centre national
de la recherche
scientifique

université
de BORDEAUX

Master Thesis In Optimal Transport

Semi-discrete Optimal Transport for Large-Scale Problems

University of Pau and Pays de l'Adour
Master Mathematics and Applications
M2 Mathematics, Modeling and Simulation

Author

Moussa Atwi

Supervisors

Nicolas Papadakis

Arthur Leclaire

Acknowledgements

First, I would like to express my sincere gratitude to my supervisors, Nicolas PAPADAKIS and Arthur LECLAIRE who offered a friendly working environment, welcomed and introduced me to this very interesting topic of mathematics, as well as for giving me such an interesting topic which is lying in the intersection of measure theory, optimal transport and convex optimization with applications in various fields, such as machine learning and image processing.

I am truly thankful for all the help, advice and motivation they gave me to overcome challenges and difficulties encountered during the internship, carefully explaining and answering my questions whenever needed.

Likewise, I extend my sincere thanks to the referees for their helpful suggestions.

Finally, I am extremely grateful to my friends and family for their encouraging and believing in me.

Contents

Page

Acknowledgements 2

Abstract 4

Introduction 5

1 Preliminaries on Optimal Transport 1
- 1.1 Optimal transport problem . . . 1
  - 1.1.1 Monge and Kantorovich formulations . . . 1
  - 1.1.2 Duality . . . 4
  - 1.1.3 Existence and form of an optimal T . . . 7
- 1.2 Duality and Optimality conditions . . . 8
  - 1.2.1 Convexity, Legendre - Fenchel transform and subdifferential . . . 8
  - 1.2.2 c-Cyclical monotonicity and duality . . . 8
- 1.3 Semi-discrete Optimal Transport . . . 10
  - 1.3.1 Power diagram or Laguerre diagram . . . 10

2 Gradient descent algorithms for convex optimization. 13
- 2.1 Convexity, smoothness and strong convexity. . . . 13
  - 2.1.1 Convexity . . . 13
  - 2.1.2 Smoothness . . . 15
  - 2.1.3 Smooth and Convex . . . 16
  - 2.1.4 Strong convexity . . . 17
  - 2.1.5 Application . . . 19
- 2.2 Gradient Descent . . . 21
  - 2.2.1 Convergence for convex and smooth functions . . . 21
  - 2.2.2 Convergence for strongly convex and smooth convex functions . . . 22
  - 2.2.3 Application . . . 23
- 2.3 Stochastic Gradient Descent . . . 27
  - 2.3.1 Convex . . . 27
  - 2.3.2 Strongly convex . . . 28
  - 2.3.3 Application: Gradient Noise . . . 29

3 Stochastic Semi-discrete Optimization. 36
- 3.1 Average stochastic gradient descent-ASGD . . . 36
  - 3.1.1 Numerical study and results of the stochastic methods for OT . . . 37

Conclusion 39

Bibliography 40

Abstract

In this report, we first offer a quick review on optimal transport problems, duality and optimality conditions to end up introducing the framework of semi-discrete optimal transport. In semi-discrete optimal transport we introduce the notion of Laguerre cells L_i^w that is beneficial in the study of existence of optimal transport map T_w for the cost c between the absolutely continuous measure μ and the discrete measure ν, because $\nu = T_{w\#}\mu$ if and only if $\mu(L_i^w) = \nu_i$.

Afterwards, we present some notions about convex optimization, such as: convexity, smoothness and strong convexity in order to introduce a detailed explanation of the gradient descent (GD) and stochastic gradient descent (SGD). Generally, we proceed by proving the convergence for convex and strongly convex smooth functions.

We then consider two applications: first one is about the rate of convergence of our gradient descent algorithm in order to reach the local/global minima of the function according to the number of iterations and learning rate, the second one is to do some experiments on stochastic gradient descent on the function f by considering (2.45) and add time-dependent Gaussian Noise to every gradient with Mean Value of zero and certain Standard Deviation Value

$$w^{t+1} = w^t - \alpha_t(\nabla f(w^t) + N(0, \sigma_t^2)) \tag{1}$$

to see the impact on the distance to the global minimizer and on the convergent of the gradient descent.

Finally, we present a numerical study of the stochastic gradient descent for semi-discrete OT to approximate the optimal solution and maximize the loss function.

Introduction

Optimal transportation [5] is a very active research topic with applications in various fields, like economics, machine learning, or image processing. Optimal transport [4] gives a framework for comparing measures μ and ν. However, the practical impact of OT is still limited because of its computational burden. We propose averaged stochastic gradient descent algorithm which will be implemented in Python and applied to a large-scale semi-discrete OT problems [17]. Essentially one pays a cost for transporting one measure to another. To illustrate this, consider the first measure μ as a pile of sand and the second measure ν as a hole we wish to fill up. We assume that both measures are probability measures on spaces X and Y respectively. Let $c : X \times Y \longrightarrow [0, +\infty]$ be a convex cost function where $c(x, y)$ measures the cost of transporting one unit of mass from $x \in X$ to $y \in Y$. The optimal transport problem is how to transport μ to ν while minimizing the cost c. There are two ways to formulate the optimal transport problem: the Monge and Kantorovich formulations. We explain both these formulations in this report. Historically the Monge formulation comes before Kantorovich which is why we introduce Monge first. The Kantorovich formulation can be seen as a generalisation of Monge.

In this report we will focus primarily on the semi-discrete case of optimal transportation [6] meaning that μ is absolutely continuous whereas ν is supported on a finite set Y. This setting leads to a finite-dimensional concave problem that can be solved with deterministic [6] and stochastic [8], [7] solvers.

This report is divided into three chapters, which are as follows:

- For a start, we shall recall in Chapter 1 some basic facts about OT problem such as Monge and Kantorovich formulations, as well as we will recall some basic notions concerning analysis over Polish space in order to prove the existence of minimizers of (KP). Afterwards, we shall consider the dual problem (DP), where we maximize a linear functional with affine constraints. Besides, we define the c-transform in order to reformulate the (DP) as an unconstrained maximization problem. In addition, we consider the case where $c(x, y) = h(x - y)$, with h is strictly convex to explain the existence and the form of an optimal transport T. Finally, we will focus on the semi-discrete optimal transport and define the Laguerre cells and prove the feasibility and optimality of the transport map T_w between μ and $T_{w\#}\mu$ where w solves the concave optimization problem
$$(KP) = \max_{w \in \mathbb{R}^d} \mathcal{K}(w)$$
where the Kantorovitch functional $\mathcal{K}$ defined by
$$\mathcal{K}(w) = \int_{\mathbb{R}^d} w^c \, d\mu - \int_{\mathbb{R}^d} w \, d\nu = \int_{\mathbb{R}^d} w^c(x) \, d\mu(x) - \sum_{i=1}^{N} w_i \, \nu_i \tag{2}$$
where, $w^c : Y \longrightarrow \overline{\mathbb{R}}$ is the c-transform of $w : X \longrightarrow \overline{\mathbb{R}}$, defined by
$$w^c(y) = \inf_{x \in X} c(x, y) - w(x).$$

- In Chapter 2 we shall present some definitions and propositions about convex optimization, with detailed proofs such as convexity, smoothness and strong convexity. Besides, we offer an application to study the type of convexity of our function by calculating its eigenvalues. We also introduce the gradient descent algorithm and present a detailed proof of the convergence of the following iterates

  $$x^{t+1} = x^t - \alpha \nabla f(x^t),$$

  under the assumption that f is convex or strongly convex. In addition, we offer an application to study the impact of learning rate and number of iterations in the rate of convergence and the speed of convergence of the proposed gradient descent algorithm. Finally, we introduce the stochastic gradient descent for convex and strongly convex cases with a detailed proof of the conditions satisfied by the iteration of stochastic method and we offer an application to do some experiments on SGD by adding a Gaussian Noise to every gradient with Mean Value of zero and certain Standard Deviation Value in order to see the impact on the distance to the global minimizer and on the convergent of the gradient descent.

