EEOR6616

Convex Optimization

Instructed by Prof. Tianyi Lin [*]
Transcribed by Satvik Saha [†]

Table of Contents

1. Basic Definitions

1.1. Convex Sets and Functions

Definition 1.1 (Convex set). We say that $\mathcal{K} \subseteq \mathbb{R}^d$ is convex if

$$\lambda x + (1 - \lambda) y \in \mathcal{K}$$

for all $x, y \in \mathcal{K}$ and $\lambda \in [0, 1]$.

Definition 1.2 (Convex function). We say that $f : \mathcal{K} \to \mathbb{R}$ is convex if $\mathcal{K}$ is convex, and

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

for all $x, y \in \mathcal{K}$ and $\lambda \in [0, 1]$.

[*]Department of Industrial Engineering and Operations Research (IEOR), Columbia University
[†]Department of Statistics, Columbia University

Proposition 1.3 (Jensen's Inequality). *f is convex if and only if*

$$f(\lambda_1 x_1 + \ldots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \ldots + \lambda_n f(x_n)$$

for all $x_1, \ldots, x_n \in \mathcal{K}$ and $\lambda_1, \ldots, \lambda_n \geq 0$ such that $\sum_k \lambda_k = 1$,

Definition 1.4 (Epigraph). The epigraph of $f : \mathcal{K} \to \mathbb{R}$ is defined as

$$\operatorname{epi}(f) = \{(x, \alpha) \in \mathcal{K} \times \mathbb{R} : f(x) \leq \alpha\}.$$

Remark. The epigraph of f is simply the region above the graph of f,

$$\Gamma(f) = \{(x, \alpha) \in \mathcal{K} \times \mathbb{R} : f(x) = \alpha\}.$$

Proposition 1.5. *f is convex if and only if $\operatorname{epi}(f)$ is convex.*

Proof. ($\Longrightarrow$) For $(x_1, \alpha_1), (x_2, \alpha_2) \in \operatorname{epi}(f)$ and $\lambda \in [0, 1]$, we have

$$\begin{aligned} f(\lambda x_1 + (1-\lambda)x_2) &\leq \lambda f(x_1) + (1-\lambda) f(x_2) \\ &\leq \lambda \alpha_1 + (1-\lambda)\alpha_2. \end{aligned}$$

($\Longleftarrow$) For $x_1, x_2 \in \mathcal{K}$ and $\lambda \in [0, 1]$, since $(x_1, f(x_1)), (x_2, f(x_2)) \in \operatorname{epi}(f)$, we have

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$ □

From now on, we will always assume that $f : \mathcal{K} \to \mathbb{R}$ is differentiable. Under this setting, we have a simpler characterization of convexity.

Proposition 1.6 (Gradient Inequality). *f is convex if and only if*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

for all $x, y \in \mathcal{K}$.

Proof. ($\Longrightarrow$) Note that for $t \in (0, 1)$, we may write

$$\begin{aligned} f(x) + \frac{f(x + t(y-x)) - f(x)}{t} &= \frac{f((1-t)x + ty) - (1-t)f(x)}{t} \\ &\leq f(y). \end{aligned}$$

Taking the limit $t \to 0$ gives the desired result.

($\Longleftarrow$) Let $x, y \in \mathcal{K}$ and $\lambda \in [0, 1]$. Setting $z = \lambda x + (1-\lambda)y$, we have

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z), \qquad f(y) \geq f(z) + \nabla f(z)^\top (y - z).$$

Combining these gives $\lambda f(x) + (1-\lambda) f(y) \geq f(z)$. □

Remark. This is often presented as

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y).$$

1.2. The Optimization Problem

Definition 1.7 (Global Minimizer)**.** We say that x^* is a global minimizer of $f : \mathcal{K} \to \mathbb{R}$ if $f(x) \geq f(x^*)$ for all $x \in \mathcal{K}$.

Definition 1.8 (Local Minimizer)**.** We say that x^* is a local minimizer of $f : \mathcal{K} \to \mathbb{R}$ if $f(x) \geq f(x^*)$ for all $x \in \mathcal{U}$ for some neighborhood $\mathcal{U} \subseteq \mathcal{K}$ of x^*.

