Convex Optimization

Brennan Gebotys

March 2020

Overview

Note: This presentation is based off of [1]

Introduction

Definition: Convex sets

A set $\mathcal{X} \subset \mathbb{R}^n$ is said to be convex if it contains all of its segments, that is

$$\forall (x, y, \gamma) \in \mathcal{X} \times \mathcal{X} \times [0,1], \ (1-\gamma)x + \gamma y \in \mathcal{X}. \tag{1.1}$$

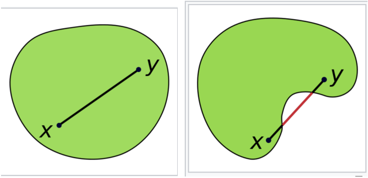


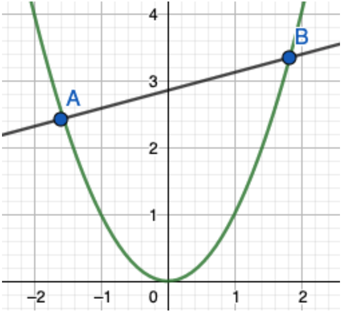
Figure: convex set vs non-convex set

Introduction

Definition: Convex functions

A function $f : \mathcal{X} \to \mathbb{R}$ is said to be convex if it always lies below its chords, that is

$$\forall (x, y, \gamma) \in \mathcal{X} \times \mathcal{X} \times [0, 1]$$
$$f((1-\gamma)x + \gamma y) \leq (1-\gamma)f(x) + \gamma f(y). \tag{1.2}$$



Objective

Objective

For a convex function f and convex set $\mathcal{X}$ find $x^* \in \mathcal{X}$ such that

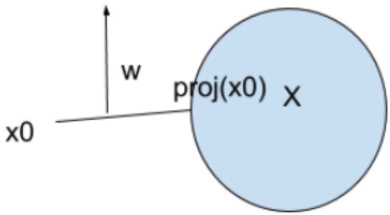
$$x^* = \operatorname{argmin} \; f(x) \tag{1.3}$$

Main Theorems

Theorem: Separation Theorem

Let $\mathcal{X} \subset \mathbb{R}^n$ be a closed convex set, and $x_0 \in \mathbb{R}^n \setminus \mathcal{X}$. Then, there exists $w \in \mathbb{R}^n$ and $t \in \mathbb{R}^n$ such that

$$w^\top x_0 < t, \text{ and } \forall x \in \mathcal{X}, w^\top x \geq t. \tag{2.1}$$

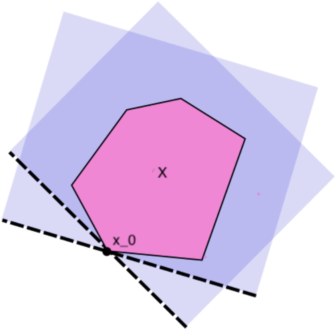


Main Theorems

Theorem: Supporting Hyperplane Theorem

Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set, and $x_0 \in \partial\mathcal{X}$. Then, there exists $w \in \mathbb{R}^n$, $w \neq 0$ such that

$$\forall x \in \mathcal{X}, w^\top x \geq w^\top x_0. \tag{2.2}$$



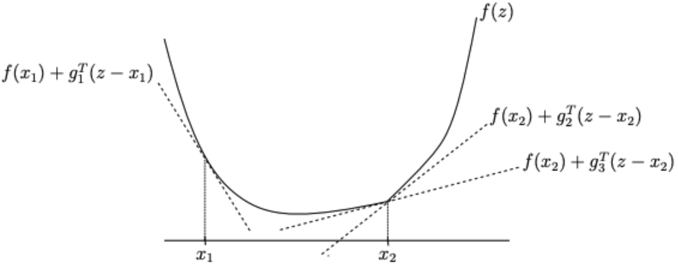
Main Theorems

Definition: Subgradients

Let $\mathcal{X} \subset \mathbb{R}^n$, and $f : \mathcal{X} \to \mathbb{R}$. Then $g \in \mathbb{R}^n$ is a subgradient of f at $x \in \mathcal{X}$ if for any $y \in \mathcal{X}$ one has

$$f(x) - f(y) \leq g^\top (x - y). \tag{2.3}$$

The set of subgradients of f at x is denoted $\partial f(x)$.



Convexity Properties

Definition: β-Smooth Convexity

We say that a continuously differentiable function f is β-smooth if the gradient ∇f is β-Lipschitz, that is

$$\frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} \leq \beta. \tag{2.4}$$

Convexity Properties

Definition: Strong Convexity

We say that $f : \mathcal{X} \to \mathbb{R}$ is α-strongly convex if it satisfies the following improved subgradient inequality:

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{\alpha}{2}\|x - y\|^2. \tag{2.5}$$

Another view:

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

$-\nabla f(x)^\top (x - y)$ must be strong enough to ensure its sum with $f(x) + \frac{\alpha}{2}\|x - y\|^2$ is $\leq f(y)$

Projected Gradient Descent

Assumptions:

$\mathcal{X}$ is contained in a Euclidean ball of radius R.

