Fundamentals of Machine Learning

Yassine Laguel

Mail : yassine.laguel@univ-cotedazur.fr

UNIVERSITÉ CÔTE D'AZUR

LESSON 8
OPTIMIZATION METHODS

1. Setting up an optimization problem

2. Recalls on Gradient Descent

3. The quadratic case

4. The smooth and strongly convex case

5. The smooth and merely convex case

6. Stochastic gradient descent

1. Setting up an optimization problem

Optimization for machine learning

Learning as optimization

The supervised learning setup

Recall the general supervised learning seeks to solve

$$\min_{f \in \mathcal{H}} r(f) = \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)],$$

where $\mathcal{H} \subset \mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ denotes a **hypothesis space**,
an infinite-dimensional space of potential **predictor** functions.
Direct minimization of r is very hard.

Excess risk decomposition

In practice, we rather decompose the risk as

$$\begin{aligned} r(f) - \inf_{g\in\mathcal{F}} r(g) \leq\ & r(f) - R(f) && \text{Statistical learning} \\ & + R(f) - R(\tilde{f}^\star) && \text{Optimization} \\ & + R(f^\star) - r(f^\star) && \text{Statistical learning} \\ & + r(f^\star) - \inf_{g\in\mathcal{F}} r(g) && \text{Approximation error} \end{aligned}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{h\in\mathcal{H}} R(h)$, and $f^\star = \operatorname*{argmin}_{h\in\mathcal{H}} r(h)$.

- Concentration theory provides powerful tools for the derivation of statistical learning bounds → see lessons 1 to 3.

Optimization for machine learning

Learning as optimization

The supervised learning setup

Recall the general supervised learning seeks to solve

$$\min_{f \in \mathcal{H}} r(f) = \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)],$$

where $\mathcal{H} \subset \mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ denotes a **hypothesis space**, an infinite-dimensional space of potential **predictor** functions.
Direct minimization of r is very hard.

Excess risk decomposition

In practice, we rather decompose the risk as

$$\begin{aligned} r(f) - \inf_{g\in\mathcal{F}} r(g) \leq\ & r(f) - R(f) && \text{Statistical learning} \\ & + R(f) - R(\tilde{f}^\star) && \text{Optimization} \\ & + R(f^\star) - r(f^\star) && \text{Statistical learning} \\ & + r(f^\star) - \inf_{g\in\mathcal{F}} r(g) && \text{Approximation error} \end{aligned}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{h\in\mathcal{H}} R(h)$, and $f^\star = \operatorname*{argmin}_{h\in\mathcal{H}} r(h)$.

- Concentration theory provides powerful tools for the derivation of statistical learning bounds → see lessons 1 to 3.
- The choice of the hypothesis space can significantly reduce the approximation error (linear models, kernel methods, etc…) → see lessons 4 to 6.
- From now on we will focus on the practical minimization of the empirical risk. We are entering the real of Optimization ✨

Formalising an optimization problem

Definition : Optimization Problem

An **optimization problem** is a mathematical problem of the form

$$\min_{x\in\mathcal{X}} \quad f(x)$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a given set, and $f : \mathbb{R}^d \to \mathbb{R}$ is a function.
f is often called the **objective function**. $\mathcal{X}$ is often called the **constraint set**.

Remarks :

- Assumptions on f and $\mathcal{X}$ are critical in the definition of $(\mathcal{P})$, as they strongly affect the type of methods required to solve it. In particular, look wheter
  - $\mathcal{X}$ has finite or infinite dimension, is convex, polyhedral, etc …
  - f enjoys some regularity properties such as smoothness, convexity, lispchitzness, etc …
  - one has access to exact oracles, approximate oracles, stochastic oracles, …

Back to basics

- Global minimas, local minimas and critical points

Definition : global and local minimas

A point $x^\star \in \mathbb{R}^d$ is called a **global minimum** for $\mathcal{P}$, if

$$f(x^\star) \leq f(x), \quad \forall x \in \mathcal{X}.$$

A point $x^\star \in \mathbb{R}^d$ is called a **local minimum** for $\mathcal{P}$, if there exists neighborhood $\mathcal{W}$ of $x^\star$ such that

$$f(x^\star) \leq f(x), \quad \forall x \in \mathcal{W} \cap \mathcal{X}.$$

Remarks :

- Global minima are local minima but the reverse is false!

- Fermat's condition

Theorem : Fermat's condition

Assume $\mathcal{X} = \mathbb{R}^d$, and f is differentiable.

Then any local minimum $x^\star$ of f satisfies $\nabla f(x^\star) = 0$.

Points x satisfying $\nabla f(x) = 0$ are called **critical points**.

Back to basics

- Global minimas, local minimas and critical points

Definition : global and local minimas

A point $x^\star \in \mathbb{R}^d$ is called a **global minimum** for $\mathcal{P}$, if

$$f(x^\star) \leq f(x), \quad \forall x \in \mathcal{X}.$$

A point $x^\star \in \mathbb{R}^d$ is called a **local minimum** for $\mathcal{P}$, if there exists neighborhood $\mathcal{W}$ of $x^\star$ such that

$$f(x^\star) \leq f(x), \quad \forall x \in \mathcal{W} \cap \mathcal{X}.$$

Remarks :

- Global minima are local minima but the reverse is false!

- Fermat's condition

Theorem : Fermat's condition

Assume $\mathcal{X} = \mathbb{R}^d$, and f is differentiable.

Then any local minimum $x^\star$ of f satisfies $\nabla f(x^\star) = 0$.