- The last Chapter is devoted to the study of the stochastic approach in order to solve the semi-discrete optimal transport problem for large-scale problems where the dimension is high. In order to minimize $-\mathcal{K}$ mentioned in the equation (2) we study the averaged stochastic gradient descent which we implement in Python to approximate the optimal solution w of the semi-discrete problem.

The main contribution in this work is to understand some notions about optimal transport and in particular semi-discrete optimal transportation. As well as, to implement gradient descent and stochastic gradient descent in order to have some observations for the convergence of the algorithms and the efficiency of such methods to estimate and approximate the optimal solution.

Chapter 1

Preliminaries on Optimal Transport

In this chapter, we present a quick review on optimal transport problems, duality and optimality conditions. Also we recall some basic notions concerning analysis over Polish space to end up introducing the framework of semi-discrete optimal transport.

1.1 Optimal transport problem

Given a Polish space (X , d) (i.e. a complete and separable metric space). We will denote by $\mathcal{P}(X)$ the set of Borel probability measures on X.

Let $X, Y \subset \mathbb{R}^d$ be two Polish spaces, $T : X \longrightarrow Y$ is a Borel map, and $\mu \in \mathcal{P}(X)$ the source measure, the target measure $T_{\#}\mu \in \mathcal{P}(Y)$, called the push forward of μ through T is defined by

$$T_{\#}\mu(A) = \mu(T^{-1}(A)),$$

for every Borel set A $\subset Y$. The push forward is characterized by the fact that,

$$\int_Y f d(T_{\#}\mu) = \int_X (f \circ T) d\mu$$

for every Borel function $f : Y \longrightarrow \mathbb{R}_+$.

We say that T is a transport map from μ to ν if $T_{\#}\mu = \nu$.

1.1.1 Monge and Kantorovich formulations

Problem 1.1.1. Given two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and a Borel cost function $c : X \times Y \longrightarrow \overline{\mathbb{R}}$ (cost to pay to move unit mass from X to Y). The Monge formulation of the non-convex optimal transport problem is:

$$(MP) \quad \inf \left\{ M(T) := \int c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\}, \tag{1.1}$$

among all transport maps T from μ to ν.

The Monge problem consists in finding a map $T : X \longrightarrow Y$ which transports the mass from μ to ν while minimizing the mass transportation cost (any transport map T achieving the minimum in (1.1) is called an *optimal transport map*).

Regardless of the choice of the cost function c, Monge's problem can be ill-posed because:

- There may not exist any admissible map T to allow mass to split (for instance, if μ is a Dirac delta δ_x, $x \in X$ and ν is not of the form δ_y for some $y \in Y$. Indeed, $T_{\#}\delta_x = \delta_{T(x)}$).
- The constraint on T ($T_{\#}\mu = \nu$) is not closed under weak convergence.

Remark 1.1.2. Any solution is said to be a feasible solution of a transportation problem if it satisfies the constraints.

To overcome these difficulties, we generalize via the Kantorovich formulation, which relaxed the Monge problem by casting problem (1.1) into a minimization over couplings $(X, Y) \sim \gamma$ (to say that $X \times Y$ is distributed according to the probability measure γ) rather than the set of maps, where γ should have marginals equal to μ and ν. We seek a transport plan that allows mass to split.

$\mu(A)$ tells us how much mass is in A and $\gamma(A \times B)$ denotes the amount of mass transported from A to B.
We need to conserve mass, choose point $x \in X$ and let us consider $\gamma(x, Y) = \mu(x)$ (total mass coming from X). More generally, if $A \subset X$, $\gamma(A, Y) = \mu(A)$, we say that μ is the marginal of γ on X. Also, we need ν to be the marginal of γ on Y. If $B \subset Y$, then $\gamma(X, B) = \nu(B)$.

Again we have a source measure μ supported on X and a target measure ν supported on Y, we want to know how much mass gets moved from x to y for any combination (x, y) in $X \times Y$. We store this in another measure called γ defined on product space $X \times Y$.

Problem 1.1.3. Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \longrightarrow [0, +\infty]$, we consider the problem where $c(x, y)$ is weighted by the amount of mass from x to y.

$$(KP) \quad \inf \left\{ K(\gamma) := \int_{X \times Y} c(x, y)\, d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}, \tag{1.2}$$

where $\Pi(\mu, \nu)$ is the set of all transport plans $\gamma \in \mathcal{P}(X \times Y)$ from μ to ν, whose marginals on X and Y are μ and ν respectively, i.e.,

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(X \times Y) : (\pi_x)_{\#}\gamma = \mu,\ (\pi_y)_{\#}\gamma = \nu\},$$

where π_x and π_y are the natural projections of $X \times Y$ onto X and Y, respectively.

Note that $\Pi(\mu, \nu)$ is a convex set, and the minimizers for this problem are called *optimal transport plans* between μ and ν. Should γ be of the form $(\mathrm{Id}, T)_{\#}\mu$ for a measurable map $T : X \longrightarrow Y$, the map T would be called the *optimal transport map* from μ to ν.

Indeed,

$$\int_{X \times Y} c\, d\gamma = \int_{X \times Y} c\, d(\mathrm{Id}, T)_{\#}\mu = \int_X c \circ (\mathrm{Id}, T)\, d\mu = \int_X c(x, T(x))\, d\mu(x)$$

$\gamma = (\mathrm{Id}, T)_{\#}\mu$ belongs to $\Pi(\mu, \nu)$ if and only if T pushes μ onto ν $\big($i.e., $T_{\#}\mu(A) = \nu(A) = \mu(T^{-1}(A))$ for any Borel set A $\big)$.

From an intuitive point of view, a minimizing γ describes how the mass of μ is to be associated with the mass of ν to make the overall transport cost minimal.

The generalized problem by Kantorovich is a much better formulation than the original one proposed by Monge because:

- There always exist transport plans in $\Pi(\mu,\nu)$, $\mu\otimes\nu \in \Pi(\mu,\nu) \neq \varnothing$ (for instance, if $\mu=\delta_x$, $\Pi(\mu,\nu)$ contains a unique element, which is $\gamma=\mu\otimes\nu=\delta_x\otimes\nu$.)
- Transport plans $\supset$ transport maps, since $T_{\#}\mu=\nu$ implies $\gamma := (\mathrm{Id},T)_{\#}\mu$ belongs to $\Pi(\mu,\nu)$.
- The set of transport plans is closed with respect to the weak topology.
- $\gamma \mapsto \int c\,d\gamma$ is linear and weakly continuous.
- minima always exist under mild assumptions on c.

In order to prove existence of minimizers of Kantorovich's problem we recall some basic notions concerning analysis over a Polish space.

Definition 1.1.4. On a metric space X, a function $f : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ is said to be lower semi-continuous (l.s.c) if for every sequence $x_n \to x$ we have $f(x) \leq \liminf_n f(x_n)$.

Definition 1.1.5. A metric space X is said to be compact if from any sequence x_n, we can extract a converging subsequence $x_{n_k} \to x \in X$.

Theorem 1.1.6. (Weierstrass). If $f : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ is l.s.c and X is compact, then f attains a minimum on X, i.e., there exists $y \in X$ such that $f(y)=\min\{f(x) : x \in X\}$.

Definition 1.1.7. A sequence $(\mu_n) \subset \mathcal{P}(X)$ is narrowly convergent to $\mu \in \mathcal{P}(X)$, if

$$\lim_{n\to+\infty}\int_X g(x)d\mu_n(x) \longrightarrow \int_X g(x)d\mu(x), \quad \forall\, g \in C_b(X),$$

$C_b(X)$ being the space of continuous and bounded functions on X.

