

Convex Optimization

Instructed by *Prof. Tianyi Lin*^{*}

Transcribed by *Satvik Saha*[†]

Table of Contents

1. Basic Definitions	1
1.1. Convex Sets and Functions	1
1.2. The Optimization Problem	3
2. Projections	4
2.1. Normals	5
2.2. Subdifferentials	7
3. Gradient Descent	8
3.1. L -Lipschitz Functions	8
3.2. ℓ -smoothness	9
Bibliography	11

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 $\lambda \in [0, 1]$.

Example 1.1.1. All linear subspaces of \mathbb{R}^d are convex sets.

^{*}Department of Industrial Engineering and Operations Research (IEOR), Columbia University

[†]Department of Statistics, Columbia University

Example 1.1.2. Consider points $x_1, \dots, x_n \in \mathbb{R}^d$. Their *convex hull*, described by

$$\text{conv}(x_1, \dots, x_n) = \left\{ \lambda_1 x_1 + \dots + \lambda_n x_n : \lambda_1, \dots, \lambda_n \geq 0, \sum_{i=1}^n \lambda_i = 1 \right\},$$

is a convex set. In fact, it is the smallest convex set containing x_1, \dots, x_n .

Definition 1.2 (Convex function). We say that $f : \mathcal{K} \rightarrow \mathbb{R}$ 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 $\lambda \in [0, 1]$.

Example 1.2.1. The map $x \mapsto x^2$ is convex.

Example 1.2.2. Indicator functions of convex sets are convex. The indicator function of $\mathcal{X} \subseteq \mathbb{R}^d$ is given by

$$I_{\mathcal{X}} : \mathbb{R}^d \rightarrow \overline{\mathbb{R}}, \quad x \mapsto \begin{cases} 0 & \text{if } x \in \mathcal{X} \\ \infty & \text{if } x \notin \mathcal{X} \end{cases}.$$

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

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

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

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

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

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 $\text{epi}(f)$ is convex.

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

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

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

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

From now on, we will always assume that $f : \mathcal{K} \rightarrow \mathbb{R}$ is differentiable, unless stated otherwise. 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. (\implies) 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 $t \rightarrow 0$ gives the desired result.

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

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

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

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 $f : \mathcal{K} \rightarrow \mathbb{R}$ 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 $f : \mathcal{K} \rightarrow \mathbb{R}$ 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 $x^* \in \text{int}(\mathcal{K})$ 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

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

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

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

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

Proposition 1.10. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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. \square

2. Projections

Definition 2.1. We say that z is a projection of a point y onto a set \mathcal{X} if $z \in \mathcal{X}$ and $\|y - z\| \leq \|y - x\|$ for all $x \in \mathcal{X}$.

In other words, z is a projection of y onto \mathcal{X} when $z \in \arg \min_{x \in \mathcal{X}} \|y - x\|$. In general, such projections of points need not exist! For instance, one can argue that a projection of $y \notin \mathcal{X}$ onto \mathcal{X} cannot lie in $\text{int}(\mathcal{X})$.

Example 2.1.1. Consider the open unit disk $\mathbb{D}^2 = \{x \in \mathbb{R}^2 : \|x\| < 1\}$ in \mathbb{R}^2 . Projections of points outside \mathbb{D}^2 onto \mathbb{D}^2 do not exist.

In Euclidean spaces \mathbb{R}^d , we may observe that closedness of (nonempty) \mathcal{X} guarantees the existence of a projection of $y \in \mathbb{R}^d$ onto \mathcal{X} . By picking some $x_0 \in \mathcal{X}$, we need only look at the compact set $\mathcal{X} \cap \overline{B_r(y)}$ where $r = \|y - x_0\|$, on which the continuous map $x \mapsto \|y - x\|$ must attain its minimum.

On the other hand, projections of points need not be unique.

Example 2.1.2. Consider the unit circle $S^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$ in \mathbb{R}^2 . Then, every point in S^1 is a projection of $0 \in \mathbb{R}^2$ onto S^1 .

Theorem 2.2 (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 $\|z_n - y\| \rightarrow \delta$. 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 $m, n \rightarrow \infty$, $\{z_n\}$ 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 $z = z'$. \square

Definition 2.3. Let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex. The projection operator onto \mathcal{K} is defined by

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

Remark. Theorem 2.2 guarantees that $\Pi_{\mathcal{K}}$ is well defined; the minimizer of $x \mapsto \|x - y\|$ on \mathcal{K} exists and is unique.

Proposition 2.4 (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. (\implies) 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 $t \rightarrow 0$ gives the desired inequality.

