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

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

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

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

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

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

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

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

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

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

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

for all $x_1, ..., x_n \in \mathcal{K}$ and $\lambda_1, ..., \lambda_n \geq 0$ 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

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

($\Longleftarrow$) 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

$$
f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2). \qquad \square
$$

From now on, we will always assume that $f : \mathcal{K} \to \mathbb{R}$ 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*

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

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

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

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

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

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

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

Remark. This is often presented as

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

1.2. Projections

Definition 1.7. We say that z is a projection of a point y onto a set $\mathcal{X}$ if $z \in \mathcal{X}$ and $\|y - z\| \le \|y - x\|$ 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 the interior of $\mathcal{X}$: given $z \in B_\delta(z) \subseteq \operatorname{int}(\mathcal{X})$, set $z_t = z + t(y - z) \in \mathcal{X}$ with $t = \delta / (2\|y - z\|)$, whence $\|y - z_t\| = (1-t)\|y - z\| < \|y - z\|$.

Example 1.7.1. 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 1.7.2. 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.

The following theorem establishes the existence and uniqueness of projections onto closed convex sets in any Hilbert space; we focus on Euclidean spaces $\mathbb{R}^d$ for simplicity.

Theorem 1.8 (Hilbert Projection). *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex. Then, for each $y \in \mathbb{R}^d$, there exists a unique projection of y onto $\mathcal{X}$.*

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

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

Remark. Theorem 1.8 guarantees that $\Pi_{\mathcal{K}}$ is well defined; the minimizer of $x \mapsto \|x - y\|$ on $\mathcal{K}$ exists and is unique.

Proposition 1.10 (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 1.11 (Pythagoras). *For all $x \in \mathcal{K}$ and $y \in \mathbb{R}^d$,*

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

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 1.10. □

Corollary 1.11.1. *For all $x, y \in \mathbb{R}^d$,*

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

1.3. Normals

A very useful property of closed convex sets $\mathcal{K}$ is that given a point $w \notin K$, 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 1.12 (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 1.10 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

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

Definition 1.13 (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 1.14 (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 $x \notin \operatorname{int}(X)$; 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 1.15. *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 $x_n \to x$, and a corresponding sequence $\{v_n\} \subset S^{d-1}$ of directions via Theorem 1.12, 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 $v_n \to v \in S^{d-1}$. Then, for $y \in K$, we have

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

□

Proposition 1.16. *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 1.10. □

2. The Convex Optimization Problem

Definition 2.1 (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 2.2 (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 2.3. *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

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

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

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

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

Proposition 2.4. *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 2.3 and Proposition 1.6. □

Corollary 2.4.1. *Local minimizers of convex functions are global minimizers.*

3. Gradient Descent

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

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

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

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

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 $f : \mathcal{K} \to \mathbb{R}$ 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 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 $x_1, \ldots, x_T$ 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

$$\begin{aligned} f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) &\leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) && \text{(Proposition 1.3)} \\ &\leq \frac{1}{T}\sum_{t=1}^{T} \nabla f(x_t)^\top (x_t - x^*) && \text{(Proposition 1.6)} \\ &= \frac{1}{T\eta}\sum_{t=1}^{T} (x_t - y_{t+1})^\top (x_t - x^*) \\ &= \frac{1}{2T\eta}\sum_{t=1}^{T} \left[\|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{\eta}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2T\eta}\sum_{t=1}^{T}\left[\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right] \\
&\le \frac{\eta L^2}{2} + \frac{1}{2T\eta}\sum_{t=1}^{T}\Big[\|x_t - x^*\|^2 - \|\underbrace{\Pi_{\mathcal{K}}(y_{t+1})}_{x_{t+1}} - x^*\|^2\Big] && \text{(Lemma 1.11)} \\
&= \frac{\eta L^2}{2} + \frac{1}{2T\eta}\left[\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2\right] \\
&\le \frac{\eta L^2}{2} + \frac{R^2}{2T\eta} \\
&= \frac{RL}{\sqrt{T}}.
\end{aligned}
$$

□

3.2. ℓ-smoothness

Definition 3.3 (ℓ-smoothness)**.** We say that $f : \mathcal{K} \to \mathbb{R}$ is ℓ-smooth for some $\ell \ge 0$ if