Definition 1.1.8. A sequence $(\mu_n) \subset \mathcal{P}(X)$ is said to be tight if for every $\epsilon>0$, there exists a compact subset $K \subset X$ such that $\mu_n(X\setminus K)<\epsilon$ for every n.

Theorem 1.1.9. (Prokhorov). Suppose that μ_n is a tight sequence of probability measures over a polish space X. Then there exists $\mu \in \mathcal{P}(X)$ and a subsequence μ_{n_k} such that $\mu_{n_k} \rightharpoonup \mu$ (in duality with $C_b(X)$).

Conversely, every sequence $\mu_n \rightharpoonup \mu$ is necessarly tight.

Theorem 1.1.10. Let X and Y be compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c : X\times Y \longrightarrow \mathbb{R}\cup\{+\infty\}$ be l.s.c and bounded from below. Then (KP) attains a minimum.

Theorem 1.1.11. Let X and Y be Polish spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X\times Y \longrightarrow [0,+\infty]$ l.s.c. Then (KP) admits a minimizer.

The proofs of Theorems 1.1.10 and 1.1.11 can be found in [1], Theorems 1.5 and 1.7 respectively.

1.1.2 Duality

The problem (KP) is a linear optimization under convex constraints. Hence, an important tool will be duality theory, for which the problems of this kind admit a natural dual problem (DP), where we maximize a linear functional with affine constraints [2].

Let us express the constraint $\gamma \in \Pi(\mu,\nu)$ in the following way. If $\gamma \in \mathcal{M}_+(X\times Y)$, then we have

$$\sup_{\phi,\psi} \int_X \phi\, d\mu + \int_Y \psi\, d\nu - \int_{X\times Y} (\phi(x)+\psi(y))\, d\gamma = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu,\nu), \\ +\infty & \text{otherwise,} \end{cases}$$

where the supremum is taken among all $(\phi,\psi) \in C_b(X)\times C_b(Y)$.

Indeed, if $\gamma \in \Pi(\mu,\nu)$, then

$$(\pi_x)_\#\gamma = \mu \quad \text{and} \quad (\pi_y)_\#\gamma = \nu,$$

thus

$$\int_X \phi\, d\mu = \int_X \phi\, d(\pi_x)_\#\gamma = \int_{X\times Y} (\phi\circ\pi_x)\, d\gamma = \int_{X\times Y} \phi(x)\, d\gamma(x,y)$$

and

$$\int_Y \psi\, d\nu = \int_Y \psi\, d(\pi_y)_\#\gamma = \int_{X\times Y} (\psi\circ\pi_y)\, d\gamma = \int_{X\times Y} \psi(y)\, d\gamma(x,y)$$

while, if $\gamma \notin \Pi(\mu,\nu)$, we can find $(\phi,\psi) \in C_b(X)\times C_b(Y)$ such that

$$\int_X \phi\, d\mu + \int_Y \psi\, d\nu - \int_{X\times Y} (\phi(x)+\psi(y))\, d\gamma \neq 0,$$

thus by multiplying (ϕ,ψ) by appropriate real numbers we have that the supremum is $+\infty$. Thus, (KP) is equivalent to

$$\min_\gamma \sup_{\phi,\psi} \int_{X\times Y} c\, d\gamma + \int_X \phi\, d\mu + \int_Y \psi\, d\nu - \int_{X\times Y} (\phi(x)+\psi(y))\, d\gamma \tag{1.3}$$

Let

$$G(\gamma,\phi,\psi) = \int_{X\times Y} c\, d\gamma + \int_X \phi\, d\mu + \int_Y \psi\, d\nu - \int_{X\times Y} (\phi(x)+\psi(y))\, d\gamma.$$

As, $\gamma \mapsto G(\gamma,\phi,\psi)$ is convex (actually affine) and $(\phi,\psi) \mapsto G(\gamma,\phi,\psi)$ is concave (actually affine), thus, $\min - \max$ principle holds and we get,

$$\inf_{\gamma\in\Pi(\mu,\nu)} \sup_{\phi,\psi} G(\gamma,\phi,\psi) = \sup_{\phi,\psi} \inf_{\gamma\in\mathcal{M}_+(X\times Y)} G(\gamma,\phi,\psi).$$

Hence, (KP) is equivalent to

$$\sup_{\phi,\psi} \int_X \phi\, d\mu + \int_Y \psi\, d\nu + \inf_\gamma \int_{X\times Y} (c(x,y) - (\phi(x)+\psi(y)))\, d\gamma. \tag{1.4}$$

Remark 1.1.12. If the spaces of ϕ,ψ and γ are compact, one can use another version of min-max principle called **Von Neumann's minimax** in order to prove that (KP) is equivalent to (1.4) .

Theorem 1.1.13. Von Neumann's minimax. Let $X, Y \subset \mathbb{R}^d$ be two compact convex sets. If $f : X\times Y \longrightarrow \mathbb{R}$ is a continuous function that is concave-convex, i.e.,

$$f(.,y) : X \longrightarrow \mathbb{R} \quad \text{is concave for fixed y,}$$

and

$$f(x,.) : Y \longrightarrow \mathbb{R} \quad \text{is convex for fixed x.}$$

Then we have that

$$\max_{x\in X} \min_{y\in Y} f(x,y) = \min_{y\in Y} \max_{x\in X} f(x,y).$$

If we come back to the maximization over (ϕ, ψ), one can rewrite the inf in γ as a constraint on ϕ and ψ :

$$\inf_{\gamma \in \mathcal{M}_+(X\times Y)} \int_{X\times Y} (c - \phi \oplus \psi)\, d\gamma = \begin{cases} 0 & \text{if } \phi \oplus \psi \leq c \text{ on } X \times Y, \\ -\infty & \text{otherwise,} \end{cases}$$

where $\phi \oplus \psi$ denotes the function defined through $(\phi \oplus \psi)(x, y) := \phi(x) + \psi(y)$.

Indeed, If $\phi(x)+\psi(y) \leq c(x, y)$ for any (x, y), then the integrand is non-negative and the inf is 0. Conversely, if $\phi(x) + \psi(y) > c(x, y)$ for some (x, y) in $X \times Y$, then use measures γ concentrated on the set where this strict inequality holds, with mass tending to ∞ i.e., choose $\gamma := n\, \delta_{(x,y)}$ with n large to get the inf is $-\infty$. This leads to the following dual optimization problem.

Problem 1.1.14. Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \longrightarrow [0, +\infty]$, we consider the problem

$$(DP) \quad \max \left\{ \int_X \phi\, d\mu + \int_Y \psi\, d\nu : \phi \in C_b(X), \psi \in C_b(Y) : \phi \oplus \psi \leq c \right\}. \tag{1.5}$$

Let $\gamma \in \Pi(\mu, \nu)$ and observe that for any admissible (ϕ, ψ) we have

$$\int_X \phi\, d\mu + \int_Y \psi\, d\nu = \int_{X\times Y} (\phi \oplus \psi)\, d\gamma \leq \int_{X\times Y} c\, d\gamma.$$

This shows that min(KP) $\geq$ sup(DP).

Now, we will define the notion of c-transform in order to reformulate (DP) as an unconstrained maximization problem.