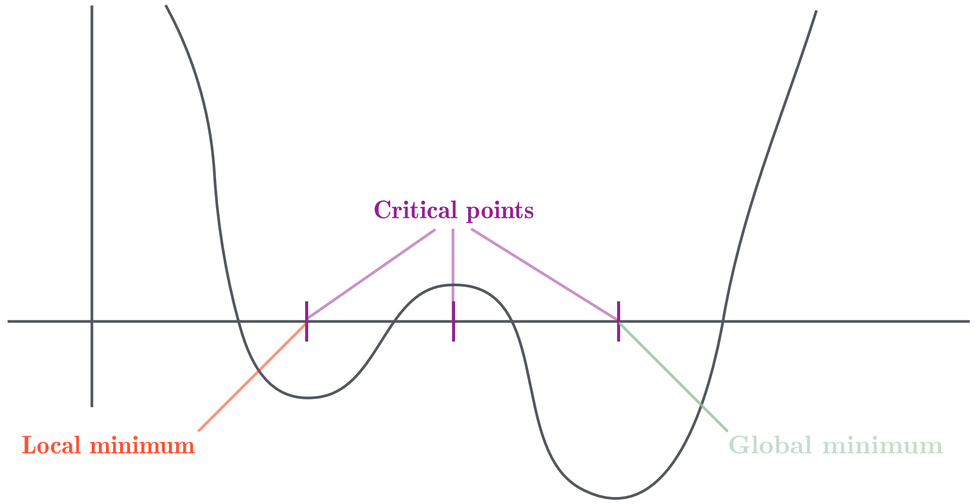
Points x satisfying $\nabla f(x) = 0$ are called **critical points**.

Remarks :

- Thus we have the implications

$$\text{Global Minima} \implies \text{Local Minima} \implies \text{Critical Point}$$

- All reverse implications are false!
- As an illustration :



First benefits of convexity

Recalls on convexity

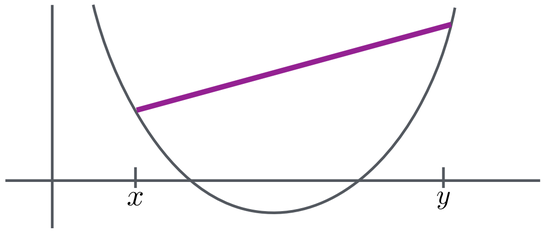
Definition : convex functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \mathbb{R}^d, \forall \lambda \in [0,1].$$

Remarks :

- Intuitively



Proposition : convexity for differentiable functions

For $f : \mathbb{R}^d \to \mathbb{R}$ differentiable, we have

$$f \text{ convex} \Leftrightarrow f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in \mathbb{R}^d.$$

Remarks :

- In other words, f is convex if and only if, the graph of f lies above its tangent lines.
- We directly deduce that for differentiable function, we have

$$\text{Global minima} \Leftrightarrow \text{Local minima} \Leftrightarrow \text{Critical point}$$

First benefits of convexity

Recalls on convexity

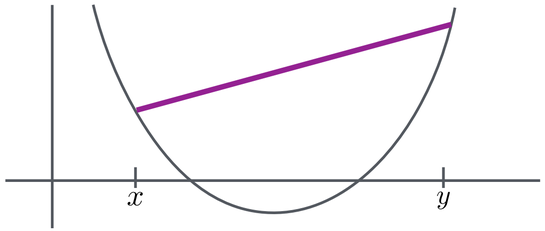
Definition : convex functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \mathbb{R}^d, \forall \lambda \in [0,1].$$

Remarks :

- Intuitively



Proposition : convexity for differentiable functions

For $f : \mathbb{R}^d \to \mathbb{R}$ differentiable, we have

$$f \text{ convex} \Leftrightarrow f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in \mathbb{R}^d.$$

Remarks :

- In other words, f is convex if and only if, the graph of f lies above its tangent lines.
- We directly deduce that for differentiable function, we have

  Global minima $\Leftrightarrow$ Local minima $\Leftrightarrow$ Critical point

Convex functions and supporting hyperplanes

Theorem : Convex functions through supporting hyperplanes

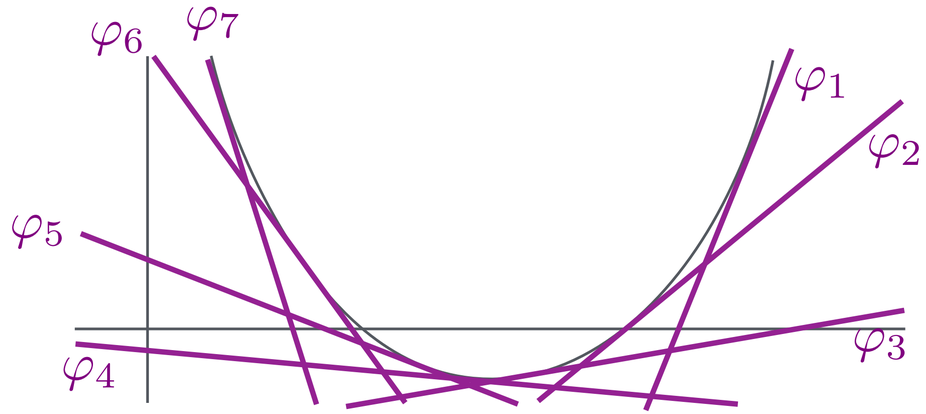
Any continuous and convex function $f : \mathbb{R} \to \mathbb{R}$, satisfies

$$f(x) = \sup_{\substack{\varphi \text{ affine} \\ \varphi \leq f}} \varphi(x).$$

Proof : on the blackboard.

Remarks :

- Intuitively



1. Setting up an optimization problem

2. Recalls on Gradient Descent

1st order methods

- Gradients provide descent directions

1st order methods

- Gradients provide descent directions
  - The differential plays a key role for linearly approximating an objective function.

1st order methods

- Gradients provide descent directions
  - The differential plays a key role for linearly approximating an objective function.
  - Gradients furthermore provide a direction of ascent

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.
- Exercise: Find the set of valid values for α when f is quadratic.

