

# Convex Optimization

Brennan Gebotys

March 2020

# Overview

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

- 1 Introduction to Convexity
- 2 Main Theorems and Definitions
- 3 PGD
  - Lemmas
  - Convergence
- 4 Mirror Descent
  - Lemmas
  - Convergence
- 5 Stochastics
  - Non-Smooth Case
    - S-MD
    - S-PGD
  - Smooth Case
    - S-MD
  - Mini-Batch SGD

Note: This presentation is based off of [1]

# Introduction

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

# Introduction

## Convex Optimization

Brennan Gebotys

### Introduction to Convexity

#### Main Theorems and Definitions

#### PGD

#### Lemmas

#### Convergence

#### Mirror Descent

#### Lemmas

#### Convergence

#### Stochastics

#### Non-Smooth Case

#### S-MD

#### S-PGD

#### Smooth Case

#### S-MD

#### Mini-Batch SGD

#### References

## 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

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

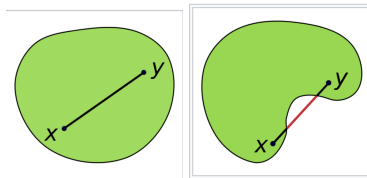


Figure: convex set vs non-convex set

# Introduction

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

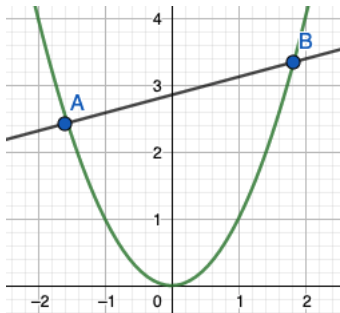
Mini-Batch SGD

References

## Definition: Convex functions

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

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



# Objective

## Convex Optimization

Brennan Gebotys

### Introduction to Convexity

### Main Theorems and Definitions

### PGD

#### Lemmas

#### Convergence

### Mirror Descent

#### Lemmas

#### Convergence

### Stochastics

#### Non-Smooth Case

##### S-MD

##### S-PGD

#### Smooth Case

##### S-MD

#### Mini-Batch SGD

### References

## Objective

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

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

# Main Theorems

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

# Main Theorems

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

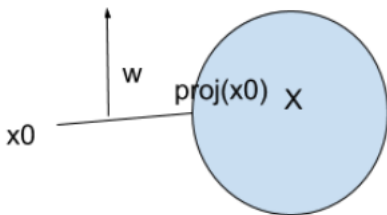
Mini-Batch SGD

References

## 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

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





# Main Theorems

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

# Main Theorems

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

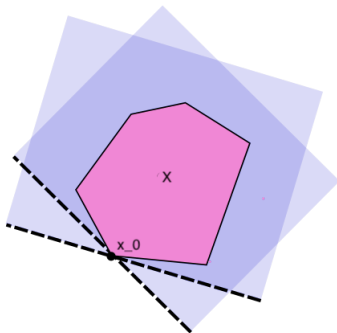
Mini-Batch SGD

References

## 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

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



# Main Theorems

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Definition: Subgradients

Let  $\mathcal{X} \subset \mathbb{R}^n$ , and  $f : \mathcal{X} \rightarrow \mathbb{R}$ . 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

$$f(x) - f(y) \leq g^\top (x - y). \quad (2.3)$$

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

# Main Theorems

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

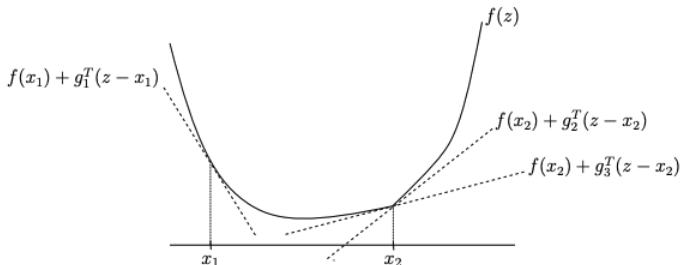
References

## Definition: Subgradients

Let  $\mathcal{X} \subset \mathbb{R}^n$ , and  $f : \mathcal{X} \rightarrow \mathbb{R}$ . 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

$$f(x) - f(y) \leq g^\top (x - y). \quad (2.3)$$

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



# Convexity Properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Definition: $\beta$ -Smooth Convexity

We say that a continuously differentiable function  $f$  is  $\beta$ -smooth if the gradient  $\nabla f$  is  $\beta$ -Lipschitz, that is