Proposition 1.9. *Let $x^* \in \operatorname{int}(\mathcal{K})$ be a local minimizer of f. Then, $\nabla f(x^*) = 0$.*

The optimization problem for convex f on a convex set $\mathcal{K}$ can be described as

$$\min_{x \in \mathcal{K}} f(x). \tag{$\mathcal{M}_{\mathcal{K}}$}$$

In the special case $\mathcal{K} = \mathbb{R}^d$, this is

$$\min_{x \in \mathbb{R}^d} f(x). \tag{$\mathcal{M}_{\mathbb{R}^d}$}$$

The convexity of f allows us to characterize solutions of $(\mathcal{M}_{\mathbb{R}^d})$ via its critical points.

Proposition 1.10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then, $x^* \in \mathbb{R}^d$ is a global minimizer of f if and only if $\nabla f(x^*) = 0$.*

Proof. Follows directly from Proposition 1.9 and Proposition 1.6. □

2. Gradient Descent

Gradient descent algorithms for solving $(\mathcal{M}_{\mathbb{R}^d})$ follow the iterative scheme

$$x_{t+1} = x_t - \eta_t \nabla f(x_t). \tag{$\mathcal{GD}$}$$

It is possible for $(\mathcal{GD})$ to take our iterates x_t outside $\mathcal{K}$; we can rectify this using projections.

2.1. Projections

Theorem 2.1 (Hilbert Projection)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex. Then, for each $y \in \mathbb{R}^d$, there exists unique $z \in \mathcal{K}$ such that $\|z - y\| \leq \|x - y\|$ for all $x \in \mathcal{K}$.*

Proof. Set $\delta = \inf_{x \in \mathcal{K}} \|x - y\|$ and pick a sequence $\{z_n\} \subset \mathcal{K}$ such that $\|z_n - y\| \to \delta$. Note that $(z_n + z_m)/2 \in \mathcal{K}$; the parallelogram law gives

$$\begin{aligned} \|z_n - z_m\|^2 &= 2\|z_n - y\|^2 + 2\|z_m - y\|^2 - 4\|(z_n + z_m)/2 - y\|^2 \\ &\leq 2\|z_n - y\|^2 + 2\|z_m - y\|^2 - 4\delta^2. \end{aligned}$$

Since this goes to 0 as $m, n \to \infty$, $\{z_n\}$ is Cauchy and hence has a limit $z \in \mathcal{K}$. Furthermore, if $\delta = \|z' - y\|$ for some other $z' \in \mathcal{K}$, then

$$\|z - z'\|^2 = 4(\delta^2 - \|(z + z')/2 - y\|)^2 \leq 0,$$

forcing $z = z'$. □

Definition 2.2. Let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex. The projection onto $\mathcal{K}$ is defined by

$$\Pi_{\mathcal{K}} : \mathbb{R}^d \to \mathcal{K}, \quad y \mapsto \arg\min_{x \in \mathcal{K}} \|x - y\|.$$

Remark. Theorem 2.1 guarantees that $\Pi_{\mathcal{K}}$ is well defined; the minimizer of $x \mapsto \|x - y\|$ on $\mathcal{K}$ exists and is unique.

Proposition 2.3 (Variational Inequality). *Let $y \in \mathbb{R}^d$ and $z \in \mathcal{K}$ for closed convex $\mathcal{K}$. Then, $z = \Pi_{\mathcal{K}}(y)$ if and only if $\langle z - y, z - x \rangle \leq 0$ for all $x \in \mathcal{K}$.*

Proof. ($\Longrightarrow$) Let $t \in (0, 1)$, and $z_t = (1 - t)\Pi_{\mathcal{K}}(y) + tx \in \mathcal{K}$. Then,

$$\|z - y\|^2 \leq \|z_t - y\|^2 = \|z - y - t(z - x)\|^2,$$

which simplifies to

$$-2\langle z - y, z - x \rangle + t\|z - x\|^2 \geq 0.$$

Taking the limit $t \to 0$ gives the desired inequality.