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.
- Exercise: Find the set of valid values for α when f is quadratic.

1st-order methods

- 1^{st}-order methods are methods that utilize the gradient of f to find a minimum.

$$x_{k+1} = \mathcal{M}\left(x_0, \ldots, x_k, \nabla f(x_0), \ldots, \nabla f(x_k)\right)$$

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.
- Exercise: Find the set of valid values for α when f is quadratic.

1st-order methods

- 1^{st}-order methods are methods that utilize the gradient of f to find a minimum.

$$x_{k+1} = \mathcal{M}\left(x_0, \ldots, x_k, \nabla f(x_0), \ldots, \nabla f(x_k)\right)$$

Definition

The simplest instantiation of such methods is the **gradient descent** algorithm. Given $x_0 \in \mathbb{R}^d$, the algorithm iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.
- Exercise: Find the set of valid values for α when f is quadratic.

1st-order methods

- 1^{st}-order methods are methods that utilize the gradient of f to find a minimum.

$$x_{k+1} = \mathcal{M}\left(x_0, \ldots, x_k, \nabla f(x_0), \ldots, \nabla f(x_k)\right)$$

Definition

The simplest instantiation of such methods is the **gradient descent** algorithm. Given $x_0 \in \mathbb{R}^d$, the algorithm iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Remarks :

- The parameter α is often called the **stepsize** of the method.

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.

Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the white board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.
- Exercise: Find the set of valid values for α when f is quadratic.

1st-order methods

- 1^{st}-order methods are methods that utilize the gradient of f to find a minimum.

$$x_{k+1} = \mathcal{M}\left(x_0, \ldots, x_k, \nabla f(x_0), \ldots, \nabla f(x_k)\right)$$

Definition

The simplest instantiation of such methods is the **gradient descent** algorithm. Given $x_0 \in \mathbb{R}^d$, the algorithm iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Remarks :

- The parameter α is often called the **stepsize** of the method.
- Setting this parameter is often of critical importance to ensure the convergence of x_k towards $x^\star$.

1st order methods

Gradients provide descent directions

- The differential plays a key role for linearly approximating an objective function.
- Gradients furthermore provide a direction of ascent

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then, for any $x \in \mathbb{R}^d$ not critical, there exists $\alpha > 0$ such that

$$f(x - \alpha \nabla f(x)) < f(x).$$

Proof : on the black board.

Remarks :

- There is extensive research on how to find α to maximize the descent of f along the direction $-\alpha \nabla f(x)$.
- Exercise: Find the set of valid values for α when f is quadratic.

1st-order methods

- 1^{st}-order methods are methods that utilize the gradient of f to find a minimum.

$$x_{k+1} = \mathcal{M}\left(x_0, \ldots, x_k, \nabla f(x_0), \ldots, \nabla f(x_k)\right)$$

Definition

The simplest instantiation of such methods is the **gradient descent** algorithm. Given $x_0 \in \mathbb{R}^d$, the algorithm iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Remarks :

- The parameter α is often called the **stepsize** of the method.
- Setting this parameter is often of critical importance to ensure the convergence of x_k towards $x^\star$.
- It is often set depending on the regularity properties of f.

1. Setting up an optimization problem

2. Recalls on Gradient Descent

3. The quadratic case

Gradient Descent on quadratics

- Quadratic optimization

  We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

  where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$

where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.
- OLS constitute a standard example of quadratic optimization.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$ where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.
- OLS constitute a standard example of quadratic optimization.

Proposition: Gradient descent on quadratics

Let $Q \in \mathcal{S}_d^{++}(\mathbb{R}), \mu = \min \mathrm{Sp}(Q),$ and $L = \max \mathrm{Sp}(Q)$.

Let $f_Q : x \mapsto \frac{1}{2}(x - x^\star)^\top Q(x - x^\star),$ where $x^\star \in \mathbb{R}^d$.

Let $(x_n)_{n\geq 0}$ be the sequence generated by (GD) on f_Q, initialized at $x_0 \in \mathbb{R}^d,$ with a constant stepsize $\gamma \leq \frac{2}{L}$.

Then, for all $n \in \mathbb{N}$,

$$\|x_n - x^\star\| \leq (1 - \gamma\mu)^n \, \|x_0 - x^\star\|.$$

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$ where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.
- OLS constitute a standard example of quadratic optimization.

Proposition: Gradient descent on quadratics

Let $Q \in \mathcal{S}_d^{++}(\mathbb{R}), \mu = \min \mathrm{Sp}(Q),$ and $L = \max \mathrm{Sp}(Q)$.
Let $f_Q : x \mapsto \frac{1}{2}(x - x^\star)^\top Q(x - x^\star),$ where $x^\star \in \mathbb{R}^d$.
Let $(x_n)_{n\geq 0}$ be the sequence generated by (GD) on f_Q,
initialized at $x_0 \in \mathbb{R}^d,$ with a constant stepsize $\gamma \leq \frac{2}{L}$.

Then, for all $n \in \mathbb{N}$,

$$\|x_n - x^\star\| \leq (1 - \gamma\mu)^n \, \|x_0 - x^\star\|.$$

Proof : on the black board.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$ where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.
- OLS constitute a standard example of quadratic optimization.

Proposition: Gradient descent on quadratics

Let $Q \in \mathcal{S}_d^{++}(\mathbb{R}), \mu = \min \mathrm{Sp}(Q),$ and $L = \max \mathrm{Sp}(Q)$.
Let $f_Q : x \mapsto \frac{1}{2}(x - x^\star)^\top Q(x - x^\star),$ where $x^\star \in \mathbb{R}^d$.
Let $(x_n)_{n\geq 0}$ be the sequence generated by (GD) on f_Q,
initialized at $x_0 \in \mathbb{R}^d,$ with a constant stepsize $\gamma \leq \frac{2}{L}$.