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

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

Lemma 3.4. *Let $f : \mathcal{K} \to \mathbb{R}$ for convex $\mathcal{K}$ be ℓ-smooth. Then,*

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

Proof. Using the Fundamental Theorem of Calculus,

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

□

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

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

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,*

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

Proof. Set $z = y + (\nabla f(x) - \nabla f(y))/\ell$. Using Proposition 1.6, Lemma 3.4,

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

□

Corollary 3.5.1. *Let f be convex and ℓ-smooth. Then,*

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

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,*

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

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

Proof. Using $\nabla f(x^*) = 0$ and Corollary 3.5.1,

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

□

Theorem 3.7. *Let f be convex and ℓ-smooth, $x^* \in \mathbb{R}^d$ be its global minimizer, and $\|x_1 - x^*\| \le R$. Further let $x_1, \ldots, x_T$ be T iterates of $(\mathcal{GD})$ with $\eta = 1/\ell$. Then,*

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

Proof. Using Lemma 3.4, note that

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

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 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. □

Remark. We have shown that

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

3.3. α-strong Convexity

Definition 3.8 (α-strong Convex Function)**.** We say that convex differentiable f is α-strongly convex for $\alpha \geq 0$ if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2$$

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

Remark. This is often presented as

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

Thus, α-strong convexity is a strengthening of the gradient inequality (Proposition 2.4).

Example 3.8.1. All convex functions are ‘0-strongly convex’.

We can improve upon Theorem 3.2 and Theorem 3.6 dramatically with this added assumption.

Theorem 3.9. *Let f be α-strongly convex and L-Lipschitz, and let $x^* \in \mathcal{K}$ be its global minimizer. Further let $x_1, \ldots, x_T$ be T iterates of* ($\mathcal{PGD}$) *with $\eta_t = 2/(\alpha(t+1))$. Then,*

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

Note that when f is both α-strongly convex and ℓ-smooth, we have

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

This also justifies that $\alpha \leq \ell$.

Lemma 3.10. *Let f be α-strongly convex and ℓ-smooth, and let $x^+ = x - \frac{1}{\ell}\nabla f(x)$. Then,*

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

Proof. Write

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

□

Theorem 3.11. *Let f be α-strongly convex and ℓ-smooth, and let $x^* \in \mathbb{R}^d$ be its global minimizer. Further let $\{x_t\}_{t\in\mathbb{N}}$ be iterates of* ($\mathcal{GD}$) *with $\eta = 1/\ell$. Then,*

$$\left\|x_{t+1} - x^*\right\|^2 \leq e^{-t\alpha/\ell}\left\|x_1 - x^*\right\|^2$$

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

Proof. Write

$$\begin{aligned}
\left\|x_{t+1} - x^*\right\|^2 &= \left\|x_{t+1} - x_t\right\|^2 + \left\|x_t - x^*\right\|^2 + 2(x_{t+1} - x_t)^\top (x_t - x^*) \\
&= \frac{1}{\ell^2}\|\nabla f(x_t)\|^2 + \|x_t - x^*\|^2 - \frac{2}{\ell}\nabla f(x_t)^\top (x_t - x^*) \\
&\leq \frac{1}{\ell^2}\|\nabla f(x_t)\|^2 + \|x_t - x^*\|^2 \\
&\qquad -\frac{2}{\ell}\left[f(x_{t+1}) - f(x^*) + \frac{1}{2\ell}\|\nabla f(x_t)\|^2 + \frac{\alpha}{2}\|x_t - x^*\|^2\right] && \text{(Lemma 3.10)} \\
&\leq \|x_t - x^*\|^2 - \frac{\alpha}{\ell}\|x_t - x^*\|^2 && (f(x_{t+1}) \geq f(x^*))
\end{aligned}$$

$$= \left(1 - \frac{\alpha}{\ell}\right) \|x_t - x^*\|^2.$$

Iterating and using $1 - s \leq e^{-s}$, we have

$$\left\|x_{t+1} - x^*\right\|^2 \leq \left(1 - \frac{\alpha}{\ell}\right)^t \|x_1 - x^*\|^2 \leq e^{-t\alpha/\ell} \|x_1 - x^*\|^2. \qquad \square$$

A version of the above still holds with regards to ($\mathcal{PGD}$).