($\Longleftarrow$) For $x \in \mathcal{K}$,

$$\|y - x\|^2 = \|y - z\|^2 + \|z - x\|^2 - 2\langle z - y, z - x \rangle \geq \|y - z\|^2.$$

□

Lemma 2.4 (Pythagoras). *For all $x \in \mathcal{K}$ and $y \in \mathbb{R}^d$,*

$$\|\Pi_{\mathcal{K}}(y) - x\|^2 \leq \|y - x\|^2 - \|y - \Pi_{\mathcal{K}}(y)\|^2.$$

Proof. It suffices to show that $\langle \Pi_{\mathcal{K}}(y) - y, \Pi_{\mathcal{K}}(y) - x \rangle \leq 0$ for all $x \in \mathcal{K}$, which holds via Proposition 2.3. □

Corollary 2.4.1. *For all $x, y \in \mathbb{R}^d$,*

$$\|\Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(y)\| \leq \|x - y\|.$$

Projected gradient descent algorithms for solving $(\mathcal{M}_{\mathcal{K}})$ follow the iterative scheme

$$\begin{aligned} y_{t+1} &= x_t - \eta_t \nabla f(x_t), \\ x_{t+1} &= \Pi_{\mathcal{K}}(y_{t+1}). \end{aligned} \tag{$\mathcal{PGD}$}$$

We can establish rates of convergence of $(\mathcal{GD})$ and $(\mathcal{PGD})$ under certain regularity conditions on f.

2.2. L-Lipschitz Functions

Definition 2.5 (L-Lipschitz)**.** We say that $f : \mathcal{K} \to \mathbb{R}$ is L-Lipschitz for some $L \geq 0$ if

$$|f(x) - f(y)| \leq L\|x - y\|$$

for all $x, y \in \mathcal{K}$.

Remark. When f is differentiable, f is L-Lipschitz if and only if $\|\nabla f\| \leq L$.

Theorem 2.6. *Let f be convex and L-Lipschitz, $x^* \in \mathcal{K}$ be its global minimizer, and $\|x_1 - x^*\| \leq R$. Further let $x_1, \dots, x_T$ be T iterates of ($\mathcal{PGD}$) with $\eta = R/L\sqrt{T}$. Then,*

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

Proof. Compute

$$\begin{aligned}
f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) &\leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) && \text{(Proposition 1.3)} \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \nabla f(x_t)^\top (x_t - x^*) && \text{(Proposition 1.6)} \\
&= \frac{1}{T\eta}\sum_{t=1}^{T} (x_t - y_{t+1})^\top (x_t - x^*) \\
&= \frac{1}{2T\eta}\sum_{t=1}^{T} \left[\|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right] \\
&= \frac{\eta}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2T\eta}\sum_{t=1}^{T} \left[\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right] \\
&\leq \frac{\eta L^2}{2} + \frac{1}{2T\eta}\sum_{t=1}^{T} \left[\|x_t - x^*\|^2 - \|\underbrace{\Pi_{\mathcal{K}}(y_{t+1})}_{x_{t+1}} - x^*\|^2\right] && \text{(Lemma 2.4)} \\
&= \frac{\eta L^2}{2} + \frac{1}{2T\eta}\left[\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2\right] \\
&\leq \frac{\eta L^2}{2} + \frac{R^2}{2T\eta} \\
&= \frac{RL}{\sqrt{T}}.
\end{aligned}$$

□

2.3. ℓ-smoothness

Definition 2.7 (ℓ-smoothness). We say that $f : \mathcal{K} \to \mathbb{R}$ is ℓ-smooth for some $\ell \geq 0$ if

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|$$

for all $x, y \in \mathcal{K}$.

Lemma 2.8. *Let $f : \mathcal{K} \to \mathbb{R}$ for convex $\mathcal{K}$ be ℓ-smooth. Then,*

$$|f(y) - f(x) - \nabla f(x)^\top (y - x)| \leq \frac{\ell}{2}\|y - x\|^2.$$

Proof. Using the Fundamental Theorem of Calculus,

$$\begin{aligned}
|f(y) - f(x) - \nabla f(x)^\top (y - x)| &= \left| \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^\top (y - x)\, dt \right| \\
&\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\|\, dt \\
&\leq \int_0^1 \ell t \|y - x\| \cdot \|y - x\|\, dt \\
&= \frac{\ell}{2}\|y - x\|^2.
\end{aligned}$$

□

When f is convex, the norm on the left hand side is redundant, giving the estimate

$$0 \leq f(y) - f(x) - \nabla f(x)^\top (y - x) \leq \frac{\ell}{2}\|y - x\|^2.$$

In fact, we can use ℓ-smoothness to improve upon the estimate in Proposition 1.6.