Then, for all $n \in \mathbb{N}$,

$$\|x_n - x^\star\| \leq (1 - \gamma\mu)^n \, \|x_0 - x^\star\|.$$

Proof : on the black board.

Remarks :

- Therefore for any positive definite matrix, the convergence of (GD) occurs at a geometric rate. Optimisers like to describe convergence rates in log scale. In our case :

$$\log(\|x_n - x^\star\|) \leq n \, \log\left(\frac{L-\mu}{L+\mu}\right) + \log\left(\, \|x_0 - x^\star\|\right).$$

We say that (GD) achieves linear convergence.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$ where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.
- OLS constitute a standard example of quadratic optimization.

Proposition: Gradient descent on quadratics

Let $Q \in \mathcal{S}_d^{++}(\mathbb{R}), \mu = \min \mathrm{Sp}(Q),$ and $L = \max \mathrm{Sp}(Q)$.
Let $f_Q : x \mapsto \frac{1}{2}(x - x^\star)^\top Q(x - x^\star),$ where $x^\star \in \mathbb{R}^d$.
Let $(x_n)_{n\geq 0}$ be the sequence generated by (GD) on f_Q,
initialized at $x_0 \in \mathbb{R}^d,$ with a constant stepsize $\gamma \leq \frac{2}{L}$.

Then, for all $n \in \mathbb{N}$,

$$\|x_n - x^\star\| \leq (1 - \gamma\mu)^n \, \|x_0 - x^\star\|.$$

Proof : on the black board.

Remarks :

- Therefore for any positive definite matrix, the convergence of (GD) occurs at a geometric rate. Optimisers like to describe convergence rates in log scale. In our case :

$$\log(\|x_n - x^\star\|) \leq n \, \log\left(\frac{L-\mu}{L+\mu}\right) + \log\left(\, \|x_0 - x^\star\|\right).$$

  We say that (GD) achieves linear convergence.
- When using the optimal stepsize $\gamma = \frac{2}{L+\mu}$, the number n_ε of operations to achieve $\left(\frac{L-\mu}{L+\mu}\right)^{n_\varepsilon} \|x_0 - x^\star\| \leq \varepsilon$ is $n_\varepsilon = \frac{1}{\left(\frac{L-\mu}{L+\mu}\right)} \log\left(\frac{\|x_0 - x^\star\|}{\varepsilon}\right)$.

Gradient Descent on quadratics

Quadratic optimization

We call **quadratic**, any function φ of the form $\varphi : x \mapsto \frac{1}{2}x^\top Qx + p\top x + c$ where $Q \in S_d(\mathbb{R})$ is symmetric definite semi positive, $p \in R^d$ and $c \in \mathbb{R}$.

Remarks :

- Quadratics are fundamental in the development of numerical algorithms.
  - When developing a numerical method, it is very common to try it first on quadratics.
  - We will see how the spectral properties of Q often dictate the speed of convergence of the method.
  - On quadratics 1^{st}-order methods reduce to linear dynamical systems.
- Exercise : Compute the gradient of φ.
- OLS constitute a standard example of quadratic optimization.

Proposition: Gradient descent on quadratics

Let $Q \in \mathcal{S}_d^{++}(\mathbb{R}), \mu = \min \mathrm{Sp}(Q),$ and $L = \max \mathrm{Sp}(Q)$.
Let $f_Q : x \mapsto \frac{1}{2}(x - x^\star)^\top Q(x - x^\star),$ where $x^\star \in \mathbb{R}^d$.
Let $(x_n)_{n\geq 0}$ be the sequence generated by (GD) on f_Q, initialized at $x_0 \in \mathbb{R}^d,$ with a constant stepsize $\gamma \leq \frac{2}{L}$.

Then, for all $n \in \mathbb{N}$,

$$\|x_n - x^\star\| \leq (1 - \gamma\mu)^n \, \|x_0 - x^\star\|.$$

Proof : on the black board.

Remarks :

- Therefore for any positive definite matrix, the convergence of (GD) occurs at a geometric rate. Optimisers like to describe convergence rates in log scale. In our case :

$$\log(\|x_n - x^\star\|) \leq n \, \log\left(\frac{L-\mu}{L+\mu}\right) + \log(\, \|x_0 - x^\star\|)\,.$$

  We say that (GD) achieves linear convergence.
- When using the optimal stepsize $\gamma = \frac{2}{L+\mu}$, the number n_ε of operations to achieve $\left(\frac{L-\mu}{L+\mu}\right)^{n_\varepsilon} \|x_0 - x^\star\| \leq \varepsilon$ is $n_\varepsilon = \frac{1}{\left(\frac{L-\mu}{L+\mu}\right)} \log\left(\frac{\|x_0 - x^\star\|}{\varepsilon}\right)$.
- When $\mu \ll L,$ a simple Taylor expansion gives $\log\left(\frac{L-\mu}{L+\mu}\right)^{-1} \sim \frac{L}{\mu}$.

  The constant $\frac{L}{\mu} = \|Q\| \cdot \|Q^{-1}\|$ is called the **condition number** of Q.

  This quantity characterizes the best speed of convergence allowed by (GD) on quadratics. We will soon generalize it to non-quadratic problems.

1. Setting up an optimization problem

2. Recalls on Gradient Descent

3. The quadratic case

4. The smooth and strongly convex case

From positive definiteness to strong convexity

- Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

From positive definiteness to strong convexity

- Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0,$ we recover the standard convexity property.

From positive definiteness to strong convexity

- Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0$, we recover the standard convexity property.
- One may easily see that if f is μ_1-strongly convex, for some $\mu_1 \geq 0$, then f is also μ_2-strongly convex, for any $\mu_2 \geq \mu_1$.