(\impliedby) 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 2.5 (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 2.4. □

Corollary 2.5.1. *For all $x, y \in \mathbb{R}^d$,*

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

2.1. Normals

A very useful property of closed convex sets \mathcal{K} is that given a point $w \notin \mathcal{K}$, one can find a hyperplane separating w from \mathcal{K} . In other words, there exists a continuous linear functional g and a constant a such that $g(x) < a < g(w)$ for all $x \in \mathcal{K}$.

Theorem 2.6 (Strict Separation). *Let $w \notin \mathcal{K}$ for closed convex \mathcal{K} . There exists $v \neq 0$ such that*

$$\sup_{x \in \mathcal{K}} \langle v, x \rangle < \langle v, w \rangle.$$

Proof. Set $v = w - \Pi_{\mathcal{K}}(w)$. Then, [Proposition 2.4](#) gives

$$\langle v, x - (w - v) \rangle = \langle w - \Pi_{\mathcal{K}}(w), x - \Pi_{\mathcal{K}}(w) \rangle \leq 0,$$

for all $x \in \mathcal{K}$, which rearranges into

$$\langle v, x \rangle + \|v\|^2 \leq \langle v, w \rangle. \quad \square$$

Definition 2.7 (Normal). Let $x \in \mathcal{K}$ for closed convex \mathcal{K} . We say that v is normal to \mathcal{K} at x if $\langle v, y \rangle \leq \langle v, x \rangle$ for all $y \in \mathcal{K}$.

Definition 2.8 (Normal Cone). Let $x \in \mathcal{K}$ for closed convex \mathcal{K} . The normal cone $N_{\mathcal{K}}(x)$ at x is the collection of normals to \mathcal{K} at x .

Note that if v is normal to \mathcal{K} at x , so is αv for $\alpha \geq 0$, hence $N_{\mathcal{K}}(x)$ is indeed a cone; it is also convex. Furthermore, $N_{\mathcal{K}}(x)$ is nontrivial only when $x \notin \text{int}(X)$; if $x \in B_{\delta}(x) \subseteq \mathcal{K}$, then for any v with $\|v\| = 1$, we have $x \pm \frac{\delta}{2}v \in B_{\delta}(x) \subseteq \mathcal{K}$, and

$$\langle v, x - \frac{\delta}{2}v \rangle = \langle v, x \rangle - \frac{\delta}{2} < \langle v, x \rangle < \langle v, x \rangle + \frac{\delta}{2} = \langle v, x + \frac{\delta}{2}v \rangle.$$

Thus, we need only look at normal cones at boundary points $x \in \partial\mathcal{K}$. At these points, nonzero $v \in N_{\mathcal{K}}(x)$ describe *supporting hyperplanes* to \mathcal{K} at x .

Proposition 2.9. Let $x \in \partial\mathcal{K}$ for closed convex $K \subseteq \mathbb{R}^d$. Then, $N_{\mathcal{K}}(x)$ is nontrivial, i.e. there exists a supporting hyperplane to \mathcal{K} at x .

Proof. Pick a sequences $\{x_n\} \subseteq \mathcal{K}^c$ such that $x_n \rightarrow x$, and a corresponding sequence $\{v_n\} \subset S^{d-1}$ of directions via [Theorem 2.6](#), such that $\sup_{y \in \mathcal{K}} \langle v_n, y \rangle < \langle v_n, x_n \rangle$. Using the compactness of S^{d-1} , descend to a subsequence and relabel so that $v_n \rightarrow v \in S^{d-1}$. Then, for $y \in K$, we have

$$\langle v, y \rangle = \lim_{n \rightarrow \infty} \langle v_n, y \rangle \leq \lim_{n \rightarrow \infty} \langle v_n, x_n \rangle = \langle v, x \rangle. \quad \square$$

Proposition 2.10. Let $x \in \mathcal{K}$ for closed convex \mathcal{K} , and let $v \in N_{\mathcal{K}}(x)$. Then, $\Pi_{\mathcal{K}}(x + \alpha v) = x$ for all $\alpha \geq 0$.

Proof. For all $y \in \mathcal{K}$, we have

$$\langle x - (x + \alpha v), x - y \rangle = \alpha \langle v, y - x \rangle \leq 0,$$

whence $x = \Pi_{\mathcal{K}}(x + \alpha v)$ by [Proposition 2.4](#). \square

2.2. Subdifferentials

Definition 2.11 (Subdifferential). Let $f : \mathcal{K} \rightarrow \mathbb{R}$ be convex. The subdifferential of f at $x \in \mathcal{K}$ is the collection of all directions v such that

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

for all $y \in \mathcal{K}$, and is denoted $\partial f(x)$.

Compare with the gradient inequality (Proposition 1.6) for differentiable convex f .

Example 2.11.1. Consider $f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto |x|$. Then,

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ \{+1\} & \text{if } x > 0 \end{cases}$$

It is clear that the subgradient $\partial f(x)$ is convex. Showing that it is nontrivial requires more work.