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

# Convexity Properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Definition: Strong Convexity

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

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

# Convexity Properties

## Convex Optimization

Brennan Gebotys

Introduction to Convexity

Main Theorems and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Definition: Strong Convexity

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

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

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

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

**PGD**

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Assumptions:

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



# Projected Gradient Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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)

# Projected Gradient Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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,

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

# Algorithm

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

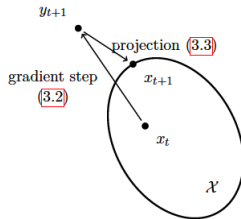
References

## Definition: PGD Algorithm

For  $t \geq 1$  :

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

$$(3.3)$$



# Algorithm

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

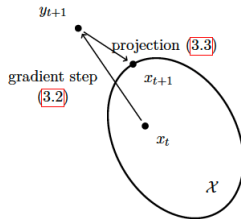
References

## Definition: PGD Algorithm

For  $t \geq 1$  :

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

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



# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

To prove this we require more convex knowledge.

# $x^*$ properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

# $x^*$ properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

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

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



# $x^*$ properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

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:

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

# $x^*$ properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

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  $\alpha$ ,

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

# Geometry Properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

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

# Geometry Properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

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

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

# Geometry Properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

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

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

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

# Geometry Properties

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

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

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

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

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

# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

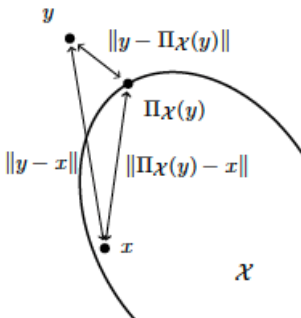
Mini-Batch SGD

References

## Lemma

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

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



# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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:



# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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:

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

# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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:

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

# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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:

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

# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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:

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

# Triangle Inequality-ish

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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:

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

# PGD

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

# PGD

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

## 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,

# PGD

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

**Lemmas**

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Lemma

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

## 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,

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



## Lemma

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

## 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,

$$\begin{aligned} \|\Pi_{\mathcal{X}}(y_{s+1}) - x^*\|^2 + \|y_{s+1} - \Pi_{\mathcal{X}}(y_{s+1})\|^2 &\leq \|y_{s+1} - x^*\|^2 \\ \|\Pi_{\mathcal{X}}(y_{s+1}) - x^*\|^2 &\leq \|y_{s+1} - x^*\|^2 - \|y_{s+1} - \Pi_{\mathcal{X}}(y_{s+1})\|^2 \end{aligned}$$

## Lemma

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

## 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,

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

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

# PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to Convexity

Main Theorems and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

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

# PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

$$f(x_s) - f(x^*) \leq \frac{1}{2\eta} (\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta}{2} \|g_s\|^2$$



# PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

$$f(x_s) - f(x^*) \leq \frac{1}{2\eta} (\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta}{2} \|g_s\|^2$$

$$\sum_{s=1}^t f(x_s) - f(x^*) \leq \frac{1}{2\eta} \|x_1 - x^*\|^2 + \frac{tL^2\eta}{2}$$

# PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Convergence

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

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

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

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

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

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $g \in \partial f(x)$  .).

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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



# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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

$$\text{PGD convergence} = \sqrt{n/t}$$

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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

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

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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

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  $x$  ( $\|\cdot\|_2$ ).

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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

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 ( $\nabla f$ ) vs Primal Space ( $x$ )

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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

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 ( $\nabla f$ ) vs Primal Space ( $x$ )

Can we find a better way?

# When PGD Breaks

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $B_{2,n}$

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

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

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 ( $\nabla f$ ) vs Primal Space ( $x$ )

Can we find a better way? Yes!