Definition 1.1.15. (c - transforms). Given a function $\varphi : X \longrightarrow \overline{\mathbb{R}}$, we define its c - transform (also called c - conjugate function) $\varphi^c : Y \longrightarrow \overline{\mathbb{R}}$ by

$$\varphi^c(y) = \inf_{x\in X} c(x, y) - \varphi(x).$$

We also define the $\overline{c}$ - transform of $\xi : Y \longrightarrow \overline{\mathbb{R}}$ by

$$\xi^{\overline{c}}(x) = \inf_{y\in Y} c(x, y) - \xi(y).$$

Moreover, we say that a function $\psi : Y \longrightarrow \overline{\mathbb{R}}$ is $\overline{c}$-concave if there exists $\varphi : X \longrightarrow \overline{\mathbb{R}}$ such that $\psi = \varphi^c$. Similarly, we say that a function $\phi : X \longrightarrow \overline{\mathbb{R}}$ is said to be c-concave if there exists $\xi : Y \longrightarrow \overline{\mathbb{R}}$ such that $\phi = \xi^{\overline{c}}$. We denote by c - conc(X) and $\overline{c}$ - conc(Y) the sets of c - and $\overline{c}$ - concave functions, respectively (when $X = Y$ and c is symmetric, this distinction between c and $\overline{c}$ will play no more role).

Definition 1.1.16. A function $f : (X, d_X) \longrightarrow (Y, d_Y)$ admits γ as modulus of continuity if and only if, $d_Y(f(x), f(y)) \leq \gamma(d_X(x, y)) \quad \forall\, x, y \in X$. For example,

$$\gamma(t) := \sup \{ d_Y(f(x), f(y)) \ : x, y \in X,\ d_X(x, y) \leq t \} \quad \forall\, t \geq 0.$$

is the modulus of continuity.

Proposition 1.1.17. Let $(f_\alpha)_\alpha$ be a family of functions, all satisfying the same condition

$$|f_\alpha(x) - f_\alpha(y)| \leq \gamma(d(x, y)).$$

Consider f defined through $f(x) = \inf_\alpha f_\alpha(x)$. Then f also satisfies the same estimate.

In particular, if the function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies $\lim_{t\to 0} \gamma(t) = 0$ (which means that the family $(f_\alpha)_\alpha$ is equicontinuous), then f continuity as the functions f_α.

In our case, if c is continuous and finite on a compact set, and hence uniformly continuous, this means that there exists an increasing continuous function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ with $\gamma(0) = 0$ such that

$$|c(x_1, y_1) - c(x_2, y_2)| \leq \gamma(d(x_1, x_2) + d(y_1, y_2)).$$

Proposition 1.1.18. Let c be a continuous and finite on a compact set, the ϕ^c shares the same continuity modulus of c.

Proof. Take the definition of φ^c, we have

$$\varphi^c(y) = \inf_x g_x(y)$$

with

$$g_x(y) := c(x, y) - \varphi(x),$$

and

$$c(x, y_1) - c(x, y_2) \leq \gamma(d(x, x) + d(y_1, y_2)),$$

i.e.

$$c(x, y_1) - \varphi(x) - c(x, y_2) + \varphi(x) \leq \gamma(d(y_1, y_2)),$$

which implies

$$g_x(y_1) - g_x(y_2) \leq \gamma(d(y_1, y_2)).$$

Interchanging y_1 and y_2, one obtains

$$|g_x(y_1) - g_x(y_2)| \leq \gamma(d(y_1, y_2)),$$

As $\lim_{t\to 0} \gamma(t) = \gamma(0) = 0$, **this proves that φ^c shares the same continuity modulus of** c. □

Proposition 1.1.19. Suppose that X and Y are compact and c is continuous. Then there exists a solution (ϕ, ψ) to (DP) and it has the form $\phi \in c - conc(X), \psi \in \overline{c} - conc(Y)$ and $\psi = \phi^c$. In particular we can reformulate (DP) as an unconstrained maximization problem:

$$\max(DP) = \max_{\phi \in c-conc(X)} \int_X \phi \, d\mu + \int_Y \phi^c \, d\nu. \tag{1.6}$$

The competitors in (DP) are called Kantorovich potentials in the transport of μ onto ν, and some optimal potentials have the form (ϕ, ϕ^c).

Definition 1.1.20. On a separable metric space X, the support of a measure γ is defined as the smallest closed set on which γ is concentrated, i.e.,

$$\mathrm{spt}(\gamma) := \bigcap \Big\{ A : \text{A is closed and } \gamma(X \setminus A) = 0 \Big\}.$$

Consider an optimal transport plan γ and a Kantorovich potential ϕ and write

$$\phi(x) + \phi^c(y) \leq c(x, y) \text{ on } X \times Y \text{ and } \phi(x) + \phi^c(y) = c(x, y) \text{ on } \mathrm{spt}(\gamma).$$

The equality on $\mathrm{spt}(\gamma)$ is a consequence of the inequality which is valid everywhere and of

$$\int_{X\times Y} c \, d\gamma = \int_X \phi \, d\mu + \int_Y \phi^c \, d\nu = \int_{X\times Y} (\phi(x) + \phi^c(y)) \, d\gamma,$$

which implies equality γ-a.e. These functions being continuous, the equality is satisfied on a closed set, i.e., on the support of the measure γ.

1.1.3 Existence and form of an optimal T

For $\gamma \subset \mathbb{R}^d$, if c is C^1, ϕ is a Kantorovich potential for the cost c and $(x_0, y_0) \in \text{spt}(\gamma)$, then $\nabla\phi(x_0) = \nabla_x c(x_0, y_0)$, provided ϕ is differentiable at x_0. Indeed,

$$\forall\, x, \quad \phi(x) + \phi^c(y_0) \leq c(x, y_0),$$

i.e.,

$$\forall\, x, \quad \phi(x) - c(x, y_0) \leq -\phi^c(y_0),$$

thus

$$x \mapsto \phi(x) - c(x, y_0) \quad \text{is minimal at}\, x = x_0.$$

In the particular case where $c(x_0, y_0) = h(x_0 - y_0)$, with h is strictly convex, **if ϕ and h are differentiable** at x_0 and $x_0 - y_0$, respectively, and $x_0 \notin \partial\Omega$, we get $\nabla\phi(x_0) = \nabla h(x_0 - y_0)$. Thus,

$$x_0 - y_0 = (\nabla h)^{-1}(\nabla\phi(x_0))$$

i.e.,

$$y_0 = x_0 - (\nabla h)^{-1}(\nabla\phi(x_0)).$$

We deduce that, for every x_0, the point y_0 such that $(x_0, y_0) \in \text{spt}(\gamma)$, is unique, i.e., γ is of the form $\gamma_T := (\text{Id}, T)_{\#}\mu$ where $T(x_0) = y_0$. Moreover, this also gives uniqueness of γ and $\nabla\phi$.

Theorem 1.1.21. Given μ and ν probability measures on a compact $\Omega \subset \mathbb{R}^d$, with μ absolutely continuous with respect to the Lebesgue measure, such that $\partial\Omega$ is negligible. There exists a unique optimal transport plan γ of the form $(\text{Id}, T)_{\#}\mu$ for the cost $c(x, y) = h(x - y)$ with h strictly convex. Moreover, there exists a Kantorovich potential ϕ, and T and the potentials ϕ are linked by

$$T(x) = x - (\nabla h)^{-1}(\nabla\phi(x)).$$

If we take $(x_0, y_0) \in \text{spt}(\gamma)$, where $x_0 \notin \partial\Omega$ and $\nabla\phi(x_0)$ exists, then necessarily we have

$$y_0 = x_0 - (\nabla h)^{-1}(\nabla\phi(x_0)).$$

The case where, $X = Y = \mathbb{R}^d$ and $c(x, y) = |x - y|^2$ deserves a special attention. There is a simple characterization and connection of c-concavity with the usual notion of convexity.