$\forall g \in \partial f(x), ||g|| \leq$ L (f is L-Lipschitz)

Define the projection operator of x on $\mathcal{X}$ as $\Pi_{\mathcal{X}}(x)$. In this case,

$$\Pi_{\mathcal{X}}(x) = \underset{y \in \mathcal{X}}{\operatorname{argmin}} ||x - y|| \tag{3.1}$$

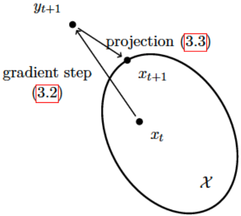
Algorithm

Definition: PGD Algorithm

For $t \geq 1$:

$$y_{t+1} = x_t - \eta g_t, \text{ where } g_t \in \partial f(x_t) \tag{3.2}$$

$$x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1}) \tag{3.3}$$



Convergence

Theorem: Convergence

The projected subgradient descent method with $\eta = \frac{R}{L\sqrt{t}}$ satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}. \tag{3.4}$$

To prove this we require more convex knowledge.

x^* properties

Lemma

Let $\mathcal{X}$ be a closed convex set and f be a convex function. Then x^* is a solution of (1.3) if and only if

$$\langle f'(x^*), x - x^* \rangle \geq 0 \tag{3.5}$$

Let x^* be a solution to (1.3). Assume $\exists x \in \mathcal{X}$ such that

$$\langle f'(x^*), x - x^* \rangle < 0$$

Consider $\phi(\alpha) = f(x^* + \alpha(x - x^*)), \alpha \in [0, 1]$.
Note:

$$\phi(0) = f(x^*), \phi'(0) = \langle f'(x^*), x - x^* \rangle < 0$$

Then for a small enough α,

$$f(x^* + \alpha(x - x^*)) = \phi(\alpha) < \phi(0) = f(x^*).\text{Contradiction.} \quad \square$$

Geometry Properties

Lemma

Let $\mathcal{X}$ be a closed convex set and $x_0 \notin \mathcal{X}$. Then for any $x \in \mathcal{X}$

$$\langle \Pi_{\mathcal{X}}(x_0) - x_0, x - \Pi_{\mathcal{X}}(x_0) \rangle \geq 0 \tag{3.6}$$

Note: For $h(x) = \frac{1}{2}||x - x_0||^2$, $\Pi_{\mathcal{X}}(x_0) \in \mathrm{argmin}_{x \in \mathcal{X}}\, h(x)$

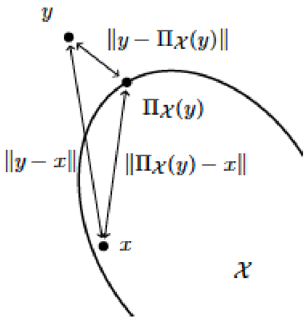
$$\langle h'(\Pi_{\mathcal{X}}(x_0)), x - \Pi_{\mathcal{X}}(x_0) \rangle \geq 0 \qquad (3.5)$$
$$\langle \Pi_{\mathcal{X}}(x_0) - x_0, x - \Pi_{\mathcal{X}}(x_0) \rangle \geq 0 \quad \square$$

Triangle Inequality-ish

Lemma

For any $x \in \mathcal{X}$ we have

$$||\Pi_{\mathcal{X}}(y) - x||^2 + ||y - \Pi_{\mathcal{X}}(y)||^2 \leq ||y - x||^2 \tag{3.7}$$



Triangle Inequality-ish

Lemma

For any $x \in \mathcal{X}$ we have

$$||\Pi_{\mathcal{X}}(y) - x||^2 + ||y - \Pi_{\mathcal{X}}(y)||^2 \leq ||y - x||^2$$

Proof:

$$\begin{aligned}
&||x - \Pi_{\mathcal{X}}(y)||^2 - ||x - y||^2 \\
&= \langle y - \Pi_{\mathcal{X}}(y), 2x - \Pi_{\mathcal{X}}(y) - y \rangle \\
&= \langle y - \Pi_{\mathcal{X}}(y), 2x - \Pi_{\mathcal{X}}(y) - y + (\Pi_{\mathcal{X}}(y) - \Pi_{\mathcal{X}}(y)) \rangle \\
&= \langle y - \Pi_{\mathcal{X}}(y), \Pi_{\mathcal{X}}(y) - y \rangle + 2 \langle y - \Pi_{\mathcal{X}}(y), x - \Pi_{\mathcal{X}}(y) \rangle \\
&= - \langle \Pi_{\mathcal{X}}(y) - y, \Pi_{\mathcal{X}}(y) - y \rangle - 2 \langle \Pi_{\mathcal{X}}(y) - y, x - \Pi_{\mathcal{X}}(y) \rangle \\
&\leq - \langle \Pi_{\mathcal{X}}(y) - y, \Pi_{\mathcal{X}}(y) - y \rangle \qquad (3.6) \qquad \square
\end{aligned}$$

PGD

Lemma

$$||x_{s+1} - x^*||^2 \leq ||y_{s+1} - x^*||^2 \tag{3.8}$$

Recall: PGD Algorithm

$$y_{t+1} = x_t - \eta g_t, \text{ where } g_t \in \partial f(x_t)$$
$$x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1})$$

Using (3.7) and $||y_{s+1} - \Pi_{\mathcal{X}}(y_{s+1})|| \geq 0$ we have,

$$||\Pi_{\mathcal{X}}(y_{s+1}) - x^*||^2 + ||y_{s+1} - \Pi_{\mathcal{X}}(y_{s+1})||^2 \leq ||y_{s+1} - x^*||^2$$
$$||\Pi_{\mathcal{X}}(y_{s+1}) - x^*||^2 \leq ||y_{s+1} - x^*||^2 - ||y_{s+1} - \Pi_{\mathcal{X}}(y_{s+1})||^2$$
$$||x_{s+1} - x^*||^2 \leq ||y_{s+1} - x^*||^2 \quad \square$$