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

**Mirror Descent**

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

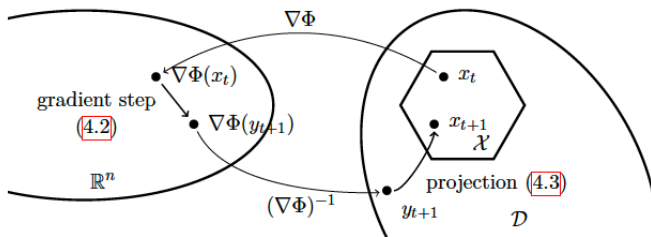
Mini-Batch SGD

References

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

# Mirror Descent

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

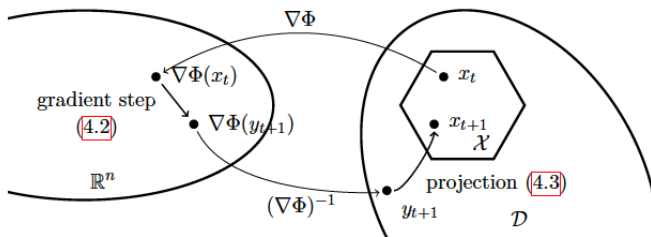


1. Map  $x_t$  to the dual,  $\nabla\Phi(x_t)$



# Mirror Descent

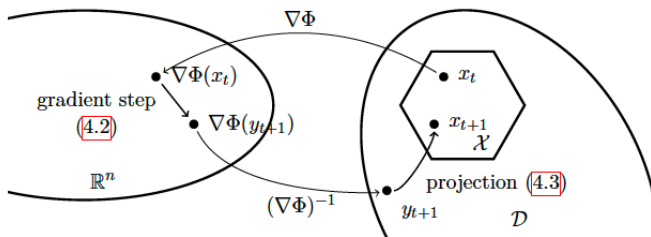
Idea: Use an invertible mapping  $\nabla\Phi$ : Primal  $\rightarrow$  Dual and optimize in the Dual



1. Map  $x_t$  to the dual,  $\nabla\Phi(x_t)$
2. Take a gradient step,  $\nabla\Phi(y_{t+1}) = \nabla\Phi(x_t) - \eta g$  ( $g \in \partial f(x_t)$ )

# Mirror Descent

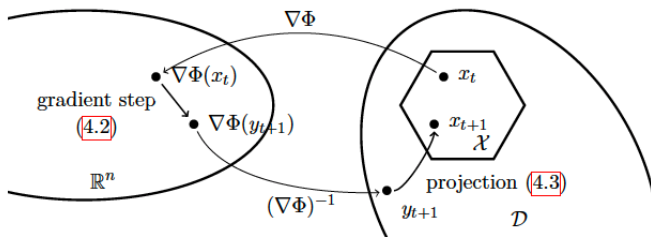
Idea: Use an invertible mapping  $\nabla\Phi$ : Primal  $\rightarrow$  Dual and optimize in the Dual



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  $y_{t+1} = \nabla\Phi^{-1} \nabla\Phi(y_{t+1})$

# 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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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  $\Phi : \mathcal{D} \rightarrow \mathbb{R}$  is a mirror map if it satisfies the following properties:

- (i)  $\Phi$  is strictly convex and differentiable.
- (ii) The gradient of  $\Phi$  takes all possible values, that is  $\nabla\phi(\mathcal{D}) = \mathbb{R}^n$ .
- (iii) The gradient of  $\Phi$  diverges on the boundary of  $\mathcal{D}$ , that is

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

# Mirror Descent: The Mapping

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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  $\Phi : \mathcal{D} \rightarrow \mathbb{R}$  is a mirror map if it satisfies the following properties:

- (i)  $\Phi$  is strictly convex and differentiable.
- (ii) The gradient of  $\Phi$  takes all possible values, that is  $\nabla\phi(\mathcal{D}) = \mathbb{R}^n$ .
- (iii) The gradient of  $\Phi$  diverges on the boundary of  $\mathcal{D}$ , that is

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

Note: (ii) gives us invertibility! Why?