Lemma 2.9. *Let f be convex and ℓ-smooth. Then,*

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{1}{2\ell}\|\nabla f(x) - \nabla f(y)\|^2.$$

Proof. Set $z = y + (\nabla f(x) - \nabla f(y))/\ell$. Using Proposition 1.6, Lemma 2.8,

$$\begin{aligned}
f(x) - f(y) &= (f(x) - f(z)) + (f(z) - f(y)) \\
&\leq \nabla f(x)^\top (x - z) + \nabla f(y)^\top (z - y) + \frac{\ell}{2}\|z - y\|^2 \\
&= \nabla f(x)^\top (x - y) + (\nabla f(y) - \nabla f(x))^\top (z - y) + \frac{\ell}{2}\|z - y\|^2 \\
&= \nabla f(x)^\top (x - y) - \frac{1}{\ell}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{1}{2\ell}\|\nabla f(x) - \nabla f(y)\|^2 \\
&= \nabla f(x)^\top (x - y) - \frac{1}{2\ell}\|\nabla f(x) - \nabla f(y)\|^2.
\end{aligned}$$

□

Corollary 2.9.1. *Let f be convex and ℓ-smooth. Then,*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{\ell}\|\nabla f(x) - \nabla f(y)\|^2.$$

Theorem 2.10. *Let f be convex and ℓ-smooth, $x^* \in \mathbb{R}^d$ be its global minimizer. Further let $\{x_t\}_{t\in\mathbb{N}}$ be iterates of $(\mathcal{GD})$ with $\eta = 1/\ell$. Then,*

$$\|x_{t+1} - x^*\| \leq \|x_t - x^*\|$$

for all $t \in \mathbb{N}$.

Proof. Using $\nabla f(x^*) = 0$ and Corollary 2.9.1,

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_{t+1} - x_t\|^2 + 2(x_{t+1} - x_t)^\top (x_t - x^*) + \|x_t - x^*\|^2 \\
&= \frac{1}{\ell^2}\|\nabla f(x_t)\|^2 - \frac{2}{\ell}\nabla f(x_t)^\top (x_t - x^*) + \|x_t - x^*\|^2 \\
&\leq \frac{1}{\ell^2}\|\nabla f(x_t)\|^2 - \frac{2}{\ell^2}\|\nabla f(x_t)\|^2 + \|x_t - x^*\|^2 \\
&= -\frac{1}{\ell^2}\|\nabla f(x_t)\|^2 + \|x_t - x^*\|^2 \\
&\leq \|x_t - x^*\|^2.
\end{aligned}$$

□

Remark. This remains true with $(\mathcal{PGD})$ as long as $x^* \in \operatorname{int}(\mathcal{K})$, via

$$\|x_{t+1} - x^*\| = \|\Pi_{\mathcal{K}}(y_{t+1}) - x^*\| \leq \|y_{t+1} - x^*\|.$$

Theorem 2.11. *Let f be convex and ℓ-smooth, $x^* \in \mathbb{R}^d$ be its global minimizer, and $\|x_1 - x^*\| \leq R$. Further let $x_1, \ldots, x_T$ be T iterates of $(\mathcal{GD})$ with $\eta = 1/\ell$. Then,*

$$f(x_T) - f(x^*) \leq \frac{2R^2\ell}{T-1}.$$

Proof. Using Lemma 2.8, note that

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\
&= -\frac{1}{2\ell}\|\nabla f(x_t)\|^2.
\end{aligned}$$

Setting $\delta_t = f(x_t) - f(x^*)$, this reads

$$\delta_{t+1} \leq \delta_t - \frac{1}{2\ell}\|\nabla f(x)\|^2.$$

Now,

$$\delta_t \leq \nabla f(x_t)^\top (x_t - x^*) \leq \|\nabla f(x_t)\|\|x_t - x^*\| \leq \|\nabla f(x_t)\|\|x_1 - x^*\|,$$

with the last inequality guaranteed by Theorem 2.10. Setting $w = 1/2\ell\|x_1 - x^*\|^2$, this is $\|\nabla f(x_t)\|^2/2\ell \geq w\delta_t^2$. Thus, $\delta_{t+1} \leq \delta_t - w\delta_t^2$, which rearranges to

$$\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \geq w\frac{\delta_t}{\delta_{t+1}} \geq w.$$

Summing over t gives $1/\delta_T \geq w(T-1)$, which is the desired estimate. □

Remark. We have shown that

$$\frac{1}{\ell}\|\nabla f(x_t)\|^2 \leq f(x_t) - f(x_{t+1}) \leq \frac{1}{2\ell}\|\nabla f(x_t)\|^2.$$