The quantity $\kappa = \ell/\alpha \geq 1$, called the *conditional number*, controls the rate of convergence of ($\mathcal{GD}$). Convergence is especially slow when κ is very high.

Example 3.11.1. Let $f(x) = \frac{1}{2}x^\top Ax$ for positive definite A. Then, ℓ and α are the largest and smallest eigenvalues of A respectively.

4. Momentum-Based Gradient Descent

4.1. Polyak's Heavy Ball Method

Polyak's heavy ball method follows the iterative scheme

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) + \beta_t(x_t - x_{t-1}). \qquad (\text{HB-}\mathcal{GD})$$

Remark. The (HB-$\mathcal{GD}$) method can be viewed as a discretized version of the *heavy ball flow*

$$\ddot{x} + \gamma \dot{x} = -\nabla f(x).$$

Lemma 4.1. *Given $M \in \mathbb{R}^{d\times d}$ and $\varepsilon > 0$, there exists a norm $\|\cdot\|_\varepsilon$ such that $\|M\|_\varepsilon \leq \rho(M) + \varepsilon$, where*

$$\rho(M) = \max\{|\lambda_1|, .., |\lambda_n|\}$$

is the spectral radius *of M, and $\lambda_1, ..., \lambda_n$ are the eigenvalues of M.*

Remark. Recall that every norm $\|\cdot\|$ on $\mathbb{R}^d$ naturally induces a matrix norm

$$\|M\| = \sup\{\|Mx\| : \|x\| = 1\}$$

on $\mathbb{R}^{d\times d}$. The spectral radius satisfies $\rho(A) \leq \|A\|$ for every natural matrix norm $\|\cdot\|$. The above lemma shows that

$$\rho(M) = \inf\{\|M\| : \|\cdot\| \text{ is a matrix norm}\}.$$

Theorem 4.2. *Let $f(x) = \frac{1}{2}x^\top Ax$ for positive definite $A \in \mathbb{R}^{d\times d}$, and let $\{x_t\}_{t\in\mathbb{N}}$ be iterates of (HB-$\mathcal{GD}$) with*

$$\eta = \left(\frac{2}{\sqrt{\ell}+\sqrt{\alpha}}\right)^2, \quad \beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2, \quad \kappa = \frac{\ell}{\alpha},$$

where ℓ, α are the largest and smallest eigenvalues of A. Then, for every $\varepsilon > 0$, there exists a norm $\|\cdot\|_\varepsilon$ such that

$$\left\|\begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix}\right\|_\varepsilon \le \left(\sqrt{\beta}+\varepsilon\right)^t \left\|\begin{pmatrix} x_1 \\ x_0 \end{pmatrix}\right\|_\varepsilon$$

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

Proof. Note that $\nabla f(x) = Ax$, so the (HB-$\mathcal{GD}$) updates read

$$x_{t+1} = x_t - \eta Ax_t + \beta(x_t - x_{t-1}) = ((1+\beta)I_d - \eta A)x_t - \beta x_{t-1},$$

which can be rewritten as

$$\begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} = \begin{pmatrix} (1+\beta)I_d - \eta A & -\beta I_d \\ I_d & 0 \end{pmatrix}\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}.$$

Notate this as

$$X_{t+1} = BX_t = B^t X_1.$$

Since $\prod_j |\nu_j| = |\det(B)| = \beta^d$ for eigenvalues $\{\nu_j\}_{j=1}^{2d}$ of B, we must have $\rho(B) = \max_j |\nu_j| \ge \sqrt{\beta}$. The eigenvalue equation for B reads

$$B\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} (1+\beta)y - \eta Ay - \beta z \\ y \end{pmatrix} = \nu\begin{pmatrix} y \\ z \end{pmatrix} \iff \begin{cases} \eta\nu Az = (\beta + (1+\beta)\nu - \nu^2)z \\ y = \nu z \end{cases}.$$

