frac{1}{T}\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \le \frac{\phi(x_0) - \phi(x_T)}{\eta T} + \frac{\eta}{2T} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2.$$

The LHS of above is the average gap between the function value on points found so far and the minimum function value, which goes to 0 when T tends to infinity (assuming other quantities is bounded). This indeed provides us a point whose function value is close to the minimum, since by the convexity of f:

$$f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) - f(x^*) \le \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*).$$

One may expect that $\|x_t - x^*\|$ is decreasing in t. However, this is not true in general.

Lipschitzness of f

The gradient descent may not converge if the derivative of f is unbounded. Therefore, we usually assume f is L-Lipschitz, meaning $\|\nabla f\| \le L$. Then (1)

becomes to

$$\begin{aligned}\sum_{t=0}^{T-1}(f(x_t)-f(x^*)) &\le \frac{\phi(x_0)-\phi(x_{T+1})}{\eta}+\frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2\\ &\le \frac{\|x_0-x^*\|^2}{2\eta}+\frac{\eta T}{2}L^2\\ &\le \|x_0-x^*\|\cdot L\sqrt{T},\end{aligned}$$

by choosing $\eta = \frac{\|x_0-x^*\|}{L\sqrt{T}}$.

Constrained Case

The above analysis is for unconstrained optimization. If we require $x \in V$ for some *closed convex set* V, then we need to modify the updating rule of the gradient descent algorithm to

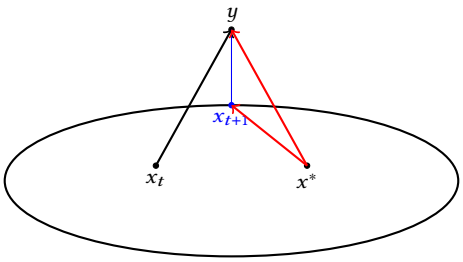
$$x_{t+1} = \Pi_V(x_t - \eta\nabla f(x_t)).$$

where $\Pi_V(\cdot)$ is the projection operator satisfying $\Pi_V(y) = \arg\min_{x\in V}\|x-y\|$. The algorithm is therefore called *projected gradient descent (PGD)*.

The bound (1) still holds for PGD. To see this, we only needs to verify that

$$\phi\left(\Pi_V(x_t-\eta\nabla f(x_t))\right) \le \phi\left(x_t-\eta\nabla f(x_t)\right), \tag{2}$$

then every step in the analysis of GD still holds.



As illustrated above, (2) is equivalent to $\|x_{t+1}-x^*\|^2 \le \|y-x^*\|^2$. Note that by the definition of the projection, $x_{t+1} = \arg\min_{x\in V} g(x)$ for $g(x) = \|x-y\|^2$. Therefore, by the first order optimality condition, we have $\langle x^*-x_{t+1}, \nabla g(x_{t+1})\rangle > 0$, which is equivalent to $\langle y-x_{t+1}, x^*-x_{t+1}\rangle \le 0$.

On the other hand, we have

$$\begin{aligned}\|y-x^*\|^2 &= \|y-x_{t+1}+x_{t+1}-x^*\|^2\\ &= \langle y-x_{t+1}+x_{t+1}-x^*, y-x_{t+1}+x_{t+1}-x^*\rangle\\ &= \|y-x_{t+1}\|^2+\|x_{t+1}-x^*\|^2-2\langle y-x_{t+1}, x^*-x_{t+1}\rangle\\ &\ge \|x_{t+1}-x^*\|^2.\end{aligned}$$

1.2 Online Gradient Descent

Recall the setting of online learning. Let $V \subseteq \mathbb{R}^d$ be the action space. The game lasts for T rounds, for each $s = 0, 1, \ldots, T-1$,

- The player picks some $x_s \in V$;
- The adversary picks some $\ell_s : V \to \mathbb{R}$;
- Pay the cost $\ell_s(x_s)$.

And our goal is to minimize $\sum_{t=0}^{T-1} \ell_t(x_t) - \sum_{t=0}^{T-1} \ell_t(x^*)$.

The *online projected gradient descent (OPGD)* algorithm is the following rule to pick x_s:

$$x_{s+1} = x_s - \eta_s \nabla \ell_t(x_s).$$

We can easily apply our previous analysis for gradient descent to this online version and obtain a similar regret bound. However, here we use a continuous approach to study it, and hopefully, get more intuition on why the bound is of that form.

To analyze the algorithm, we define a *continuous version* of it. We first fix some notations.

- For every $s = 0, 1, \ldots, T$, $\mathcal{T}_s := \sum_{i=0}^{s-1} \eta_i$;
- For every $0 \le t \le \mathcal{T}_T$, we let s_t be the unique integer s satisfying $\mathcal{T}_s \le t < \mathcal{T}_{s+1}$.
- For every $0 \le t \le \mathcal{T}_T$, let $g_t = \nabla \ell_{s_t}(x_{s_t})$ and $\hat{\eta}_t = \ell_{s_t}$.

A continuous version of the algorithm is

- $y_0 = x_0$;
- $\frac{\mathrm{d}y_t}{\mathrm{d}t} = -g_t$.