PGD Convergence

Theorem: Convergence

The projected subgradient descent method with $\eta = \frac{R}{L\sqrt{t}}$ satisfies

$$f(\frac{1}{t}\sum_{s=1}^{t} x_s) - f(x^*) \leq \frac{RL}{\sqrt{t}}. \tag{3.9}$$

Proof: Using definition the of subgradients, (3.8) and the identity, $2a^\top b = ||a||^2 + ||b||^2 - ||a-b||^2$ we get,

$$\begin{aligned} f(x_s) - f(x^*) &\leq g_s^\top (x_s - x^*) \\ &= \frac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x^*) \qquad (3.2) \\ &= \frac{1}{2\eta}(||x_s - x^*||^2 - ||y_{s+1} - x^*||^2 + ||x_s - y_{s+1}||^2) \\ &\leq \frac{1}{2\eta}(||x_s - x^*||^2 - ||x_{s+1} - x^*||^2) + \frac{\eta}{2}||g_s||^2 (3.8) \end{aligned}$$

PGD Convergence

Theorem: Convergence

The projected subgradient descent method with $\eta = \frac{R}{L\sqrt{t}}$ satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}. \tag{3.10}$$

$$\begin{aligned} f(x_s) - f(x^*) &\leq \frac{1}{2\eta}\left(||x_s - x^*||^2 - ||x_{s+1} - x^*||^2\right) + \frac{\eta}{2}||g_s||^2 \\ \sum_{s=1}^{t} f(x_s) - f(x^*) &\leq \frac{1}{2\eta}||x_1 - x^*||^2 + \frac{tL^2\eta}{2} \\ &\leq \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} \end{aligned}$$

PGD Convergence

Theorem: Convergence

The projected subgradient descent method with $\eta = \frac{R}{L\sqrt{t}}$ satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}. \tag{3.11}$$

$$\sum_{s=1}^{t} f(x_s) - f(x^*) \leq \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2}$$

Using Jensen's Inequality, $f(\frac{1}{t}\sum_{s=1}^{t} x_s) \leq \frac{1}{t}\sum_{s=1}^{t} f(x_s)$

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}} \quad \square$$

When PGD Breaks

However, convergence of $\frac{RL}{\sqrt{t}}$ is only possible when f and $\mathcal{X}$ are well-behaved in Euclidean norm ($\|x\|_2$ and $\|g\|_2$ are independent of the ambient dimension n for $x \in \mathcal{X}$ and all $g \in \partial f(x)$.).

Example:

f on the Euclidean ball $B_{2,n}$

$\|\nabla f(x)\|_\infty \leq 1, \forall x \in B_{2,n}$. (f is 1-Lipschitz in $||\cdot||_\infty$)

Then, $||\nabla f(x)||_2 < \sqrt{n}$

PGD convergence = $\sqrt{n/t}$ (large n will be bad)

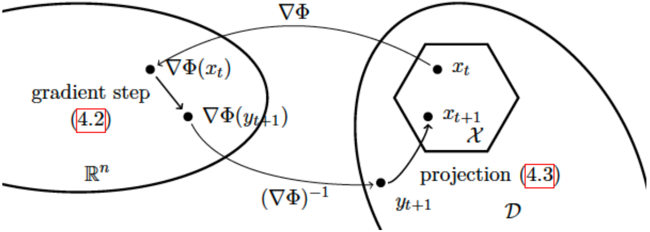
f has nice properties in $||\cdot||_\infty$ but is in a different vector space than x ($||\cdot||_2$).

Dual Space (∇f) vs Primal Space (x)

Can we find a better way? Yes!

Mirror Descent

Idea: Use an invertible mapping $\nabla\Phi$: Primal $\rightarrow$ Dual and optimize in the Dual



1. Map x_t to the dual, $\nabla\Phi(x_t)$
2. Take a gradient step, $\nabla\Phi(y_{t+1}) = \nabla\Phi(x_t) - \eta g$ ($g \in \partial f(x_t)$)
3. Map back to the primal $y_{t+1} = \nabla\Phi^{-1}\nabla\Phi(y_{t+1})$
4. Project into $\mathcal{X}$, $x_{t+1} = \Pi_{\mathcal{X}}^{\Phi}(y_{t+1})$

Mirror Descent: The Mapping

Definition: Mirror Maps

Let $\mathcal{D} \subset \mathbb{R}^n$ be a convex open set such that $\mathcal{X}$ is included in its closure, that is $\mathcal{X} \subset \overline{\mathcal{D}}$, and $\mathcal{X} \cap \mathcal{D} \neq \emptyset$. We say that $\Phi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it safisfies the following properties:

(i) Φ is strictly convex and differentiable.

(ii) The gradient of Φ takes all possible values, that is $\nabla \phi(\mathcal{D}) = \mathbb{R}^n$.

(iii) The gradient of Φ diverges on the boundary of $\mathcal{D}$, that is

$$\lim_{x \to \partial \mathcal{D}} \|\nabla \Phi(x)\| = +\infty.$$

Note: (ii) gives us invertibility! Why?

Hint: $\nabla \Phi(x_t) - \eta g = v \in \mathbb{R}^n \stackrel{?}{=}$

Mirror Descent: The Mapping

Definition: Mirror Maps

Let $\mathcal{D} \subset \mathbb{R}^n$ be a convex open set such that $\mathcal{X}$ is included in its closure, that is $\mathcal{X} \subset \overline{\mathcal{D}}$, and $\mathcal{X} \cap \mathcal{D} \neq \emptyset$. We say that $\Phi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it safisfies the following properties:

(i) Φ is strictly convex and differentiable.