Thus, the eigenvalues $\{\lambda_i\}_{i=1}^d$ of A and $\{\nu_{2i-1}, \nu_{2i}\}_{i=1}^d$ of B are related via $\eta\lambda\nu = \beta + (1+\beta)\nu - \nu^2$, or

$$\nu_{2i-1,2i} = \frac{1}{2}\left(1 + \beta - \eta\lambda_i \pm \sqrt{(1+\beta-\eta\lambda_i)^2 - 4\beta}\right).$$

Note that when $\Delta_i = (1+\beta-\eta\lambda_i)^2 - 4\beta \le 0$, we have $|\nu_{2i-1}| = |\nu_{2i}| = \sqrt{\beta}$. Thus, for $\rho(B)$ to achieve the lower bound $\sqrt{\beta}$, we need $\left(1-\sqrt{\beta}\right)^2 \le \eta\lambda_i \le \left(1+\sqrt{\beta}\right)^2$ for all i, which holds when

$$\left(1-\sqrt{\beta}\right)^2 \le \eta\alpha \le \eta\ell \le \left(1+\sqrt{\beta}\right)^2.$$

Plugging in our choice of η, β, this is indeed true.

We now have $\rho(B) = \sqrt{\beta}$. Pick a norm $\|\cdot\|_\varepsilon$ such that $\|B\|_\varepsilon \le \sqrt{\beta}+\varepsilon$ using Lemma 4.1, whence

$$\|X_{t+1}\|_\varepsilon \le \|B^t\|_\varepsilon \|X_1\|_\varepsilon \le \left(\sqrt{\beta}+\varepsilon\right)^t \|X_1\|_\varepsilon.$$ □

Remark. Given $f(x) = \frac{1}{2}(x-x^*)^\top A(x-x^*)$ for positive definite, symmetric A, set $y = P(x-x^*)$ where $A = P^\top \Lambda P$ is the diagonalization of A. Minimizing f is now equivalent to minimizing $g(y) = y^\top \Lambda y$.

4.2. Nesterov's Accelerated Gradient Descent

Nesterov's accelerated gradient descent follows the iterative scheme

$$\begin{aligned} y_t &= x_t + \beta_t(x_t - x_{t-1}), \\ x_{t+1} &= y_t - \eta_t \nabla f(y_t). \end{aligned} \tag{N-$\mathcal{AGD}$}$$

Theorem 4.3. *Let f be α-strongly convex and ℓ-smooth, and let x^* be its global minimizer. Further let $\{x_t\}_{t\in\mathbb{N}}$ be iterates of (N-$\mathcal{AGD}$) with*

$$\eta = \frac{1}{\ell}, \quad \beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \quad \kappa = \frac{\ell}{\alpha}.$$

Then,

$$f(x_t) - f(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \left(\frac{l+m}{2}\right) \|x_0 - x^*\|^2$$

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

Theorem 4.4. *Let f be convex and ℓ-smooth, x^* be its global minimizer, and $\|x_0 - x^*\| \leq R$. Further let $x_1, \ldots, x_T$ be T iterates of (N-$\mathcal{AGD}$) with*

$$\eta = \frac{1}{\ell}, \quad \lambda_{t+1} = \frac{1 + \sqrt{1 + 4\lambda_t^2}}{2}, \quad \beta_{t+1} = \frac{\lambda_t - 1}{\lambda_{t+1}},$$

where $\lambda_0 = \beta_0 = 0$. Then,

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

5. Subdifferentials

Definition 5.1 (Subdifferential)**.** Let $f : \mathcal{K} \to \mathbb{R}$ 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 5.1.1. Consider $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|$. Then,

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

Example 5.1.2. Consider $f : \mathbb{R}^d \to \mathbb{R}$, $x \mapsto \|x\|_1 = \sum_{i=1}^d |x_i|$. Then,

$$\partial f(x) = \{v : v_i \in \partial|x_i| \text{ for all } 1 \le i \le d\}.$$

Example 5.1.3. Let $\mathcal{K}$ be closed and convex. The subdifferential of the indicator function $I_{\mathcal{K}}$ at $x \in \mathcal{K}$ is the normal cone $N_{\mathcal{K}}(x)$.