Proposition 2.12. Let $f : \mathcal{K} \rightarrow \mathbb{R}$ be convex. Then, $\partial f(x)$ is nonempty for all $x \in \text{ri}(\mathcal{K})$.

Proof. Note that $\text{epi}(f)$ is convex via Proposition 1.5. Use Proposition 2.9 to find a supporting hyperplane to $\text{epi}(f)$ at $(x^\top f(x))^\top$, i.e. $(v^\top s)^\top \neq 0$ such that for all $(y^\top \alpha)^\top \in \text{epi}(f)$,

$$v^\top(y - x) + s(\alpha - f(x)) \leq 0.$$

By considering $y = x$ and $\alpha > f(x)$, we must have $s \leq 0$. If $s = 0$, we would need $v^\top(y - x) \leq 0$ for all $y \in \mathcal{K}$, which would force $v = 0$ since $x \in \text{ri}(\mathcal{K})$. Thus, $s < 0$; putting $\alpha = f(y)$, we have

$$f(y) \geq f(x) - \frac{v^\top}{s}(y - x),$$

whence $-v^\top/s \in \partial f(x)$. □

The next result follows immediately from the definition of the subdifferential; compare this with Proposition 1.10.

Proposition 2.13. Let $f : \mathcal{K} \rightarrow \mathbb{R}$ be convex. Then, $x^* \in \mathcal{K}$ is a global minimizer of f if and only if $0 \in \partial f(x^*)$.

When f is differentiable at $x \in \text{int}(\mathcal{K})$, the subgradient reduces to the usual gradient, with $\partial f(x) = \{\nabla f(x)\}$. Indeed, Proposition 1.6 shows that $\nabla f(x) \in \partial f(x)$. To check that there are no other elements, pick $v \in \partial f(x)$, and note that for $\lambda \geq 0$,

$$v^\top u \leq \frac{f(x + \lambda u) - f(x)}{\lambda} \rightarrow \nabla f(x)^\top u \quad \text{as } \lambda \rightarrow 0,$$

hence $(\nabla f(x) - v)^\top u \geq 0$ for all directions u . This forces $v = \nabla f(x)$.

The converse of the above result also holds, in the following form.

Theorem 2.14. Let $f : \mathcal{K} \rightarrow \mathbb{R}$ be convex and $x \in \text{int}(\mathcal{K})$. If f is differentiable at x , then $\partial f(x) = \{\nabla f(x)\}$. Conversely, if $\partial f(x) = \{v\}$, then f is differentiable at x with $\nabla f(x) = v$.

3. Gradient Descent

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

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

It is possible for (\mathcal{GD}) to take our iterates x_t outside \mathcal{K} ; we can rectify this using projections. Projected gradient descent algorithms for solving $(\mathcal{M}_{\mathcal{K}})$ follow the iterative scheme

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

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

3.1. L -Lipschitz Functions

Definition 3.1 (L -Lipschitz). We say that $f : \mathcal{K} \rightarrow \mathbb{R}$ 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 3.2. 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

$$f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^T f(x_t) - f(x^*) \quad (\text{Proposition 1.3})$$

$$\leq \frac{1}{T} \sum_{t=1}^T \nabla f(x_t)^\top (x_t - x^*) \quad (\text{Proposition 1.6})$$

$$= \frac{1}{T\eta} \sum_{t=1}^T (x_t - y_{t+1})^\top (x_t - x^*)$$

$$= \frac{1}{2T\eta} \sum_{t=1}^T \left[\|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2 \right]$$

$$\begin{aligned}
&= \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2T\eta} \sum_{t=1}^T \left[\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2 \right] \\
&\leq \frac{\eta L^2}{2} + \frac{1}{2T\eta} \sum_{t=1}^T \left[\|x_t - x^*\|^2 - \underbrace{\|\Pi_{\mathcal{K}}(y_{t+1}) - x^*\|^2}_{x_{t+1}} \right] \quad (\text{Lemma 2.5}) \\
&= \frac{\eta L^2}{2} + \frac{1}{2T\eta} \left[\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2 \right] \\
&\leq \frac{\eta L^2}{2} + \frac{R^2}{2T\eta} \\
&= \frac{RL}{\sqrt{T}}. \quad \square
\end{aligned}$$

3.2. ℓ -smoothness

Definition 3.3 (ℓ -smoothness). We say that $f : \mathcal{K} \rightarrow \mathbb{R}$ 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 3.4. Let $f : \mathcal{K} \rightarrow \mathbb{R}$ 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,

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

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 3.5. 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 3.4,

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

Corollary 3.5.1. 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 3.6. 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 3.5.1,

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

Theorem 3.7. 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, \dots, 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 3.4, 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

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

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 3.6. 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. \square

Remark. We have shown that

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

Bibliography

- [1] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.
- [2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [3] Y. Nesterov, *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2018.