(ii) The gradient of Φ takes all possible values, that is $\nabla\phi(\mathcal{D}) = \mathbb{R}^n$.

(iii) The gradient of Φ diverges on the boundary of $\mathcal{D}$, that is

$$\lim_{x \to \partial\mathcal{D}} \|\nabla\Phi(x)\| = +\infty.$$

Note: (ii) gives us invertibility! Why?
Hint: $\nabla\Phi(x_t) - \eta g = v \in \mathbb{R}^n = \nabla\Phi(y_{t+1})$ for some $y_{t+1} \in \mathcal{D}$ and (i) gives a 1-to-1 mapping.

Mirror Descent: The Projection

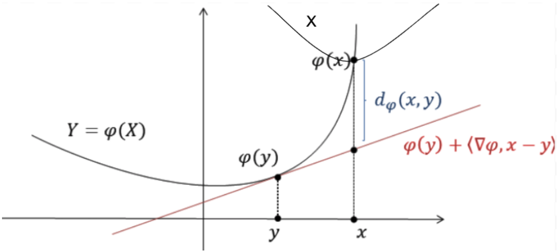
Definition: Bregman divergence

the Bregman divergence associated to f is

$$D_f(x,y) = f(x) - f(y) - \nabla f(y)^\top (x-y) \tag{4.1}$$

We then define the projection operation

$$\Pi^{\Phi}_{\mathcal{X}}(y) = \operatorname*{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} D_{\Phi}(x,y).$$



Bergman Divergence Example

Example:

Taking $\Phi(x) = \frac{1}{2}||x||_2^2$ on $\mathcal{D} = \mathbb{R}^n$

$$\begin{aligned}
D_\Phi(x, y) &= f(x) - f(y) - \langle \nabla f(y), x - y \rangle \\
&= \frac{1}{2}(||x||_2^2 - ||y||_2^2 - 2\langle y, x - y \rangle) \\
&= \frac{1}{2}(||x||_2^2 - ||y||_2^2 - 2\langle y, x \rangle + 2\langle y, y \rangle) \\
&= \frac{1}{2}(||x||_2^2 + ||y||_2^2 - 2\langle y, x \rangle) \\
&= \frac{1}{2}||x - y||_2^2
\end{aligned}$$

Bergman Divergence Example

Example:

Taking $\Phi(x) = \frac{1}{2}||x||_2^2$ on $\mathcal{D} = \mathbb{R}^n$

$$\begin{aligned}
D_\Phi(x, y) &= f(x) - f(y) - \langle \nabla f(y), x - y \rangle \\
&= \frac{1}{2}(||x||_2^2 - ||y||_2^2 - 2\langle y, x - y \rangle) \\
&= \frac{1}{2}(||x||_2^2 - ||y||_2^2 - 2\langle y, x \rangle + 2\langle y, y \rangle) \\
&= \frac{1}{2}(||x||_2^2 + ||y||_2^2 - 2\langle y, x \rangle) \\
&= \frac{1}{2}||x - y||_2^2
\end{aligned}$$

In this case Mirror Descent will be equivalent to PGD since $\nabla \Phi(x_t) = x_t$ and $\Pi_{\mathcal{X}}^{\Phi}(y) = \text{argmin}_{x \in \mathcal{X}} \frac{1}{2}||x - y||_2^2$

Mirror Descent

Definition: Mirror Descent Algorithm

Let $x_1 \in \operatorname{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} \Phi(x)$. Then for $t \geq 1$, let $y_{t+1} \in \mathcal{D}$ such that

$$\nabla \Phi(y_{t+1}) = \nabla \Phi(x_t) - \eta g_t, \text{ where } g_t \in \partial f(x_t), \tag{4.2}$$

and

$$x_{t+1} \in \Pi_{\mathcal{X}}^{\Phi}(y_{t+1}). \tag{4.3}$$

Theorem: Mirror Descent Convergence

Let Φ be a mirror map ρ-strongly convex on $\mathcal{X} \cap \mathcal{D}$ w.r.t. $\|\cdot\|$. Let $R^2 = \sup_{x \in \mathcal{X} \cap \mathcal{D}} \Phi(x) - \Phi(x_1)$, and f be convex and L-Lipschitz w.r.t. $\|\cdot\|$. Then mirror descent with $\eta = \frac{R}{L}\sqrt{\frac{2\rho}{t}}$ satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RL\sqrt{\frac{2}{\rho t}}. \tag{4.4}$$

But first, we need more lemmas!

Lemma

$$(\nabla f(x) - \nabla f(y))^\top (x - z) = D_f(x, y) + D_f(z, x) - D_f(z, y) \quad (4.5)$$

Proof:

$$\begin{aligned} &D_f(x, y) + D_f(z, x) - D_f(z, y) \\ &= (f(x) - f(y) - \nabla f(y)^\top (x - y)) + (f(z) - f(x) - \nabla f(z)^\top (z - x)) \\ &- (f(z) - f(y) - \nabla f(y)^\top (z - y)) \\ &= (\nabla f(x) - \nabla f(y))^T (x - z) \quad \square \end{aligned}$$

Lemma

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y) \tag{4.6}$$

Proof:

$$\begin{aligned}
\nabla_x D_f(x, y) &= \nabla_x (f(x) - f(y) - \langle \nabla f(y), x - y \rangle) \\
&= \nabla_x f(x) - \nabla_x \langle \nabla f(y), x - y \rangle) \\
&= \nabla f(x) - \nabla f(y) \quad \square
\end{aligned}$$