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

Proposition 5.2. *Let $f : \mathcal{K} \to \mathbb{R}$ be convex. Then, $\partial f(x)$ is nonempty for all $x \in \mathrm{ri}(\mathcal{K})$.*

Proof. Note that $\mathrm{epi}(f)$ is convex via Proposition 1.5. Use Proposition 1.15 to find a supporting hyperplane to $\mathrm{epi}(f)$ at $(x^\top\ f(x))^\top$, i.e. $(v^\top\ s)^\top \neq 0$ such that for all $(y^\top\ \alpha)^\top \in \mathrm{epi}(f)$,

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

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

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

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

The next result follows immediately from the definition of the subdifferential; compare this with Proposition 2.4.

Proposition 5.3. *Let $f : \mathcal{K} \to \mathbb{R}$ 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 $x \in \mathrm{int}(\mathcal{X})$, the subdifferential 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 $\lambda \ge 0$,

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

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

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

Theorem 5.4. *Let $f : \mathcal{K} \to \mathbb{R}$ be convex and $x \in \operatorname{int}(\mathcal{K})$. 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 $\nabla f(x) = v$.*

Proof. See [1, Theorem 25.1]. ☐

Example 5.4.1. Let f_1, f_2 be convex and differentiable, and let $f = \max\{f_1, f_2\}$. At points x where $f_1(x) \neq f_2(x)$, we have $\partial f(x) = \{\nabla f(x)\}$. Otherwise, $\partial f(x) = \operatorname{conv}(\nabla f_1(x), \nabla f_2(x))$.

Lemma 5.5.

1. $\partial(\alpha f) = \alpha\, (\partial f)$ *for* $\alpha > 0$.
2. $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.
3. $\partial g(x) = A^\top \partial f(Ax + b)$ *for* $g(x) = f(Ax + b)$.

Lemma 5.6. *Let* $f = \max\{f_1, \ldots, f_n\}$*. Then,*

$$\partial f(x) = \operatorname{conv}\left(\bigcup_{i : f(x) = f_i(x)} \partial f_i(x) \right)$$

5.1. Subgradient Descent

The subgradient descent algorithm follows the iterative scheme

$$x_{t+1} = x_t - \eta_t v_t, \quad v_t \in \partial f(x_t). \tag{$\mathcal{SGD}$}$$

Theorem 5.7. *Let f be convex and L-Lipschitz, x^* be its global minimizer, and $\|x_1 - x^*\| \leq R$. Further let $x_1, \ldots, x_T$ be T iterates of ($\mathcal{SGD}$). Then,*

$$\min_{1 \leq t \leq T} f(x_t) - f(x^*) \leq \frac{R^2 + L^2 \sum_{t=1}^T \eta_t^2}{2 \sum_{t=1}^T \eta_t}.$$

Proof. Write

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

This gives

$$\begin{aligned}
2\left(\sum_{t=1}^T \eta_t\right)\left(\min_{1 \leq t \leq T} f(x_t) - f(x^*)\right) &\leq \sum_{t=1}^T 2\eta_t (f(x_t) - f(x^*)) \\
&\leq \|x_1 - x^*\|^2 - \left\|x_{T+1} - x^*\right\|^2 + \sum_{t=1}^T \eta_t^2 L^2
\end{aligned}$$

$$\leq R^2 + L^2 \sum_{t=1}^{T} \eta_t^2. \qquad \square$$

Remark. We would like $\sum_t \eta_t \to \infty$ while $\sum_t \eta_t^2 < \infty$; this is achieved by step sizes of the form $\eta_t = 1/t$.

6. Exponential Gradient Descent

Consider the standard d-simplex

$$\Delta^d = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_i \geq 0 \text{ for all } 1 \leq i \leq d \right\}.$$

Members of Δ^d can be naturally identified with discrete probability distributions on $\{1, \ldots, k\}$. Given convex f, we examine the optimization problem

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

The exponential gradient descent algorithm follows the iterative scheme

$$\begin{aligned} z^{(t)} &= \sum_{i=1}^{d} x_i^{(t)} \, e^{-\eta \nabla f(x^{(t)})_i}, \\ x_i^{(t+1)} &= \frac{1}{z^{(t)}} \, x_i^{(t)} \, e^{-\eta \nabla f(x^{(t)})_i} \end{aligned} \tag{$\mathcal{EGD}$}$$

Since we are effectively dealing with probability distributions, we will use the Kullback-Leibler divergence instead of a Euclidean norm to describe convergence in Δ^d.