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

# Mirror Descent: The Mapping

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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  $\Phi : \mathcal{D} \rightarrow \mathbb{R}$  is a mirror map if it satisfies the following properties:

- (i)  $\Phi$  is strictly convex and differentiable.
- (ii) The gradient of  $\Phi$  takes all possible values, that is  $\nabla\phi(\mathcal{D}) = \mathbb{R}^n$ .
- (iii) The gradient of  $\Phi$  diverges on the boundary of  $\mathcal{D}$ , that is

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

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

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

**Mirror Descent**

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Definition: Bregman divergence

the Bregman divergence associated to  $f$  is

$$D_f(x, y) = f(x) - f(y) - \nabla f(y)^\top (x - y) \quad (4.1)$$

# Mirror Descent: The Projection

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

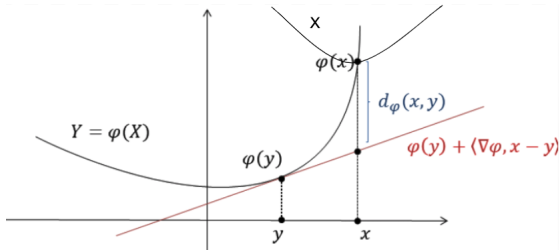
## Definition: Bregman divergence

the Bregman divergence associated to  $f$  is

$$D_f(x, y) = f(x) - f(y) - \nabla f(y)^\top (x - y) \quad (4.1)$$

We then define the projection operation

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





# Bergman Divergence Example

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

**Mirror Descent**

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Example:

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

# Bergman Divergence Example

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Example:

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

$$D_{\Phi}(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

# Bergman Divergence Example

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Example:

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

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

# Bergman Divergence Example

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case  
S-MD

S-PGD

Smooth Case  
S-MD

Mini-Batch SGD

References

Example:

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

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

# Bergman Divergence Example

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Example:

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

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

# Bergman Divergence Example

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

$$\nabla \Phi(x_t) =$$

# Bergman Divergence Example

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

$$\nabla \Phi(x_t) = x_t \text{ and } \Pi_{\mathcal{X}}^{\Phi}(y) =$$



# Bergman Divergence Example

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

$$\nabla \Phi(x_t) = x_t \text{ and } \Pi_{\mathcal{X}}^\Phi(y) = \operatorname{argmin}_{x \in \mathcal{X}} \frac{1}{2}\|x - y\|_2^2$$

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

**Mirror Descent**

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

and

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

# Mirror Descent

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## 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

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

and

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

## Theorem: Mirror Descent Convergence

Let  $\Phi$  be a mirror map  $\rho$ -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

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

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

**Lemmas**

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

But first,

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

**Lemmas**

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

But first, we need more lemmas!

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

**Lemmas**  
Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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:

# Mirror Descent

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case  
S-MD

S-PGD

Smooth Case  
S-MD

Mini-Batch SGD

References

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:

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

# Mirror Descent

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent

Lemmas  
Convergence

Stochastics

Non-Smooth Case  
S-MD

S-PGD

Smooth Case  
S-MD

Mini-Batch SGD

References

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:

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



## Lemma

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y) \quad (4.6)$$

Proof:

## Lemma

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y) \quad (4.6)$$

Proof:

$$\nabla_x D_f(x, y) = \nabla_x (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

## Lemma

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y) \quad (4.6)$$

Proof:

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

## Lemma

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y) \quad (4.6)$$

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

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

## Lemma

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

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

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

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

## Lemma

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

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

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

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

$$(\nabla f(\pi) - \nabla f(y))^{\top} (\pi - w) \leq 0 \quad (4.6)$$

## Lemma

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

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

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

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

$$(\nabla f(\pi) - \nabla f(y))^{\top}(\pi - \mathbf{w}) \leq 0 \quad (4.6)$$

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



## Lemma

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

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

## Lemma

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

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

Proof:

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

## Lemma

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

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

Proof:

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

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

## Lemma

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

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

Proof:

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

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

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

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Mirror Descent Convergence

... mirror descent ... satisfies

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

Proof:

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

# Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Mirror Descent Convergence

... mirror descent ... satisfies

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

Proof:

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

# Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Mirror Descent Convergence

... mirror descent ... satisfies

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

Proof:

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

$$\leq \frac{1}{\eta} (D_\Phi(x_s, y_{s+1}) + D_\Phi(x^*, x_s) - D_\Phi(x^*, y_{s+1})) \quad (4.5)$$

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Mirror Descent Convergence

... mirror descent ... satisfies

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

Proof:

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

$$\leq \frac{1}{\eta} (D_\Phi(x_s, y_{s+1}) + D_\Phi(x^*, x_s) - D_\Phi(x^*, y_{s+1})) \quad (4.5)$$

$$\leq \frac{1}{\eta} (D_\Phi(x_s, y_{s+1}) + D_\Phi(x^*, x_s) - D_\Phi(x^*, x_{s+1})) \quad (4.8)$$



# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Then, summing over  $s$

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

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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:

$$D_{\Phi}(x_s, y_{s+1}) = \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}), x_s - y_{s+1} \rangle$$

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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:

$$\begin{aligned} D_{\Phi}(x_s, y_{s+1}) &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}), x_s - y_{s+1} \rangle \\ &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}) + (\nabla \Phi(x_s) - \nabla \Phi(y_{s+1})), x_s - y_{s+1} \rangle \end{aligned}$$

# Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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:

$$\begin{aligned} D_{\Phi}(x_s, y_{s+1}) &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}), x_s - y_{s+1} \rangle \\ &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}) + (\nabla \Phi(x_s) - \nabla \Phi(y_{s+1})), x_s - y_{s+1} \rangle \\ &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(x_s), x_s - y_{s+1} \rangle \\ &\quad - \langle \nabla \Phi(y_{s+1}) - \nabla \Phi(x_s), x_s - y_{s+1} \rangle \end{aligned}$$

# 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:

$$\begin{aligned} D_{\Phi}(x_s, y_{s+1}) &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}), x_s - y_{s+1} \rangle \\ &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(y_{s+1}) + (\nabla \Phi(x_s) - \nabla \Phi(y_{s+1})), x_s - y_{s+1} \rangle \\ &= \Phi(x_s) - \Phi(y_{s+1}) - \langle \nabla \Phi(x_s), x_s - y_{s+1} \rangle \\ &\quad - \langle \nabla \Phi(y_{s+1}) - \nabla \Phi(x_s), x_s - y_{s+1} \rangle \\ &= \Phi(x_s) - \Phi(y_{s+1}) + \langle \nabla \Phi(x_s), y_{s+1} - x_s \rangle \\ &\quad + \langle \nabla \Phi(x_s) - \nabla \Phi(y_{s+1}), x_s - y_{s+1} \rangle \end{aligned}$$

# 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:

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

# Convergence

We will also have to use some facts

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

**Convergence**

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

We will also have to use some facts

Fact1 (Holder Inequality):

For  $w \in V$  and  $z \in V^*$

$$\langle z, w \rangle \leq \|w\| \cdot \|z\|_* \quad (4.10)$$

where  $V^*$  is the dual of  $V$ .



# Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

We will also have to use some facts

Fact1 (Holder Inequality):

For  $w \in V$  and  $z \in V^*$

$$\langle z, w \rangle \leq \|w\| \cdot \|z\|_* \quad (4.10)$$

where  $V^*$  is the dual of  $V$ .

Fact2:

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

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

We will also have to use some facts

Fact1 (Holder Inequality):

For  $w \in V$  and  $z \in V^*$

$$\langle z, w \rangle \leq \|w\| \cdot \|z\|_* \quad (4.10)$$

where  $V^*$  is the dual of  $V$ .

Fact2:

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

Then,

$$D_\Phi(x_s, y_{s+1}) \leq -\frac{\rho}{2} \|x_s - y_{s+1}\|^2 + \eta \langle g, x_s - y_{s+1} \rangle$$

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

We will also have to use some facts

Fact1 (Holder Inequality):

For  $w \in V$  and  $z \in V^*$

$$\langle z, w \rangle \leq \|w\| \cdot \|z\|_* \quad (4.10)$$

where  $V^*$  is the dual of  $V$ .

Fact2:

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

Then,

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

# Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

We will also have to use some facts

Fact1 (Holder Inequality):

For  $w \in V$  and  $z \in V^*$

$$\langle z, w \rangle \leq \|w\| \cdot \|z\|_* \quad (4.10)$$

where  $V^*$  is the dual of  $V$ .