Lemma

For any $y \in \mathbb{R}^n$, let $\pi = \Pi_{\mathcal{X}}^{\Phi}(y)$ then

$$(\nabla f(y) - \nabla f(\pi))^\top (w - \pi) \leq 0 \qquad \forall w \in \mathcal{X} \tag{4.7}$$

Proof: Recall $\pi = \operatorname{argmin}_{x \in \mathcal{X}} D_f(x, y)$, then by optimality,

$$\nabla D_f(\pi, y)^\top (\pi - w) \leq 0 \qquad \forall w \in \mathcal{X} \qquad (3.5)$$

$$(\nabla f(\pi) - \nabla f(y))^\top (\pi - w) \leq 0 \qquad (4.6)$$

$$(\nabla f(y) - \nabla f(\pi))^\top (w - \pi) \leq 0 \quad \square$$

Lemma

For any $y \in \mathbb{R}^n$, let $\pi = \Pi_{\mathcal{X}}^{\Phi}(y)$ then

$$D_f(w, y) \geq D_f(w, \pi) \qquad \forall w \in \mathcal{X} \tag{4.8}$$

Proof:

$$D_f(\pi, y) + D_f(w, \pi) - D_f(w, y) = (\nabla f(\pi) - \nabla f(y))^\top (\pi - w) \leq 0 \text{(4.5)}$$

$$D_f(\pi, y) + D_f(w, \pi) \leq D_f(w, y)$$

$$D_f(w, \pi) \leq D_f(w, y), \text{using } D_f(\pi, y) \geq 0 \qquad \square$$

Convergence

Theorem: Mirror Descent Convergence

... mirror descent ... satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RL\sqrt{\frac{2}{\rho t}}. \tag{4.9}$$

Proof:

$$\begin{aligned} f(x_s) - f(x^*) &\leq g^\top (x_s - x^*) \\ &= \frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))^\top (x_s - x^*) && (4.2) \\ &\leq \frac{1}{\eta}(D_\Phi(x_s, y_{s+1}) + D_\Phi(x^*, x_s) - D_\Phi(x^*, y_{s+1})) && (4.5) \\ &\leq \frac{1}{\eta}(D_\Phi(x_s, y_{s+1}) + D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) && (4.8) \end{aligned}$$

Convergence

Then, summing over s

$$\sum_{s=1}^{t}(f(x_s) - f(x^*)) \leq \frac{1}{\eta}(D_\Phi(x^*, x_1) + \sum_{s=1}^{t} D_\Phi(x_s, y_{s+1}))$$

To derive the bounds on the last term:

$$\begin{aligned}
&D_\Phi(x_s, y_{s+1}) = \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla\Phi(y_{s+1}), x_s - y_{s+1}\rangle \\
&= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla\Phi(y_{s+1}) + (\nabla\Phi(x_s) - \nabla\Phi(x_s)), x_s - y_{s+1}\rangle \\
&= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla\Phi(x_s), x_s - y_{s+1}\rangle \\
&- \langle \nabla\Phi(y_{s+1}) - \nabla\Phi(x_s), x_s - y_{s+1}\rangle \\
&= \Phi(x_s) - \Phi(y_{s+1}) + \langle \nabla\Phi(x_s), y_{s+1} - x_s\rangle \\
&+ \langle \nabla\Phi(x_s) - \nabla\Phi(y_{s+1}), x_s - y_{s+1}\rangle \\
&\leq -\frac{\rho}{2}||x_s - y_{s+1}||^2 + \eta\,\langle g, x_s - y_{s+1}\rangle \qquad (\rho\text{-convexity (2.5) and (4.2)})
\end{aligned}$$

Convergence

We will also have to use some facts
Fact1 (Holder Inequality):
For $w \in V$ and $z \in V^*$

$$\langle z, w \rangle \leq ||w|| \cdot ||z||_* \tag{4.10}$$

where V^* is the dual of V.
Fact2:

$$az - bz^2 \leq \frac{a^2}{4b} \qquad \forall z \in \mathbb{R} \tag{4.11}$$

Then,

$$\begin{aligned} D_\Phi(x_s, y_{s+1}) &\leq -\frac{\rho}{2}||x_s - y_{s+1}||^2 + \eta \langle g, x_s - y_{s+1} \rangle \\ &\leq \eta||g||_*||x_s - y_{s+1}|| - \frac{\rho}{2}||x_s - y_{s+1}||^2 (4.10) \\ &\leq \frac{\eta^2||g||_*^2}{2\rho} \qquad (4.11) \end{aligned}$$

Convergence

Then we have,

$$\sum_{s=1}^{t}(f(x_s) - f(x^*)) \leq \frac{1}{\eta}(D_\Phi(x^*, x_1) + t\frac{\eta^2||g||_*^2}{2\rho})$$

$$f(\frac{1}{t}\sum_{i=1}^{t} x_s) - f(x^*) \leq \frac{R}{\eta t} + \frac{\eta L^2}{2\rho} = RL\sqrt{\frac{2}{\rho t}} \quad \square$$