From positive definiteness to strong convexity

- Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0$, we recover the standard convexity property.
- One may easily see that if f is μ_1-strongly convex, for some $\mu_1 \geq 0$, then f is also μ_2-strongly convex, for any $\mu_2 \geq \mu_1$.
- In non-smooth optimization, a popular class of functions is the set of **weakly-convex** functions, for which there exists some $\mu \geq 0$ such that $x \mapsto f(x) + \frac{\mu}{2}\|x\|_2^2$ is convex.

From positive definiteness to strong convexity

Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0$, we recover the standard convexity property.
- One may easily see that if f is μ_1-strongly convex, for some $\mu_1 \geq 0$, then f is also μ_2-strongly convex, for any $\mu_2 \geq \mu_1$.
- In non-smooth optimization, a popular class of functions is the set of **weakly-convex** functions, for which there exists some $\mu \geq 0$ such that $x \mapsto f(x) + \frac{\mu}{2}\|x\|_2^2$ is convex.

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then

$$f \text{ is } \mu\text{-strongly convex}$$
$$\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

From positive definiteness to strong convexity

- Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0$, we recover the standard convexity property.
- One may easily see that if f is μ_1-strongly convex, for some $\mu_1 \geq 0$, then f is also μ_2-strongly convex, for any $\mu_2 \geq \mu_1$.
- In non-smooth optimization, a popular class of functions is the set of **weakly-convex** functions, for which there exists some $\mu \geq 0$ such that $x \mapsto f(x) + \frac{\mu}{2}\|x\|_2^2$ is convex.

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then

$$f \text{ is } \mu\text{-strongly convex}$$
$$\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Proof : On the white board!

From positive definiteness to strong convexity

Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0,$ we recover the standard convexity property.
- One may easily see that if f is μ_1-strongly convex, for some $\mu_1 \geq 0,$ then f is also μ_2-strongly convex, for any $\mu_2 \geq \mu_1$.
- In non-smooth optimization, a popular class of functions is the set of **weakly-convex** functions, for which there exists some $\mu \geq 0$ such that $x \mapsto f(x) + \frac{\mu}{2}\|x\|_2^2$ is convex.

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then

$$f \text{ is } \mu\text{-strongly convex}$$
$$\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Proof : On the white board!

Remarks :

- When $\mu > 0,$ this means that at any point $x,$ f can be lowerbounded by a quadratic tangent to f in x and of dominant coefficient $\frac{\mu}{2}\|y\|^2$

From positive definiteness to strong convexity

Strong convexity

Definition

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$is said to be **$\mu$-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Remarks :

- When $\mu = 0$, we recover the standard convexity property.
- One may easily see that if f is μ_1-strongly convex, for some $\mu_1 \geq 0$, then f is also μ_2-strongly convex, for any $\mu_2 \geq \mu_1$.
- In non-smooth optimization, a popular class of functions is the set of **weakly-convex** functions, for which there exists some $\mu \geq 0$ such that $x \mapsto f(x) + \frac{\mu}{2}\|x\|_2^2$ is convex.

Proposition

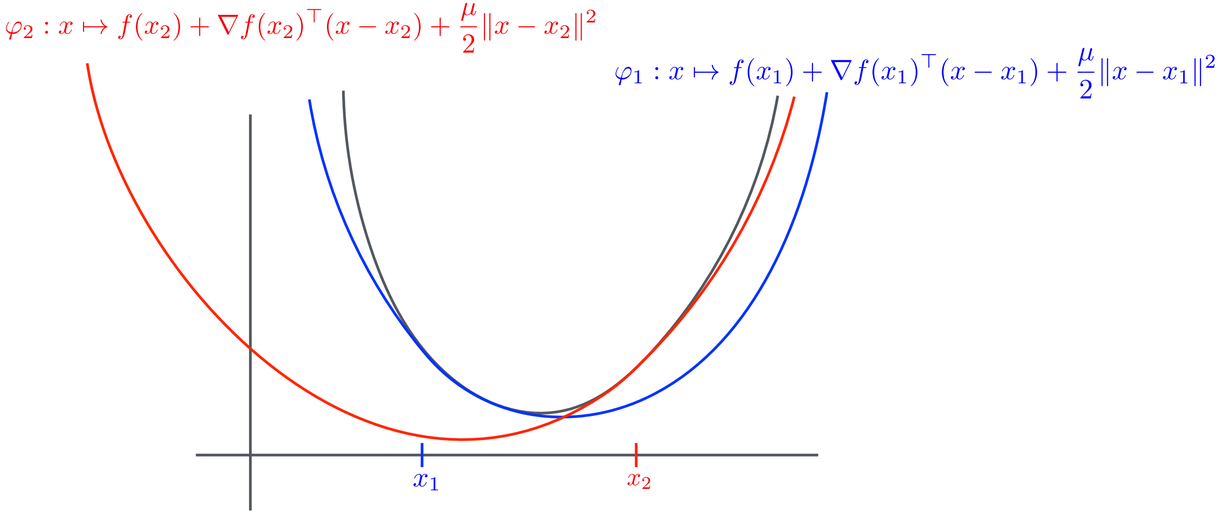
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
Then

$$f \text{ is } \mu\text{-strongly convex}$$
$$\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Proof : On the white board!