Fact2:

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

Then,

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

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Then we have,

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

# Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Then we have,

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

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

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

$$x_{t+1} = \operatorname{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} D_{\Phi}(x, y_{t+1})$$

(4.12)

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

(4.12)



# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

(4.12)

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

(4.12)

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# 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  $\mathbb{E}\widetilde{g}(x) \in \partial f(x)$

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

**Stochastics**

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Stochastics

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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  $x \in \mathcal{X}$ .

# Stochastics

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# Stochastics

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

and

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

# Stochastic Mirror Descent

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Stochastic Mirror Descent

Let  $\Phi$  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

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

Recall the Mirror Descent proof...

## Mirror Descent Proof

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

...

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

Recall the Mirror Descent proof...

## Mirror Descent Proof

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

...

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

## Corollary

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

# SMD Convergence Proof

Proof:

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

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

**S-MD**

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# SMD Convergence Proof

Proof:

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

$$\begin{aligned}\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^t x_s\right) - f(x^*) &\leq \frac{1}{t}\mathbb{E}\sum_{s=1}^t (f(x_s) - f(x^*)) \\ &\leq \frac{1}{t}\mathbb{E}\sum_{s=1}^t \mathbb{E}(\tilde{g}(x_s)|x_s)^\top (x_s - x^*)\end{aligned}$$

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

**S-MD**

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# SMD Convergence Proof

Proof:

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

$$\begin{aligned}\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^t x_s\right) - f(x^*) &\leq \frac{1}{t}\mathbb{E}\sum_{s=1}^t (f(x_s) - f(x^*)) \\ &\leq \frac{1}{t}\mathbb{E}\sum_{s=1}^t \mathbb{E}(\tilde{g}(x_s)|x_s)^\top (x_s - x^*) \\ &= \frac{1}{t}\mathbb{E}\sum_{s=1}^t \tilde{g}(x_s)^\top (x_s - x^*).\end{aligned}$$

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

**S-MD**

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# SMD Convergence Proof

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

Proof:

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

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



# SMD Convergence Proof

Proof:

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

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

# Stochastic PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: PGD Convergence

Let  $f$  be  $\alpha$ -strongly convex, and assume that the stochastic oracle is such that  $\mathbb{E}\|\tilde{g}(x)\|_*^2 \leq B^2$ . 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:

# Stochastic PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: PGD Convergence

Let  $f$  be  $\alpha$ -strongly convex, and assume that the stochastic oracle is such that  $\mathbb{E}\|\tilde{g}(x)\|_*^2 \leq B^2$ . 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:

$$f(x_s) - f(x^*) \leq \mathbb{E} \tilde{g}^\top(x_s - x^*) - \frac{\alpha}{2} \|x_s - x^*\|^2$$

# Stochastic PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: PGD Convergence

Let  $f$  be  $\alpha$ -strongly convex, and assume that the stochastic oracle is such that  $\mathbb{E}\|\tilde{g}(x)\|_*^2 \leq B^2$ . 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:

$$\begin{aligned} f(x_s) - f(x^*) &\leq \mathbb{E} \tilde{g}^\top (x_s - x^*) - \frac{\alpha}{2} \|x_s - x^*\|^2 \\ &\dots \end{aligned}$$

# Stochastic PGD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: PGD Convergence

Let  $f$  be  $\alpha$ -strongly convex, and assume that the stochastic oracle is such that  $\mathbb{E}\|\tilde{g}(x)\|_*^2 \leq B^2$ . 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:

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

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

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



# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

$$\begin{aligned} &(0x_1 - 2x_2) + 2(x_2 - 3x_3) + 3(2x_3 - 4x_4) \dots + t((t-1)x_t - (t+1)x_{t+1}) \\ &= -t(t+1)x_{t+1} \end{aligned}$$

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

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

Then taking the sum from  $s = 1, \dots, t$

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

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

Then taking the sum from  $s = 1, \dots, t$

$$\sum_{s=1}^t s(f(x_s) - f(x^*)) \leq \frac{tB^2}{\alpha} - \frac{\alpha}{4} t(t+1) \|x_{t+1} - x^*\|^2$$

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

Then taking the sum from  $s = 1, \dots, t$

$$\sum_{s=1}^t s(f(x_s) - f(x^*)) \leq \frac{tB^2}{\alpha} - \frac{\alpha}{4} t(t+1) \|x_{t+1} - x^*\|^2$$