Mirror Descent

To lead to the next topic we observe that we can rewrite mirror descent as follows,

$$\begin{aligned} x_{t+1} &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\operatorname{argmin}} D_\Phi(x, y_{t+1}) \\ &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\operatorname{argmin}} \Phi(x) - \Phi(y) - \nabla \Phi(y)^\top (x - y) \\ &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\operatorname{argmin}} \Phi(x) - \nabla \Phi(y)^\top x \\ &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\operatorname{argmin}} \Phi(x) - (\nabla \Phi(x_t) - \eta g_t)^\top x \\ &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\operatorname{argmin}} \eta g_t x + \Phi(x) - \nabla \Phi(x_t)^\top x \\ &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\operatorname{argmin}} \eta g_t x + D_\Phi(x, x_t) \end{aligned} \tag{4.12}$$

Stochastics

We now consider a stochastic oracle which takes as input $x \in \mathcal{X}$ and outputs a random variable $\widetilde{g(x)}$ such that $\mathbb{E}\widetilde{g(x)} \in \partial f(x)$

Assumptions:
Non-smooth case: there exists $B > 0$ such that $\mathbb{E}\|\widetilde{g}(x)\|_*^2 \leq B^2$ for all $x \in \mathcal{X}$.
Smooth case: there exists $\sigma > 0$ such that $\mathbb{E}\|\widetilde{g}(x) - \nabla f(x)\|_*^2 \leq \sigma^2$ for all $x \in \mathcal{X}$.

We are now interested in the minimization of

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

A more familiar view,

$$Loss(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(x_i, \theta)$$

Stochastic Mirror Descent

We'll look at convergence with the non-smooth assumption first.

Definition: Stochastic Mirror Descent Algorithm

Let $x_1 \in \operatorname{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} \Phi(x)$. Then for $t \geq 1$, let $y_{t+1} \in \mathcal{D}$ such that

$$\nabla \Phi(y_{t+1}) = \nabla \Phi(x_t) - \eta \widetilde{g}_t, \text{ where } \mathbb{E}(\widetilde{g}_t) \in \partial f(x_t), \tag{5.1}$$

and

$$x_{t+1} \in \Pi_{\mathcal{X}}^{\Phi}(y_{t+1}). \tag{5.2}$$

Stochastic Mirror Descent

Theorem: Stochastic Mirror Descent

Let Φ be a mirror map 1-strongly convex on $\mathcal{X} \cap \mathcal{D}$ with respect to $\|\cdot\|$, and let $R^2 = \sup_{x \in \mathcal{X} \cap \mathcal{D}} \Phi(x) - \Phi(x_1)$. Let f be convex. Furthermore assume that the stochastic oracle is such that $\mathbb{E}\|\widetilde{g(x)}\|_*^2 \leq B^2$. Then S-MD with $\eta = \frac{R}{B}\sqrt{\frac{2}{t}}$ satisfies

$$\mathbb{E} f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RB\sqrt{\frac{2}{t}}.$$

SMD

Recall the Mirror Descent proof...

Mirror Descent Proof

$$f(x_s) - f(x^*) \leq g^\top (x_s - x^*)$$

$$\dots$$

$$\sum_{s=1}^{t} (f(x_s) - f(x^*)) \leq \frac{1}{\eta} (D_\Phi(x^*, x_1) + \sum_{s=1}^{t} \frac{\eta^2 ||g||_*^2}{2\rho})$$

Corollary

$$\sum_{s=1}^{t} g_s^\top (x_s - x^*) \leq \frac{R^2}{\eta} + \frac{\eta}{2\rho} \sum_{s=1}^{t} \|g_s\|_*^2. \tag{5.3}$$

SMD Convergence Proof

Proof:

Using Jensen's Inequality, $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$ and (5.3) we have

$$\begin{aligned}
\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) &\leq \frac{1}{t}\mathbb{E}\sum_{s=1}^{t}(f(x_s) - f(x^*)) \\
&\leq \frac{1}{t}\mathbb{E}\sum_{s=1}^{t}\mathbb{E}(\tilde{g}(x_s)|x_s)^\top(x_s - x^*) \\
&= \frac{1}{t}\mathbb{E}\sum_{s=1}^{t}\tilde{g}(x_s)^\top(x_s - x^*). \\
&\leq \frac{1}{t}\mathbb{E}\left(\frac{R^2}{\eta} + \frac{\eta}{2\rho}\sum_{s=1}^{t}\|g_s\|_*^2\right) && (5.3) \\
&\leq RB\sqrt{\frac{2}{t}} \quad \square
\end{aligned}$$

Stochastic PGD Convergence

Theorem: PGD Convergence

Let f be α-strongly convex, and assume that the stochastic oracle is such that $\mathbb{E}\|\widetilde{g}(x)\|_*^2 \leq B^2$. Then PGD with $\eta_s = \frac{2}{\alpha(s+1)}$ satisfies

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) \leq \frac{2B^2}{\alpha(t+1)}.$$

The proof follows similar to our first PGD proof:

$$\begin{aligned}
f(x_s) - f(x^*) &\leq \mathbb{E}\widetilde{g}^\top (x_s - x^*) - \frac{\alpha}{2}||x_s - x^*||^2 \\
&\dots \\
&\leq \frac{1}{2\eta_s}(||x_s - x^*||^2 - ||x_{s+1} - x^*||^2) + \frac{\eta_s}{2}\mathbb{E}||\widetilde{g}(x_s)||_*^2 \\
&\quad -\frac{\alpha}{2}||x_s - x^*||^2
\end{aligned}$$

PGD Convergence

$$\begin{aligned} \leq \quad & \frac{1}{2\eta_s}(||x_s - x^*||^2 - ||x_{s+1} - x^*||^2) + \frac{\eta_s}{2}B^2 \\ & -\frac{\alpha}{2}||x_s - x^*||^2 \\ = \quad & (\frac{1}{2\eta_s} - \frac{\alpha}{2})||x_s - x^*||^2 - \frac{1}{2\eta_s}||x_{s+1} - x^*||^2 + \frac{\eta_s}{2}B^2 \end{aligned}$$