Remarks :

- When $\mu > 0$, this means that at any point x, f can be lowerbounded by a quadratic tangent to f in x and of dominant coefficient $\frac{\mu}{2}\|y\|^2$
- Intuitively



Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

  - The fundamental theorem of Calculus actually gives a further characterization :

$$f(y) = f(x) + \int_{s=0}^{1} \nabla f(x + s(y - x))^\top (y - x) \mathrm{d}s$$

Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

  - The fundamental theorem of Calculus actually gives a further characterization :

$$f(y) = f(x) + \int_{s=0}^{1} \nabla f(x + s(y - x))^\top (y - x) \mathrm{d}s$$

- Smooth functions

Definition

Let $\ell \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be ℓ-smooth if f is differentiable on $\mathbb{R}^d$, and ∇f is ℓ-lipschitz on $\mathbb{R}^d$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

  - The fundamental theorem of Calculus actually gives a further characterization :

$$f(y) = f(x) + \int_{s=0}^{1} \nabla f(x + s(y - x))^\top (y - x) \mathrm{d}s$$

- Smooth functions

Definition

Let $\ell \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be ℓ-smooth if f is differentiable on $\mathbb{R}^d$, and ∇f is ℓ-lipschitz on $\mathbb{R}^d$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
If f is ℓ-smooth, then

$$\left|f(y) - \left[f(x) + \nabla f(x)^\top (y - x)\right]\right| \leq \frac{\ell}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

  - The fundamental theorem of Calculus actually gives a further characterization :

$$f(y) = f(x) + \int_{s=0}^{1} \nabla f(x + s(y - x))^\top (y - x) \mathrm{d}s$$

- Smooth functions

Definition

Let $\ell \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be ℓ-smooth if f is differentiable on $\mathbb{R}^d$, and ∇f is ℓ-lipschitz on $\mathbb{R}^d$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
If f is ℓ-smooth, then

$$\left|f(y) - \left[f(x) + \nabla f(x)^\top (y - x)\right]\right| \leq \frac{\ell}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Proof : On the white board!

Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

  - The fundamental theorem of Calculus actually gives a further characterization :

$$f(y) = f(x) + \int_{s=0}^{1} \nabla f(x + s(y - x))^\top (y - x) \mathrm{d}s$$

- Smooth functions

Definition

Let $\ell \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be ℓ-smooth if f is differentiable on $\mathbb{R}^d$, and ∇f is ℓ-lipschitz on $\mathbb{R}^d$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
If f is ℓ-smooth, then

$$\left|f(y) - \left[f(x) + \nabla f(x)^\top (y - x)\right]\right| \leq \frac{\ell}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Proof : On the white board!

Remarks :

- If $\ell = 0$, then f is an affine function.

Smoothness

- Recall that...
  - Continuously differentiable functions can locally be approximated by their first-order expansion :

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o\left(\|y - x\|^2\right)$$

  - The fundamental theorem of Calculus actually gives a further characterization :

$$f(y) = f(x) + \int_{s=0}^{1} \nabla f(x + s(y - x))^\top (y - x) \mathrm{d}s$$

- Smooth functions

Definition

Let $\ell \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be ℓ-smooth if f is differentiable on $\mathbb{R}^d$, and ∇f is ℓ-lipschitz on $\mathbb{R}^d$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

Proposition

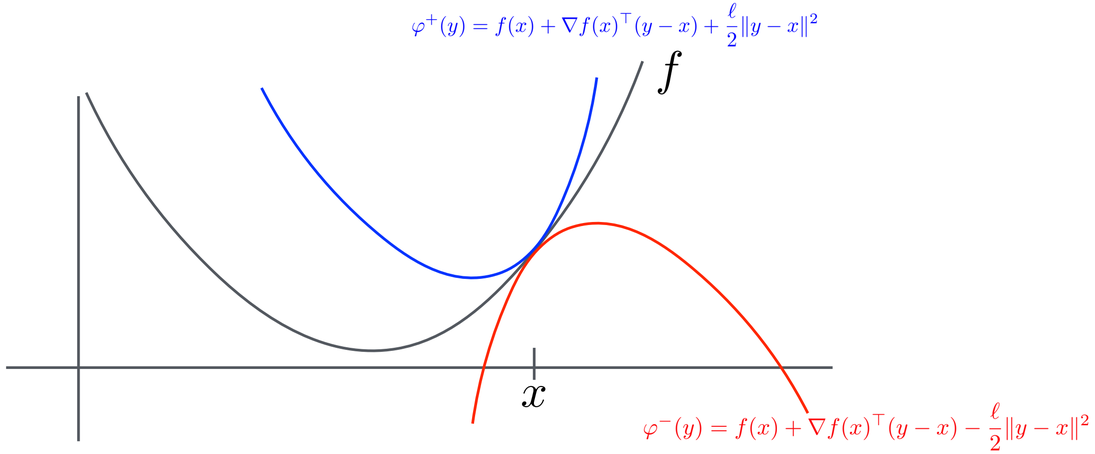
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.
If f is ℓ-smooth, then

$$\left|f(y) - \left[f(x) + \nabla f(x)^\top (y - x)\right]\right| \leq \frac{\ell}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d$$

Proof : On the white board!

Remarks :

- If $\ell = 0$, then f is an affine function.
- In general, ℓ-smooth function can be upperbounded and lower-bounded by quadratics.



Gradient Descent on Smooth and Strongly Convex Functions

- Gradient Descent on S_μ^ℓ

  Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Gradient Descent on Smooth and Strongly Convex Functions

Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S_\mu^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Gradient Descent on Smooth and Strongly Convex Functions

Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S_\mu^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$\|x_t - x^\star\|^2 \leq \left(1 - \frac{\mu}{\ell}\right)^t \|x_0 - x_\star\|^2$$

Gradient Descent on Smooth and Strongly Convex Functions

Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S_\mu^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$\|x_t - x^\star\|^2 \leq \left(1 - \frac{\mu}{\ell}\right)^t \|x_0 - x_\star\|^2$$

Proof : On the white board!

Gradient Descent on Smooth and Strongly Convex Functions

Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S_\mu^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t\geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t\geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$\|x_t - x^\star\|^2 \leq \left(1 - \frac{\mu}{\ell}\right)^t \|x_0 - x_\star\|^2$$

Proof : On the white board!

Remarks :

- In other words, for smooth and strongly convex function, the gradient descent method converges at a geometric rate! Machine learning scientists often consider $\log\left(\|x_t - x^\star\|^2\right)$ and refer the to the speed of convergence of the gradient descent method as linear.