Definition 6.1 (Kullback-Leibler Divergence). The Kullback-Leibler (KL) divergence of $p \in \Delta^d$ with respect to $q \in \Delta^d$ is defined by

$$\mathrm{KL}(p \,\|\, q) = \mathbb{E}_p \left[\log \left(\frac{p}{q} \right) \right] = \sum_{i=1}^{d} p_i \log \left(\frac{p_i}{q_i} \right).$$

Note that for any $x^* \in \Delta^d$, the concavity of the logarithm gives

$$\mathrm{KL}\left(x^* \,\middle\|\, \frac{1}{d} \mathbf{1} \right) = \sum_{i=1}^{d} x_i^* \log(x_i^* \cdot d) \leq \log \left(\sum_{i=1}^{d} (x_i^*)^2 d \right) \leq \log(d).$$

This is often useful in bounding the ‘diameter’ of Δ^d.

Lemma 6.2. *For iterates of* ($\mathcal{EGD}$),

$$\mathrm{KL}(x^* \,\|\, x^{(t)}) - \mathrm{KL}(x^* \,\|\, x^{(t+1)}) = -\eta \nabla f(x^{(t)})^\top x^* - \log(z^{(t)}).$$

Theorem 6.3. *Let f be convex such that $\|\nabla f\|_\infty \leq \ell$ on Δ^d, and let $x^* \in \mathcal{K}$ be its global minimizer. Further let $x^{(1)}, ..., x^{(T)}$ be T iterates of* ($\mathcal{EGD}$) *with*

$$\eta = \frac{1}{\ell}\sqrt{\frac{\log(d)}{T}}, \quad x^{(1)} = \frac{1}{d}\mathbf{1}.$$

Then,

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x^{(t)}\right) - f(x^*) \leq 2\ell\sqrt{\frac{\log(d)}{T}}.$$

Proof. It will suffice to show that

$$\sum_{t=1}^{T} \nabla f\left(x^{(t)}\right)^\top \left(x^{(t)} - x^*\right) \leq \frac{\mathrm{KL}\left(x^* \,\|\, x^{(1)}\right)}{\eta} + \eta^2 \ell T,$$

from which the result will follow using Proposition 1.3 and Proposition 1.6, much like in the proof of Theorem 3.2. Indeed, checking that $e^{-x} \leq 1 + x + x^2$ for $x \leq 1$ and noting that $\eta\|\nabla f\|_\infty \leq 1$ for sufficiently large T,

$$\begin{aligned}
\log\left(z^{(t)}\right) &= \log\left(\sum_{i=1}^{d} x_i^{(t)} e^{-\eta\left(\nabla f\left(x^{(t)}\right)\right)_i}\right) \\
&\leq \log\left(\sum_{i=1}^{d} x_i^{(t)} \left(1 - \eta \nabla f\left(x^{(t)}\right)_i + \eta^2 \nabla f\left(x^{(t)}\right)_i^2\right)\right) \\
&= \log\left(1 - \eta \nabla f\left(x^{(t)}\right)^\top x^{(t)} + \sum_{i=1}^{d} \eta^2 \nabla f\left(x^{(t)}\right)_i^2 x_i^{(t)}\right) \\
&\leq \log\left(1 - \eta \nabla f\left(x^{(t)}\right)^\top x^{(t)} + \eta^2 \ell^2\right) \\
&\leq -\eta \nabla f\left(x^{(t)}\right)^\top x^{(t)} + \eta^2 \ell^2.
\end{aligned}$$

Thus, by Lemma 6.2,

$$\eta \nabla f\left(x^{(t)}\right)^\top \left(x^{(t)} - x^*\right) - \eta^2 \ell^2 \leq \mathrm{KL}\left(x^* \,\|\, x^{(t)}\right) - \mathrm{KL}\left(x^* \,\|\, x^{(t+1)}\right).$$

Summing over t and rearranging gives the desired result. □

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.