Setting $\eta_s = \frac{2}{\alpha(s+1)}$ and multiplying both sides by s leads to

$$s(f(x_s)-f(x^*)) \leq \frac{B^2}{\alpha}+\frac{\alpha}{4}\left(s(s-1)\|x_s-x^*\|^2-s(s+1)\|x_{s+1}-x^*\|^2\right)$$

Note: Expanding $\sum_{s=1}^{t} s((s-1)x_s - (s+1)x_{s+1})$ we get

$$(0x_1 - 2x_2) + 2(x_2 - 3x_3) + 3(2x_3 - 4x_4) ... + t((t-1)x_t - (t+1)x_{t+1})$$

$$= -t(t+1)x_{t+1}$$

PGD Convergence

$$s(f(x_s)-f(x^*)) \leq \frac{B^2}{\alpha}+\frac{\alpha}{4}\left(s(s-1)\|x_s-x^*\|^2-s(s+1)\|x_{s+1}-x^*\|^2\right)$$

Then taking the sum from $s = 1, ..., t$

$$\begin{aligned}
\sum_{s=1}^{t} s(f(x_s) - f(x^*)) &\leq \frac{tB^2}{\alpha} - \frac{\alpha}{4} t(t+1)||x_{t+1} - x^*||^2 \\
\sum_{s=1}^{t} \frac{2s}{t(t+1)}(f(x_s) - f(x^*)) &\leq \frac{2B^2}{(t+1)\alpha} - \frac{\alpha}{2}||x_{t+1} - x^*||^2 \\
&\leq \frac{2B^2}{\alpha(t+1)} \\
f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) &\leq \frac{2B^2}{\alpha(t+1)} \quad \square
\end{aligned}$$

Non-Smooth Convergence

Theorem: Stochastic PGD

Let f be α-strongly convex, and assume that the stochastic oracle is such that $\mathbb{E}\|\widetilde{g}(x)\|_*^2 \leq B^2$. Then PGD with $\eta_s = \frac{2}{\alpha(s+1)}$ satisfies

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) \leq \frac{2B^2}{\alpha(t+1)}.$$

Following the previous proof's structure its easy to show,

Theorem: PGD Convergence

f be α-strongly convex and L-Lipschitz on $\mathcal{X}$. Then projected subgradient descent with $\eta_s = \frac{2}{\alpha(s+1)}$ satisfies

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

Non-Smooth Convergence

Similarly, comparing our results derived for Mirror Descent

Theorem: Stochastic Mirror Descent

...

$$\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RB\sqrt{\frac{2}{t}}.$$

Theorem: Mirror Descent Convergence

...

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RL\sqrt{\frac{2}{t}}.$$

There is basically no cost for having a stochastic oracle compared to an exact oracle!

Smooth S-MD Convergence

Now we investigate the convergence with the smooth assumption ($\mathbb{E}\|\widetilde{g}(x) - \nabla f(x)\|_*^2 \leq \sigma^2$).

Theorem: Smooth S-MD

Let Φ be a mirror map 1-strongly convex on $\mathcal{X} \cap \mathcal{D}$ w.r.t. $\|\cdot\|$, and let $R^2 = \sup_{x \in \mathcal{X} \cap \mathcal{D}} \Phi(x) - \Phi(x_1)$. Let f be convex and β-smooth w.r.t. $\|\cdot\|$. Furthermore assume that the stochastic oracle is such that $\mathbb{E}\|\nabla f(x) - \widetilde{g}(x)\|_*^2 \leq \sigma^2$. Then S-MD with stepsize $\frac{1}{\beta + 1/\eta}$ and $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{t}}$ satisfies

$$\mathbb{E} f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*) \leq R\sigma\sqrt{\frac{2}{t}} + \frac{\beta R^2}{t}.$$

Unfortunately, smoothness doesn't improve the general stochastic oracle :(but it's result can be useful

Smooth S-MD Convergence

One quick lemma!

Lemma

$$\eta D_\Phi(x_{s+1}, x_s) \leq 1/\eta(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) - \widetilde{g}_s^\top (x_{s+1} - x^*) \tag{5.4}$$

Proof:

$$\begin{aligned}
&\eta \widetilde{g}(x_s)^\top (x_{s+1} - x^*) \\
&= (\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))(x_{s+1} - x^*) \\
&\leq (\nabla\Phi(x_s) - \nabla\Phi(x_{s+1}))(x_{s+1} - x^*) \\
&= D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}) - D_\Phi(x_{s+1}, x_s)
\end{aligned}$$

$$\eta \widetilde{g}_s^\top (x_{s+1} - x^*) \leq D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}) - D_\Phi(x_{s+1}, x_s)$$

$$\widetilde{g}_s^\top (x_{s+1} - x^*) \leq 1/\eta(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}) - D_\Phi(x_{s+1}, x_s))$$

$$\eta D_\Phi(x_{s+1}, x_s) \leq 1/\eta(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) - \widetilde{g}_s^\top (x_{s+1} - x^*) \qquad \square$$

Smooth S-MD Convergence

Theorem: Smooth S-MD

...