Proposition 1.1.22. Let $\phi : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{-\infty\}$. Then ϕ is c-concave if and only if $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$ which is defined by $x \mapsto \Phi(x) := \dfrac{1}{2}|x|^2 - \phi(x)$ is convex and l.s.c.

As a consequence of Proposition (1.1.22), we can particularize Theorem (1.1.21) to the quadratic case , thus getting the existence of an optimal transport map

$$T(x) = x - \nabla\phi(x) = \nabla\left(\frac{|x|^2}{2} - \phi(x)\right) = \nabla u(x)$$

for a convex function u.

Theorem 1.1.23. (Brenier Theorem). Let μ and ν be probabilities over $\mathbb{R}^d$ and $c(x, y) = |x - y|^2$ with $\displaystyle\int |x|^2\, dx, \int |y|^2\, dy < +\infty$, which implies $\min(KP) < +\infty$. Suppose that μ does not give mass to sets of dimension at most $(d-1)$ (in particular if μ is absolutely continuous with respect to the Lebesgue measure). Then, there exists a unique optimal transport map T from μ to ν, and it is of the form $T = \nabla u$ for a convex function u.

1.2 Duality and Optimality conditions

1.2.1 Convexity, Legendre - Fenchel transform and subdifferential

Definition 1.2.1. A function $f : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$ is convex if and only if for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$, we have

$$f((1-t)x + ty) \leq (1-t)f(x) + t\, f(y).$$

If f_α is a family of convex functions, then

$$f(x) := \sup_\alpha f_\alpha(x)$$

is also convex.

Definition 1.2.2. For any given function $\phi : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$, we can define its Fenchel transform

$$\forall\, v \in \mathbb{R}^d, \quad \phi^*(v) := \sup\{x \in \mathbb{R}^d,\ \langle v\,, x\rangle - \phi(x)\}$$

Theorem 1.2.3. A function ϕ is convex and l.s.c if and only if there exists ψ such that $\phi = \psi^*$.

Proposition 1.2.4. A function ϕ is convex and l.s.c if and only if $\phi^{**} = \phi$.

Definition 1.2.5. For every convex function ϕ, we define its subdifferential at x as the set

$$\partial\phi(x) = \{v \in \mathbb{R}^d : \phi(x) + \langle v\,, z - x\rangle\ \leq \phi(z)\ \forall z \in \mathbb{R}^d\}.$$

For the subdifferential of the convex functions ϕ and ϕ^*, we have

$$v \in \partial\phi(x) \iff x \in \partial\phi^*(v) \iff \phi(x) + \phi^*(v) = \langle x\,, v\rangle.$$

This helps in proving ϕ is C^1 if and only if ϕ^* is strictly convex.

Note that, the subdifferential of a convex function satisfies this monotonicity property:

$$\text{If } v_1 \in \partial\phi(x_1) \text{ and } v_2 \in \partial\phi(x_2), \text{then } \langle v_1 - v_2\,, x_1 - x_2\rangle\ \geq 0.$$

Moreover, we define the graph of the subdifferential of a convex function as

$$\text{Graph}(\partial\phi) := \{(x, v) : v \in \partial\phi(x)\} = \{(x, v) : \phi(x) + \phi^*(v) = \langle x\,, v\rangle\},$$

which is monotone, i.e., $(x_i, v_i) \in \text{Graph}(\partial\phi)$ for $i = 1, 2$ implies

$$\langle v_2 - v_1, x_2 - x_1\rangle\ \geq 0$$

1.2.2 c-Cyclical monotonicity and duality

Definition 1.2.6. A set $B \subset \mathbb{R}^d \times \mathbb{R}^d$ is said to be cyclically monotone if for any $N \in \mathbb{N}$, $(x_i, v_i) \in B$ and σ permutation of the set $\{1, ..., N\}$ we have

$$\sum_{i=1}^{N} \langle x_i\,, v_i\rangle\ \geq \sum_{i=1}^{N} \langle x_i\,, v_{\sigma_i}\rangle.$$

The word "cyclical" refers to the fact that, since every σ is the disjoint composition of cycles, it is enough to check this property for cyclical permutations, i.e., replacing

$$\sum_{i=1}^{N} \langle x_i \, , v_{\sigma_i} \rangle$$

with

$$\sum_{i=1}^{N} \langle x_i \, , v_{i+1} \rangle$$

in the definition.

Note that if we take $N = 2$ we get the usual definition of monotonicity, since

$$\langle x_1 \, , v_1 \rangle + \langle x_2 \, , v_2 \rangle \geq \langle x_1 \, , v_2 \rangle + \langle x_2 \, , v_1 \rangle$$

i.e.,

$$\langle v_1 - v_2 \, , x_1 - x_2 \rangle \ \geq 0.$$

Theorem 1.2.7. Rockafellar. Every cyclically monotone set is contained in the graph of the subdifferential of a convex function.

This theorem will be seen as a particular case of Theorem (1.2.9) on a c-concave functions.

Definition 1.2.8. Once a continuous function $c : X \times Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ is given, we say that a set $\gamma \subset X \times Y$ is c-cyclically monotone (briefly c-CM) if $(x_i, v_i) \in \gamma, \ 1 \leq i \leq N$, implies

$$\sum_{i=1}^{N} c(x_i, v_i) \leq \sum_{i=1}^{N} c(x_i, v_{\sigma_i}).$$

Definition 1.2.9. Let $\phi : X \longrightarrow \mathbb{R} \cup \{-\infty\}$ be a c-concave function. The c-superdifferential $\partial^c \phi \subset X \times Y$ is defined as

$$\partial^c \phi := \big\{(x, y) \in X \times Y : \phi(x) + \phi^c(y) = c(x, y)\big\}.$$

The following theorem is a generalization of Theorem 1.2.7 in convex analysis.

Theorem 1.2.10. If $\gamma \neq \emptyset$ is a c-CM set in $X \times Y$ and $c : X \times Y \longrightarrow \mathbb{R}$, then there exists a c-concave function $\phi : X \longrightarrow \mathbb{R} \cup \{-\infty\}$ such that $\gamma \subset \partial^c \phi$.

Theorem 1.2.11. If γ is an optimal transport plan for the cost c and c is continuous, then $\mathrm{spt}(\gamma)$ is a c-CM set.

Corollary 1.2.12. Let ϕ be a maximizer of (1.6) and $\gamma \in \Pi(\mu, \nu)$ a transport plan. Then the two assertions are equivalent

- γ is an optimal transport plan
- $\mathrm{spt}(\gamma) \subset \partial^c \phi$.

Theorem 1.2.13. (Strong duality). If X and Y are Polish spaces and $c : X \times Y \longrightarrow \mathbb{R}$ is uniformly continuous and bounded. Then (DP) admits a solution (ϕ, ϕ^c) and we have

$$\min(KP) = \max(DP)$$

.

A proof of Theorem (1.2.13) can be found in ([1], Theorem 1.39). (Proof [1], Theorem 1.42).

Theorem 1.2.14. If X and Y are Polish spaces and $c : X \times Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ is l.s.c and bounded from below, then the duality formula

$$\min(KP) = \sup(DP)$$

holds.

1.3 Semi-discrete Optimal Transport

In this section, we will focus on the semi-discrete case, which means that the source measure μ is an absolutely continuous distribution that has a bounded probability density function ρ and that the target space is finite such that a target measure ν is finitely atomic, i.e. $\nu = \sum_{i=1}^{N} \nu_i \, \delta_{y_i}$.

For simplicity, in this paragraph we will only consider the quadratic cost in $\mathbb{R}^d$, i.e., $c(x, y) = |x - y|^2$.