Gradient Descent on Smooth and Strongly Convex Functions

Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S_\mu^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t\geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t\geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$\|x_t - x^\star\|^2 \leq \left(1 - \frac{\mu}{\ell}\right)^t \|x_0 - x_\star\|^2$$

Proof : On the white board!

Remarks :

- In other words, for smooth and strongly convex function, the gradient descent method converges at a geometric rate! Machine learning scientists often consider $\log\left(\|x_t - x^\star\|^2\right)$ and refer the to the speed of convergence of the gradient descent method as linear.
- The constant $\kappa = \frac{\ell}{\mu}$ is often called the condition number of the problem : the greater κ is, the slower Gradient Descent converges.

  Hence κ quantifies the hardness to minimize f through Gradient Descent.

Gradient Descent on Smooth and Strongly Convex Functions

Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S_\mu^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t\geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t\geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$\|x_t - x^\star\|^2 \leq \left(1 - \frac{\mu}{\ell}\right)^t \|x_0 - x_\star\|^2$$

Proof : On the white board!

Remarks :

- In other words, for smooth and strongly convex function, the gradient descent method converges at a geometric rate! Machine learning scientists often consider $\log\left(\|x_t - x^\star\|^2\right)$ and refer the to the speed of convergence of the gradient descent method as linear.
- The constant $\kappa = \frac{\ell}{\mu}$ is often called the condition number of the problem : the greater κ is, the slower Gradient Descent converges.

  Hence κ quantifies the hardness to minimize f through Gradient Descent.

- There are many variants of Gradient Descent aiming at having better dependency in their rates with respect to κ.

  For instance the accelerated gradient descent method (Nesterov 1984) exhibits a rate of the form

$$\|x_t - x^\star\|^2 \leq \left(1 - \sqrt{\frac{1}{\kappa}}\right)^t \|x_0 - x^\star\|^2$$

  which brings tremendous improvements for ill-conditioned problems.

1. Setting up an optimization problem

2. Recalls on Gradient Descent

3. The quadratic case

4. The smooth and strongly convex case

5. The smooth and merely convex case

Gradient Descent on Smooth and Merely Convex Functions

- Gradient Descent on S^ℓ

  Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$f(x_t) - f(x^\star) \leq \frac{2\ell \|x_0 - x^\star\|^2}{t}$$

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$f(x_t) - f(x^\star) \leq \frac{2\ell \|x_0 - x^\star\|^2}{t}$$

Proof : On the white board!

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$f(x_t) - f(x^\star) \leq \frac{2\ell \|x_0 - x^\star\|^2}{t}$$

Proof : On the white board!

Remarks :

- The convergence is **sub-linear**!

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$f(x_t) - f(x^\star) \leq \frac{2\ell \|x_0 - x^\star\|^2}{t}$$

Proof : On the white board!

Remarks :

- The convergence is **sub-linear**!
- κ is no longer the metric of complexity dictating the performance of (GD), but ℓ, the smoothness constant still does.

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$f(x_t) - f(x^\star) \leq \frac{2\ell \|x_0 - x^\star\|^2}{t}$$

Proof : On the white board!

Remarks :

- The convergence is **sub-linear**!
- κ is no longer the metric of complexity dictating the performance of (GD), but ℓ, the smoothness constant still does.
- The metric of convergence is now the functional gap $f(x_k) - f(x^\star)$. It is a common metric in non-strongly convex settings. Alternatively, we often encounter the norm of the gradient $\|\nabla f(x_k)\|$.

Gradient Descent on Smooth and Merely Convex Functions

Gradient Descent on S^ℓ

Let S^ℓ be the set of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the gradient descent method initialized at $x_0 \in \mathbb{R}^d$ on a function $f \in S^\ell$:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

for a given sequence of stepsize $(\alpha_t)_{t \geq 0}$.

Proposition

Assume the sequence $(\alpha_t)_{t \geq 0}$ is set to be constant, equal to $\alpha = \frac{1}{\ell}$. Then, for all $t > 0$,

$$f(x_t) - f(x^\star) \leq \frac{2\ell \|x_0 - x^\star\|^2}{t}$$

Proof : On the white board!

Remarks :

- The convergence is **sub-linear**!
- κ is no longer the metric of complexity dictating the performance of (GD), but ℓ, the smoothness constant still does.
- The metric of convergence is now the functional gap $f(x_k) - f(x^\star)$. It is a common metric in non-strongly convex settings. Alternatively, we often encounter the norm of the gradient $\|\nabla f(x_k)\|$.
- Here again, there exists accelerated variants of (GD) that achieve a much faster rate, of order $\mathcal{O}\left(\frac{1}{t^2}\right)$.

1. Setting up an optimization problem

2. Recalls on Gradient Descent

3. The quadratic case

4. The smooth and strongly convex case

5. The smooth and merely convex case

6. Stochastic gradient descent

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

  - The gradient method on the above problem writes

$$w_{t+1} = w_t - \alpha \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] \left(w_t\right)$$

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

  - The gradient method on the above problem writes

$$\begin{aligned} w_{t+1} &= w_t - \alpha \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] \left(w_t\right) \\ &= w_t - \alpha \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \left(w_t\right) \end{aligned}$$

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

  - The gradient method on the above problem writes

$$\begin{aligned} w_{t+1} &= w_t - \alpha \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] \left(w_t\right) \\ &= w_t - \alpha \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \left(w_t\right) \end{aligned}$$

  - In modern ML applications, n is very large, which which makes GD or SubGradient impractical.

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

  - The gradient method on the above problem writes

$$\begin{aligned} w_{t+1} &= w_t - \alpha \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] \left(w_t\right) \\ &= w_t - \alpha \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \left(w_t\right) \end{aligned}$$

  - In modern ML applications, n is very large, which which makes GD or SubGradient impractical.