It is not hard to verify that:

Proposition 1 $x_s = y_{\mathcal{T}_s}$

Consider the function $\phi(y) = \frac{1}{2}\|y - x^*\|^2$. We can compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(y_t) = \langle y_t - x^*, \frac{\mathrm{d}}{\mathrm{d}t} y_t \rangle = \langle y_t - x^*, -g_t \rangle.$$

For every $s = 0, 1, \ldots, T-1$, integrate from $\mathcal{T}_s$ to $\mathcal{T}_{s+1}$:

$$\phi(y_{\mathcal{T}_{s+1}}) - \phi(y_{\mathcal{T}_s}) = \int_{\mathcal{T}_s}^{\mathcal{T}_{s+1}} \langle y_t - x^*, -g_t \rangle \mathrm{d}t.$$

Summing up above for $s = 0, 1, \ldots T-1$ and noting that $\mathcal{T}_0 = 0$, we have

$$\int_0^{\mathcal{T}_T} \langle g_t, y_t - x^* \rangle \mathrm{d}t = \phi(y_0) - \phi(y_{\mathcal{T}_T}).$$

Remember that we aim at analyzing the discrete process, and so we will compare $\sum_{s=0}^{T-1} \eta_s \langle \nabla \ell_s(x_s), x_s \rangle$ with $\int_0^{\mathcal{T}_T} \langle g_t, y_t \rangle \mathrm{d}t$. For every fixed $s = 0, 1, \ldots, T-1$, noting that for all $t \in [\mathcal{T}_s, \mathcal{T}_{s+1}]$, it holds $g_t = \nabla \ell_s(x_s)$.

Therefore we have

$$\begin{aligned}\eta_s\langle\nabla\ell_s(x_s), x_s\rangle - \int_{\mathcal{T}_s}^{\mathcal{T}_{s+1}}\langle g_t, y_t\rangle \mathrm{d}t &= \int_{\mathcal{T}_s}^{\mathcal{T}_{s+1}}\langle\nabla\ell_s(x_s), x_s - y_t\rangle \mathrm{d}t \\ &= \int_{\mathcal{T}_s}^{\mathcal{T}_{s+1}}\langle\nabla\ell_s(x_s), (t-\mathcal{T}_s)\nabla\ell_s(x_s)\rangle \mathrm{d}t \\ &= \|\nabla\ell_s(x_s)\|^2\int_{\mathcal{T}_s}^{\mathcal{T}_{s+1}}(t-\mathcal{T}_s)\mathrm{d}t \\ &= \frac{\eta_s^2\cdot\|\nabla\ell_s(x_s)\|^2}{2}.\end{aligned}$$

This is equivalent to

$$\langle\nabla\ell_s(x_s), x_s - x^*\rangle = \frac{\phi(x_s)-\phi(x_{s+1})}{\eta_s} + \frac{\eta_s\|\nabla\ell_s(x_s)\|^2}{2}.$$

The regret bound follows from the convexity of ℓ_s

$$\ell_s(x_s) - \ell_s(x^*) \le \langle\nabla\ell_s(x_s), x_s - x^*\rangle \le \frac{\phi(x_s)-\phi(x_{s+1})}{\eta_s} + \frac{\eta_s\|\nabla\ell_s(x_s)\|^2}{2}.$$

If we take all η_s to be η, and assume ℓ is L-Lipschitz, the regret bound turns into

$$\sum_{s=0}^{T-1}\frac{\phi(x_s)-\phi(x_{s+1})}{\eta_s} + \frac{\eta_s\|\nabla\ell_s(x_s)\|^2}{2} \le \mathrm{diam}(V)\cdot L\cdot\sqrt{T},$$

which is much worse than the performance of FTL algorithm when applied to the number guessing game.

Strongly Convex Function

In the above, we obtain the regret bound via the inequality

$$\ell_s(x_s) - \ell_s(x^*) \le \langle\nabla\ell_s(x_s), x_s - x^*\rangle$$

which follows from the convexity of ℓ_s. The above inequality is tight when ℓ_s is linear. If the function is *more convex* than the linear function, we can obtain stronger bounds.

We say a function $f : \mathbb{R}^n \to \mathbb{R}^n$ is μ-strongly convex if for every $x, y \in \mathbb{R}^n$, it holds that

$$f(y) \ge f(x) + \nabla f(x)^\top(y-x) + \frac{\mu}{2}\|y-x\|^2.$$

If ℓ_s is μ-strongly convex, then it holds that

$$\begin{aligned}\ell_s(x_s) - \ell_s(x^*) &\le \langle\nabla\ell_s(x_s), x_s - x^*\rangle - \frac{\mu}{2}\|x_s - x^*\|^2 \\ &\le \frac{\phi(x_s)-\phi(x_{s+1})}{\eta_s} + \frac{\eta_s\|\nabla\ell_s(x_s)\|^2}{2} - \frac{\mu}{2}\|x_s - x^*\|^2 \\ &= \left(\frac{1}{2\eta_s} - \frac{\mu}{2}\right)\|x_s - x^*\|^2 - \frac{1}{2\eta_s}\|x_{s+1} - x^*\|^2 + \frac{\eta_s}{2}\|\nabla\ell_s(x_s)\|^2\end{aligned}$$

Summing over $s = 0, 1, \ldots, T-1$ yields

$$\sum_{s=0}^{T-1} \ell_s(x_s) - \ell_s(x^*) \leq \left(\frac{1}{2\eta_0} - \frac{\mu}{2}\right) \|x_0 - x^*\|^2 - \frac{1}{2\eta_{T-1}} \|x_T - x^*\|^2 + \sum_{s=1}^{T-1} \left(\frac{1}{2\eta_s} - \frac{1}{2\eta_{s-1}} - \frac{\mu}{2}\right) \|x_s - x^*\|^2 + \sum_{s=0}^{T-1} \frac{\eta_s}{2} \|\nabla \ell_s(x_s)\|^2.$$

Therefore, in order for the RHS to cancel out, we can set $\frac{1}{\eta_s} = s \cdot \mu + \mu$. Then if all ℓ_s are L-Lipschitz, then

$$\sum_{s=0}^{T-1} \ell_s(x_s) - \ell_s(x^*) \leq \frac{L^2}{2\mu} H_T.$$

Applying the bound to the number guessing game with quadratic loss we met last week, we can match the performace of the follow-the-leader algorithm.