1.3.1 Power diagram or Laguerre diagram

Definition 1.3.1. Given a set P of N discrete points in $\mathbb{R}^d$,

$$P = \{y_i \in \mathbb{R}^d \, : i = 1, ..., N\}$$

and a set of weights

$$W = \{w_i \in \mathbb{R} \, : i = 1, ..., N\}.$$

The power diagram or Laguerre diagram of P with respect to W is

$$L_i^w = \big\{x \in \mathbb{R}^d : c(x, y_i) - w_i \leq c(x, y_j) - w_j \quad \forall \, j = 1, ..., N\big\}.$$

We get a map T_w that maps $\mathbb{R}^d$ into set of points P, defined almost everywhere, and its pre-images L_i^w define a partition of $\mathbb{R}^d$ up to a negligible set, such that

$$T_w(x) = y_i \quad \text{if } x \in L_i^w,$$

and weight function $w : P \longrightarrow \mathbb{R}$ such that $w(y_i) = w_i$ (with a slight abuse of notation).

We need to find w such that the μ measure of each power cell corresponds to the ν measure of the associated point, and T_w is an optimal transport map for the cost c between μ and the measure ν if and only if $\nu = T_{w\#}\mu$ if and only if $\mu(L_i^w) = \nu_i$, i.e., $\nu = \sum_{i=1}^{N} \mu(L_i^w) \, \delta_{y_i}$.

The dual problem (DP) amounts to maximizing the Kantorovitch functional given by

$$\mathcal{K}(w) = \int_{\mathbb{R}^d} w^c \, d\mu + \int_{\mathbb{R}^d} w \, d\nu = \sum_{i=1}^{N} \int_{L_i^w} \big(c(x, y_i) - w_i\big) \, d\mu(x) + \int_{\mathbb{R}^d} w \, d\nu \tag{1.7}$$

Look at the terms in the Laguerre cells and define

$$\begin{aligned} f(w, T) &= \int \big(c(x, T(x)) - w(T(x)\big) \, d\mu(x) \\ &= \int c(x, T(x)) \, d\mu(x) - \sum_{i=1}^{N} \int_{T(x)=y_i} w_i \, d\mu(x) \\ &= \int c(x, T(x)) \, d\mu(x) - \sum_{i=1}^{N} w_i \, \mu(T^{-1}(y_i)) \end{aligned} \tag{1.8}$$

If we fix w : what map T minimizes (1.8) ? In Definition 1.3.1 we have defined a map T_w such that

$$c(x, T_w(x)) - w(T_w(x)) \leq c(x, y_i) - w_i \quad \forall \, i,$$

i.e.,

$$\begin{aligned} g(w) &= \min_T f(w,T) \\ &= f(w,T_w) \\ &= \int \big(c(x,T_w(x))\, d\mu(x) - \sum_{i=1}^{N} w_i\, \mu(L_i^w) \end{aligned}$$

Each $f(w,T)$ is affine in w, thus g is concave. Besides, let $h(w) = g(w) + \sum_{i=1}^{N} \nu_i\, w_i$ then h is concave, admitting that h is differentiable and that

$$\frac{\partial h}{\partial w_i} = -\mu(L_i^w) + \nu_i.$$

Thus, w being a maximizer of h if and only if

$$\nabla h(w) = 0$$

if and only if

$$\mu(L_i^w) = \nu_i \quad \text{for every } i = 1, ..., N,$$

if and only if

$$T_{w\#}\mu = \nu \,(\text{feasibility condition}) \text{ and } T_w \text{ is optimal.}$$

Let us now prove the optimality of T_w in the OT problem between μ and $T_{w\#}\mu$.

Proposition 1.3.2. Given an absolutely continuous measure μ and a set of weights W and set of points P, we can define the target measure $T_{w\#}\mu$, then T_w is optimal transport map for the cost c between μ and the measure $T_{w\#}\mu$.

Proof. Let us consider another measurable map S such that $S_\#\mu = T_{w\#}\mu$ and show that the cost attained by T_w is $\leq$ the cost of S. By definition of T_w , T_w was defined to be a minimizer of this object thus, one has

$$c(x,T_w(x)) - w(T_w(x)) \leq c(x,S(x)) - w(S(x)).$$

Integrating the above inequality gives

$$\int_X (c(x,T_w(x)) - w(T_w(x)))\, d\mu(x) \leq \int_X (c(x,S(x)) - w(S(x)))\, d\mu(x),$$

applying change of variable formulas we get

$$\int_X w(T_w(x)))\, d\mu(x) = \int_Y w(y)\, d\, T_{w\#}\mu(y) = \int_Y w(y)\, d\, S_\#\mu(y) = \int_X w(S(x))\, d\mu(x).$$

Subtracting this equality from the inequality above shows that the map T_w is optimal in the optimal transport:

$$\int_X c(x,T_w(x)) d\mu(x) \leq \int_X c(x,S(x)) d\mu(x).$$

□

We conclude, maximizing the concave function

$$\begin{aligned} h(w) &= \int_{\mathbb{R}^d} c(x, T_w(x))\, d\mu(x) + \sum_{i=1}^{N} (\nu_i - \mu(L_i^w)) w_i \\ &= \int_{\mathbb{R}^d} w^c(x)\, d\mu(x) + \sum_{i=1}^{N} w_i\, \nu_i. \end{aligned}$$

is equivalent to doing semi-discrete OT. The maximization of h is a finite-dimensional issue, since (thanks to ν being finitely atomic) the only values of w which are relevant are those taken at the points y_i. Hence, we can consider $w \in \mathbb{R}^d$, and look at the derivatives of h.

We recall that the Kantorovich formulation is given by

$$(KP) \qquad \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\gamma(x,y),$$

and

$$(KP) = \max_{w \in \mathbb{R}^d} \mathcal{K}(w), \tag{1.9}$$

where $\mathcal{K}$ is the Kantorovich functional [3] given by

$$\begin{aligned} \int_{\mathbb{R}^d} w^c\, d\mu + \int_{\mathbb{R}^d} w\, d\nu &= \sum_{i=1}^{N} \int_{L_i^w} \big(c(x, y_i) - w_i\big)\, d\mu(x) + \sum_{i=1}^{N} w_i\, \nu_i \\ &= \int_{\mathbb{R}^d} w^c(x)\, d\mu(x) + \sum_{i=1}^{N} w_i\, \nu_i \\ &= h(w). \end{aligned}$$

Theorem 1.3.3. Assume that μ is absolutely continuous with respect to the Lebesgue measure with density function ρ and ν is a discrete measure with finite support and consider the Kantorovitch functional $\mathcal{K}$ defined by

$$\mathcal{K}(w) = \int_{\mathbb{R}^d} w^c\, d\mu + \int_{\mathbb{R}^d} w\, d\nu,$$

Then the semi-discrete optimal transport problem (KP) admits solutions of the form T_w, where w solves the concave optimization problem (1.9), i.e.,

- $\mathcal{K}$ is concave and C^1-smooth and its gradient is

$$\frac{\partial\, \mathcal{K}}{\partial\, w_i} = \nu_i - \mu(L_i^w).$$

- $\mathcal{K}$ attains its maximum over $\mathbb{R}^d$ if and only if $\nu_i = \mu(L_i^w)\, \forall\, i$, i.e., $T_{w\#}\mu = \nu$.

Theorem 1.3.3 provides the solution of (1.1) which is uniquely defined for almost every x even if there exists several different w defining the same assignment T_w.

Chapter 2

Gradient descent algorithms for convex optimization.

In this chapter [18] we are going to mention the basic definitions and properties related to L-smoothness, convexity, strong convexity and proofs of the convergence of gradient and stochastic gradient descent type method in order to solve the OT problem presented in chapter 1.

2.1 Convexity, smoothness and strong convexity.

2.1.1 Convexity

Modes of convexity can be equivalently characterized in terms of the monotonicity properties of the gradient mapping $\nabla f : \mathbb{R}^d \longrightarrow \mathbb{R}^d$.