$$\mathbb{E} f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*) \leq R\sigma\sqrt{\frac{2}{t}} + \frac{\beta R^2}{t}.$$

$$\begin{aligned} & f(x_{s+1}) - f(x_s) \\ & \leq \widetilde{g}_s^\top (x_{s+1} - x_s) \\ & \leq \widetilde{g}_s^\top (x_{s+1} - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \widetilde{g}_s\|_*^2 + (\beta + 1/\eta)\frac{1}{2}\|x_{s+1} - x_s\|^2 \\ & \leq \widetilde{g}_s^\top (x_{s+1} - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \widetilde{g}_s\|_*^2 + (\beta + 1/\eta) D_\Phi(x_{s+1}, x_s). \end{aligned}$$

Smooth S-MD Convergence

$$
\begin{aligned}
& f(x_{s+1}) - f(x_s) \\
& \leq \tilde{g}_s^\top (x_{s+1} - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 + (\beta + 1/\eta) D_\Phi(x_{s+1}, x_s) \\
& \leq \tilde{g}_s^\top (x_{s+1} - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 \\
& \qquad + 1/(\beta + 1/\eta)(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) \\
& \qquad\quad - \tilde{g}_s^\top (x_{s+1} - x^*) \qquad (5.4) \\
& \leq \tilde{g}_s^\top (x^* - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 \\
& \qquad + 1/(\beta + 1/\eta)(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}))
\end{aligned}
$$

$$
\begin{aligned}
f(x_{s+1}) \leq f(x_s) + \tilde{g}_s^\top (x^* - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 \\
+ 1/(\beta + 1/\eta)(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}))
\end{aligned}
$$

Smooth S-MD Convergence

Note: $f(x_s) \leq f(x^*) + \widetilde{g}^\top (x_s - x^*) = f(x^*) - \widetilde{g}^\top (x^* - x_s)$

$$\begin{aligned}
f(x_{s+1}) &\leq f(x_s) + \widetilde{g}_s^\top (x^* - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \widetilde{g}_s\|_*^2 \\
&\quad + 1/(\beta + 1/\eta)(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) \\
&\leq f(x^*) - \widetilde{g}^\top (x^* - x_s) + \widetilde{g}_s^\top (x^* - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \widetilde{g}_s\|_*^2 \\
&\quad + 1/(\beta + 1/\eta)(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) \\
&\leq f(x^*) + \frac{\eta}{2}\|\nabla f(x_s) - \widetilde{g}_s\|_*^2 \\
&\quad + 1/(\beta + 1/\eta)(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}))
\end{aligned}$$

Smooth S-MD Convergence

$$\mathbb{E}f(x_{s+1}) - f(x^*) \leq \frac{\sigma^2 \eta}{2} + 1/(\beta + 1/\eta)\mathbb{E}(D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1}))$$

$$\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}f(x_{s+1}) - f(x^*) \leq \frac{\sigma^2 \eta}{2} + \frac{R^2}{t(\beta + 1/\eta)}$$

Using Jensen's Inequality and $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{t}}$

$$\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*) \leq R\sigma\sqrt{\frac{2}{t}} + \frac{\beta R^2}{t} \qquad \square$$

Mini-Batches

Let $m \in \mathbb{N}$ and $\widetilde{g}_i(x_t), i = 1, \ldots, m$ be independent random variables obtained from the stochastic oracle then mini-batch SGD iterates the following equation:

$$x_{t+1} = \Pi_{\mathcal{X}}\left(x_t - \frac{\eta}{m}\sum_{i=1}^{m}\widetilde{g}_i(x_t)\right).$$

With a few assumptions:

- f is β-smooth
- $\|\widetilde{g}(x)\|_2 \leq B$

Convergence?
Using the previous theorem we can prove mini-batch SGD has a convergence of

$$\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^{t}x_{s+1}\right) - f(x^*) \leq 2\frac{RB}{\sqrt{t}} + \frac{m\beta R^2}{t}.$$

Mini-Batches

Using the property of independence we get the following,

$$
\begin{aligned}
&\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\widetilde{g}_i(x) - \nabla f(x)\|_2^2 \\
&= \frac{1}{m^2}\sum_{j=1}^{m}\sum_{i=1}^{m}\mathbb{E}\left\langle \widetilde{g}_i(x) - \nabla f(x), \widetilde{g}_j(x) - \nabla f(x)\right\rangle \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\|\widetilde{g}_i(x) - \nabla f(x)\|_2^2 \\
&\leq \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\|\widetilde{g}_i(x)\|_2^2 \\
&\leq \frac{mB^2}{m^2} \\
&\leq \frac{B^2}{m}
\end{aligned}
$$

Mini-Batches

$$\mathbb{E}\|\frac{1}{m}\sum_{i=1}^{m}\tilde{g}_i(x)-\nabla f(x)\|_2^2 \leq \frac{B^2}{m}$$

Then we can apply the previous theorem to get

$$R\sqrt{\frac{2B^2}{m}}\sqrt{\frac{2}{t/m}}+\frac{\beta R^2}{t/m}=2\frac{RB}{\sqrt{t}}+\frac{m\beta R^2}{t} \qquad \square$$

When would you want to use Minibatch-SGD?

> When computation can be distributed between multiple processors

The End

Thanks for joining!
Questions?

References:

[1] Sébastien Bubeck. Convex optimization: Algorithms and complexity, 2014.

[2] Nicholas Harvey. Machine learning theory, lecture 20: Mirror descent, 2018.

[3] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic, 2004.

[4] Ioannis Mitliagkas. Gradients for smooth and for strongly convex functions, 2019.

[5] S. Boyd and L. Vandenberghe. Subgradients, 2018.