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

Then taking the sum from  $s = 1, \dots, t$

$$\sum_{s=1}^t s(f(x_s) - f(x^*)) \leq \frac{tB^2}{\alpha} - \frac{\alpha}{4} t(t+1) \|x_{t+1} - x^*\|^2$$

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

# PGD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

Then taking the sum from  $s = 1, \dots, t$

$$\sum_{s=1}^t s(f(x_s) - f(x^*)) \leq \frac{tB^2}{\alpha} - \frac{\alpha}{4} t(t+1) \|x_{t+1} - x^*\|^2$$

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

$$\leq \frac{2B^2}{\alpha(t+1)}$$

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



# Non-Smooth Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

**S-PGD**

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Stochastic PGD

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

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

# Non-Smooth Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Stochastic PGD

Let  $f$  be  $\alpha$ -strongly convex, and assume that the stochastic oracle is such that  $\mathbb{E}\|\tilde{g}(x)\|_*^2 \leq B^2$ . 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  $\alpha$ -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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

...

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

# Non-Smooth Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

## Theorem: Smooth S-MD

Let  $\Phi$  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  $\beta$ -smooth w.r.t.  $\|\cdot\|$ . Furthermore assume that the stochastic oracle is such that  $\mathbb{E}\|\nabla f(x) - \tilde{g}(x)\|_*^2 \leq \sigma^2$ . Then S-MD with stepsize  $\frac{1}{\beta+1/\eta}$  and  $\eta = \frac{R}{\sigma} \sqrt{\frac{2}{t}}$  satisfies

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

## Theorem: Smooth S-MD

Let  $\Phi$  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  $\beta$ -smooth w.r.t.  $\|\cdot\|$ . Furthermore assume that the stochastic oracle is such that  $\mathbb{E}\|\nabla f(x) - \tilde{g}(x)\|_*^2 \leq \sigma^2$ . 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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

One quick lemma!

## Lemma

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

Proof:

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



# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

One quick lemma!

## Lemma

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

Proof:

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

One quick lemma!

## Lemma

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

Proof:

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

One quick lemma!

## Lemma

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

Proof:

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

One quick lemma!

## Lemma

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

Proof:

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

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

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

One quick lemma!

## Lemma

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

Proof:

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

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

## Theorem: Smooth S-MD

...

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

$$\begin{aligned} f(x_{s+1}) - f(x_s) \\ \leq \tilde{g}_s^\top (x_{s+1} - x_s) \end{aligned}$$

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Smooth S-MD

...

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

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

# Smooth S-MD Convergence

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

## Theorem: Smooth S-MD

...

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

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



# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent  
Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas  
Convergence

Mirror Descent  
Lemmas

Convergence

Stochastics  
Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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



# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

**S-MD**

Mini-Batch SGD

References

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

# Smooth S-MD Convergence

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

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

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

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

# Mini-Batches

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

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

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

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

With a few assumptions:

$f$  is  $\beta$ -smooth

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

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

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

With a few assumptions:

$f$  is  $\beta$ -smooth

$$\|\tilde{g}(x)\|_2 \leq B$$

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

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

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

With a few assumptions:

$f$  is  $\beta$ -smooth

$$\|\tilde{g}(x)\|_2 \leq B$$

Convergence?

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References



# Mini-Batches

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

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

With a few assumptions:

$f$  is  $\beta$ -smooth

$$\|\tilde{g}(x)\|_2 \leq B$$

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,

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

Using the property of independence we get the following,

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

Using the property of independence we get the following,

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

Using the property of independence we get the following,

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

Convex  
Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

# Mini-Batches

Using the property of independence we get the following,

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

# Mini-Batches

Using the property of independence we get the following,

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

# Mini-Batches

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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

Then we can apply the previous theorem to get

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

When would you want to use Minibatch-SGD?

When computation can be distributed between multiple processors



# The End

## Convex Optimization

Brennan Gebotys

Introduction to  
Convexity

Main Theorems  
and Definitions

PGD

Lemmas

Convergence

Mirror Descent

Lemmas

Convergence

Stochastics

Non-Smooth Case

S-MD

S-PGD

Smooth Case

S-MD

Mini-Batch SGD

References

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.