Definition 2.1.1. We say that a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is convex if

$$f(tx + (1-t)y) \leq t\, f(x) + (1-t) f(y), \quad \forall x, y \in \mathbb{R}^d,\, t \in [0,1]. \tag{2.1}$$

Proposition 2.1.2. Assume that f is differentiable. Then f is convex if and only if every tangent line to the graph of f lower bounds the function values, that is

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x \rangle, \quad \forall x, y \in \mathbb{R}^d. \tag{2.2}$$

Proof. Assume f is convex, divide (2.1) by t and re-arrange to get

$$\frac{f(y + t(x-y)) - f(y)}{t} \leq f(x) - f(y).$$

Now taking the limit $t \to 0$ gives

$$\langle \nabla f(y),\, x - y \rangle \leq f(x) - f(y).$$

Assume now (2.2) holds for all $x, y \in \mathbb{R}^d$. Take an arbitrary $u, v \in \mathbb{R}^d$ and $t \in [0,1]$. Let

$$x := t\, u + (1-t)\, v.$$

Then (2.2) implies that

$$f(u) \geq f(x) + \langle \nabla f(x),\, u - x \rangle, \tag{2.3}$$
$$f(v) \geq f(x) + \langle \nabla f(x),\, v - x \rangle. \tag{2.4}$$

Note that

$$u - x = (1-t)(u-v)$$

and

$$v - x = t(v-u).$$

Thus, if we multiply (2.3) by t and (2.4) by $1-t$, and then add the two inequalities, we obtain

$$\begin{aligned} t\,f(u) + (1-t)\,f(v) &\geq f(x) + \langle t\,\nabla f(x),\,(1-t)(u-v)\rangle + \langle (1-t)\,\nabla f(x),\,t(v-u)\rangle \\ &= f(t\,u + (1-t)\,v). \end{aligned}$$

□

Remark 2.1.3. Assume that $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is twice differentiable, we know that the directional derivative of g in the direction of v on the point x is defined as

$$D_v\, g(x) = \nabla g(x).v = \langle \nabla g(x),\, v\rangle,$$

then the directional derivative in the v direction on the point x of $\langle \nabla f(x),\, y-x\rangle$ is as follows

$$\begin{aligned} \big\langle \nabla \langle \nabla f(x),\, y-x\rangle,\, v\big\rangle &= \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \langle \nabla f(x),\, y-x\rangle\, v_i \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \Big(\sum_{j=1}^{n} \frac{\partial}{\partial x_j} f(x)\,(y_j - x_j)\Big)\, v_i \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x)\,(y_j - x_j)\,v_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial}{\partial x_j} f(x)\,\frac{\partial}{\partial x_i}\,(y_j - x_j)\,v_i \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x)\,v_i\,(y_j - x_j) - \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f(x)\,v_i \\ &= \big\langle \nabla^2 f(x)\,v,\, y-x\big\rangle - \big\langle \nabla f(x),\, v\big\rangle \end{aligned} \tag{2.5}$$

Proposition 2.1.4. If $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is twice differentiable and convex, modes of convexity can be equivalently characterized in terms of the definiteness of the Hessians $\nabla^2 f(x),\ \forall\, x \in \mathbb{R}^d$.

f is convex if and only if

$$\nabla^2 f(x) \geq 0,\ \forall\, x \in \mathbb{R}^d.$$

Proof. Assume f is convex and let

$$g(y) = f(x) + \langle \nabla f(x),\, y-x\rangle - f(y), \tag{2.6}$$

we know from (2.2) that $g(y) \leq 0$ and as $g(x) = 0$, taking the directional derivative of (2.6) in the direction of v in (2.2) on the point x to get

$$\big\langle \nabla f(x),\, v\big\rangle + \big\langle \nabla\langle \nabla f(x),\, y-x\rangle,\, v\big\rangle - 0 \leq 0$$

using (2.5) in the last inequality gives

$$\big\langle \nabla f(x),\, v\big\rangle + \big\langle \nabla^2 f(x)\,v,\, y-x\big\rangle - \big\langle \nabla f(x),\, v\big\rangle \leq 0$$

therefore as

$$\langle \nabla^2 f(x)\,v,\, y-x\rangle = \langle \nabla f(x)\,,\, v\rangle + \langle \nabla^2 f(x)\,v,\, y-x\rangle - \langle \nabla f(x)\,, v\rangle, \tag{2.7}$$

we get

$$\langle \nabla^2 f(x)\, v,\, y - x \rangle \leq 0, \quad \forall\, x, y, v \in \mathbb{R}^d. \tag{2.8}$$

Setting $y = x - v$ then gives

$$\langle \nabla^2 f(x)\, v,\, v \rangle \geq 0, \quad \forall\, x, v \in \mathbb{R}^d. \tag{2.9}$$

Now assume that the Hessian $\nabla^2 f(x)$ of f is positive semi-definite for all $x \in \mathbb{R}^d$. Then Taylor's theorem implies that

$$f(y) = f(x) + \langle \nabla f(x),\, y - x \rangle + \frac{1}{2} \langle y - x,\, \nabla^2 f(x)(y - x) \rangle + o(\|y - x\|^2.)$$

Since $\nabla^2 f(x)$ is everywhere positive semi-definite, we get

$$\langle y - x,\, \nabla^2 f(x)(y - x) \rangle \geq 0$$

thus

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x \rangle$$

□

2.1.2 Smoothness

Theorem 2.1.5. Mean Value Theorem for multivariate functions.
Let $U \subset \mathbb{R}^d$ be open, $g : U \longrightarrow \mathbb{R}^m$ is continuously differentiable and $x \in U$, $h \in \mathbb{R}^d$ vectors such that the line segment $x + t\,h$, $0 \leq t \leq 1$ remains in U. Then we have

$$g(x + h) = g(x) + \big(\int_0^1 Dg(x + t\,h)\, dt \big).h,$$

where Dg denotes the Jacobian matrix of g.

Definition 2.1.6. Let f be a differentiable function. f is said to be L-smooth if its gradients are Lipschitz continuous, that is

$$\forall\, x, y, \quad \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \tag{2.10}$$

Proposition 2.1.7. If f is L-smooth and twice differentiable, then

$$f(y) \leq f(x) + \langle \nabla f(x),\, y - x \rangle + \frac{L}{2} \|y - x\|_2^2.$$

Proof. Applying Theorem 2.1.5 with $g = \nabla f$, we obtain

$$\nabla f(x + \alpha\, d) = \nabla f(x) + \int_0^\alpha \nabla^2 f(x + td)\, d\, dt$$

Followed by taking the norm gives

$$\big\| \int_0^\alpha \nabla^2 f(x + t\, d)\, d \big\|_2 \leq L\, \alpha\, \big\| d \big\|_2.$$

Dividing by α

$$\frac{\big\| \int_0^\alpha \nabla^2 f(x + t\, d)\, d \big\|_2}{\alpha} \leq L\, \big\| d \big\|_2$$

then dividing through by $\|d\|_2$ with $d \neq 0$ and taking the limit as $\alpha \to 0$ we have that

$$\frac{\left\| \int_0^\alpha \nabla^2 f(x + t\, d)\, d \right\|_2}{\alpha \, \|d\|_2} = \frac{\left\| \alpha \, \nabla^2 f(x)\, d \right\|_2}{\alpha \, \|d\|_2} + \mathcal{O}(\alpha) \overset{\alpha \to 0}{=} \frac{\left\| \nabla^2 f(x)\, d \right\|_2}{d} \leq L, \quad \forall\, d \neq 0 \in \mathbb{R}^d.$$

Taking the supremum over $d \neq 0 \in \mathbb{R}^d$ in the above gives

$$\nabla^2 f(x) \leq L\, I.$$

Furthermore, using the Taylor expansion of $f(x)$ and the uniform bound over Hessian we have that

$$f(y) \leq f(x) + \langle \nabla f(x),\, y - x \rangle + \frac{L}{2} \|y - x\|_2^2. \tag{2.11}$$

□

Some direct consequences of the smoothness are given in the following lemma.