- From batch to mini-batch methods
  - Stochastic approximation methods look for a cheap way to approximate the full batch gradient

$$\frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] \left(w_t\right)$$

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

  - The gradient method on the above problem writes

$$\begin{aligned} w_{t+1} &= w_t - \alpha \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] (w_t) \\ &= w_t - \alpha \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) (w_t) \end{aligned}$$

  - In modern ML applications, n is very large, which which makes GD or SubGradient impractical.

- From batch to mini-batch methods
  - Stochastic approximation methods look for a cheap way to approximate the full batch gradient

$$\frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] (w_t)$$

  - A simple way to approximate the full batch gradient is to randomly select one datapoint, and replace the full gradient by the gradient at this single point

$$\frac{\partial}{\partial w} \left[\mathcal{L}\left(y_{\boldsymbol{i}}, f\left(w, x_{\boldsymbol{i}}\right)\right) \right] (w_t) \simeq \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] (w_t),$$

$$\boldsymbol{i} \sim U\{1, \ldots, n\}$$

Large-Scale Stochastic Optimization

- Limitations of batch gradient methods
  - Recall that ERM aims at solving

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

  - The gradient method on the above problem writes

$$\begin{aligned} w_{t+1} &= w_t - \alpha \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] (w_t) \\ &= w_t - \alpha \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) (w_t) \end{aligned}$$

  - In modern ML applications, n is very large, which which makes GD or SubGradient impractical.

- From batch to mini-batch methods
  - Stochastic approximation methods look for a cheap way to approximate the full batch gradient

$$\frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] (w_t)$$

  - A simple way to approximate the full batch gradient is to randomly select one datapoint, and replace the full gradient by the gradient at this single point

$$\frac{\partial}{\partial w} \left[\mathcal{L}\left(y_{\boldsymbol{i}}, f\left(w, x_{\boldsymbol{i}}\right)\right) \right] (w_t) \simeq \frac{\partial}{\partial w} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f\left(w, x_i\right)\right) \right] (w_t),$$

$$\boldsymbol{i} \sim U\left\{1, \ldots, n\right\}$$

  - We can replace a single datapoints with an average over a small number of randomly selected datapoints. This problem is called **mini-batching**.

Stocastic Gradient Method

- Stochastic Gradient Descent on S_μ^ℓ

  Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Stocastic Gradient Method

- Stochastic Gradient Descent on S_μ^ℓ

  Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

  Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

  $$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f(x_t)$$

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f\left(x_t\right)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f\left(x_t\right)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Then, for all $\gamma \leq \frac{2}{\ell}$, gradient descent run with constant stepsize $\alpha_t = \gamma$ satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] \leq (1-\gamma\mu)^n \|w_0 - w^\star\|^2 + \frac{\gamma G^2}{\mu}.$$

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f\left(x_t\right)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Then, for all $\gamma \leq \frac{2}{\ell}$, gradient descent run with constant stepsize $\alpha_t = \gamma$ satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] \leq (1-\gamma\mu)^n \|w_0 - w^\star\|^2 + \frac{\gamma G^2}{\mu}.$$

Proof : On the white board!

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f\left(x_t\right)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Then, for all $\gamma \leq \frac{2}{\ell}$, gradient descent run with constant stepsize $\alpha_t = \gamma$ satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] \leq (1-\gamma\mu)^n \|w_0 - w^\star\|^2 + \frac{\gamma G^2}{\mu}.$$

Proof : On the white board!

Remarks :

- We observe here a bias-variance trade-off.

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f\left(x_t\right)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Then, for all $\gamma \leq \frac{2}{\ell}$, gradient descent run with constant stepsize $\alpha_t = \gamma$ satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] \leq (1-\gamma\mu)^n \|w_0 - w^\star\|^2 + \frac{\gamma G^2}{\mu}.$$

Proof : On the white board!

Remarks :

- We observe here a bias-variance trade-off.
- The stepsize γ plays a key role in this trade-off.

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f\left(x_t\right)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Then, for all $\gamma \leq \frac{2}{\ell}$, gradient descent run with constant stepsize $\alpha_t = \gamma$ satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] \leq (1-\gamma\mu)^n \|w_0 - w^\star\|^2 + \frac{\gamma G^2}{\mu}.$$

Proof : On the white board!

Remarks :

- We observe here a bias-variance trade-off.
- The stepsize γ plays a key role in this trade-off.
- What effect does minibatching have on this algorithm ?

Stocastic Gradient Method

Stochastic Gradient Descent on S_μ^ℓ

Let S_μ^ℓ be the set of ℓ-smooth and μ-strongly convex functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Consider the stochastic gradient descent method initialized at $x_0 \in \mathbb{R}^d$ initialized on a function $f \in S_\mu^\ell$

$$x_{t+1} = x_t - \alpha_t \widetilde{\nabla} f(x_t)$$

Proposition

Assume $\widetilde{\nabla} f(x_t)$ satisfies $\mathbb{E}\left[\widetilde{\nabla} f(x_t) - \nabla f(x_t)\right] = 0$ and $\mathbb{E}\left[\|\widetilde{\nabla} f(x_t) - \nabla f(x_t)\|^2\right] \leq G^2$.

Then, for all $\gamma \leq \frac{2}{\ell}$, gradient descent run with constant stepsize $\alpha_t = \gamma$ satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] \leq (1-\gamma\mu)^n \|w_0 - w^\star\|^2 + \frac{\gamma G^2}{\mu}.$$

Proof : On the white board!

Remarks :

- We observe here a bias-variance trade-off.
- The stepsize γ plays a key role in this trade-off.
- What effect does minibatching have on this algorithm ?

Exercise :

- For a fixed number of iterations n, what γ could help minimize the right-hand side of our convergence result ?

IMPLEMENTATION