Lemma 2.1.8. If f is L-smooth with x^* is the global minimizer then

$$f\big(x - \frac{1}{L} \nabla f(x)\big) - f(x) \leq -\frac{1}{2L} \left\| \nabla f(x) \right\|_2^2, \tag{2.12}$$

and

$$f(x^*) - f(x) \leq -\frac{1}{2L} \left\| \nabla f(x) \right\|_2^2, \tag{2.13}$$

hold for all $x \in \mathbb{R}^d$.

Proof. The first inequality (2.12) follows by inserting $y = x - \frac{1}{L} \nabla f(x)$ in the definition of smoothness (2.10) since

$$\begin{aligned} f\big(x - \frac{1}{L} \nabla f(x)\big) &\leq f(x) - \frac{1}{L} \langle \nabla f(x),\, \nabla f(x) \rangle + \frac{L}{2} \left\| \frac{1}{L} \nabla f(x) \right\|_2^2 \\ &= f(x) - \frac{1}{2L} \left\| \nabla f(x) \right\|_2^2. \end{aligned}$$

Furthermore, by using (2.12) combined with $f(x^*) \leq f(y)\ \forall\, y$, we get (2.13). Indeed since

$$f(x^*) - f(x) \leq f\big(x - \frac{1}{L} \nabla f(x)\big) - f(x) \leq -\frac{1}{2L} \left\| \nabla f(x) \right\|_2^2. \tag{2.14}$$

□

2.1.3 Smooth and Convex

There are many problems in optimization where the function is both smooth and convex. Furthermore, such a combination results in some interesting consequences and Lemmas that we will then use to prove convergence of the Gradient method.

Lemma 2.1.9. If f is convex and L-smooth then

$$f(y) - f(x) \leq \langle \nabla f(y),\, y - x \rangle \leq -\frac{1}{2L} \left\| \nabla f(y) - \nabla f(x) \right\|_2^2. \tag{2.15}$$

$$\langle \nabla f(y) - \nabla f(x),\, y - x \rangle \geq \frac{1}{L} \left\| \nabla f(x) - \nabla f(y) \right\|_2^2. \tag{2.16}$$

Proof. To prove (2.15), it follows that

$$f(y) - f(x) = f(y) - f(z) + f(z) - f(x),$$

using (2.2) and (2.11) we get

$$f(y) - f(x) \leq \langle \nabla f(y),\, y - z\rangle + \langle \nabla f(x),\, z - x\rangle + \frac{L}{2}\|z - x\|_2^2. \tag{2.17}$$

Minimizing in z we have that

$$z = x - \frac{1}{L}\Big(\nabla f(x) - \nabla f(y)\Big). \tag{2.18}$$

Substituting (2.18) in (2.17) gives

$$\begin{aligned} f(y) - f(x) &\leq \Big\langle \nabla f(y),\, y - x + \frac{1}{L}\Big(\nabla f(x) - \nabla f(y)\Big)\Big\rangle \\ &\quad - \frac{1}{L}\Big\langle \nabla f(x),\, \nabla f(x) - \nabla f(y)\Big\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\ &\leq \Big\langle \nabla f(y),\, y - x\Big\rangle - \frac{1}{L}\Big\|\nabla f(x) - \nabla f(y)\Big\|_2^2 + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \qquad (2.19) \\ &\leq \Big\langle \nabla f(y),\, y - x\Big\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \qquad (2.20) \end{aligned}$$

Finally (2.16) follows from applying (2.15) once

$$f(y) - f(x) \leq \Big\langle \nabla f(y),\, y - x\Big\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2, \tag{2.21}$$

then interchanging the roles of x and y to get

$$f(x) - f(y) \leq \Big\langle \nabla f(x),\, x - y\Big\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2. \tag{2.22}$$

Finally adding (2.21) and (2.22) gives

$$0 \leq \Big\langle \nabla f(y) - \nabla f(x),\, y - x\Big\rangle - \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|_2^2. \tag{2.23}$$

□

2.1.4 Strong convexity

We can strengthen the notion of convexity by defining μ-strong convexity.

Definition 2.1.10. We say that a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is μ-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2 \tag{2.24}$$

Minimizing both sides of (2.24) in y proves the following lemma

Lemma 2.1.11. If f is μ-strongly convex then it also satisfies the **Polyak-Lojasiewicz (PL)** condition, that is

$$\|\nabla f(x)\|_2^2 \geq 2\,\mu\,(f(x) - f(x^*)), \tag{2.25}$$

where x^* is the global minimizer.

Proof. Multiplying (2.24) by minus and substituting $y = x^*$ we have that

$$\begin{aligned} f(x) - f(x^*) &\leq \langle \nabla f(x),\, x - x^* \rangle - \frac{\mu}{2}\|x^* - x\|_2^2 \\ &= -\frac{1}{2}\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\|_2^2 + \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \\ &\leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2. \end{aligned}$$

□

Proposition 2.1.12. If the function f is twice continuously differentiable, then it is μ-strongly convex if and only if

$$\nabla^2 f(x) \succcurlyeq \mu\, I,$$

for all x in the domain, where I is the identity and $\nabla^2 f$ is the Hessian matrix, and the inequality $\succcurlyeq$ means that $\nabla^2 f(x) - \mu\, I$ is positive semi-definite. This equivalent to requiring that the minimum eigenvalue of $\nabla^2 f(x)$ be at least μ for all x.

Remark 2.1.13. A twice differentiable function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex if and only if there exists $m > 0$ and $M > 0$ such that:

$$mI_n \preccurlyeq \nabla^2 f(x) \preccurlyeq MI_n, x \in \mathbb{R}^n.$$

Lemma 2.1.14. A function f is μ-strongly convex if and only if

$$x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$$

is convex.

Proposition 2.1.15. Let A be a symmetric matrix, we say that A is positive semi-definite if and only if all the eigenvalues of A are non-negative, and that A is positive definite if and only if all the eigenvalues of A are positive.

Proof. **Without loss of generality, we may assume that $\|\mathbf{x}\|^2 = \mathbf{x^T x} = \mathbf{1}$.**

Suppose that A is positive semi-definite, let λ be an eigenvalue of A then, there exists a non-zero eigenvector x such that $Ax = \lambda x$. We have

$$0 \;\leq\; x^T Ax = x^T \lambda x = \lambda x^T x = \lambda.$$

In the other hand, as A is symmetric and has non-negative eigenvalues, then, by the spectral theorem for symmetric matrices, there exists an orthogonal matrix Q such that

$$A = Q^T DQ$$

where

$$D = diag(\lambda_1, \ldots, \lambda_n).$$

Q is invertible such that $Q^{-1} = Q^T$ and preserves the norm, i.e.,

$$\|Qx\| = \|x\|, \quad \forall\, x \in \mathbb{R}^n.$$

Indeed,

$$\|Qx\|^2 = (Qx)^T(Qx) = x^T Q^T Qx = x^T x = \|x\|^2 = 1,$$

this implies for any y such that $\|y\| = 1$, there exists a unique x with $\|x\| = 1$ and $Qx = y \neq 0$. Hence,

$$x^T Ax = x^T(Q^T DQ)x = x^T Q^T DQx = y^T Dy = \sum_{i=1}^{n} \lambda_i y_i^2 \leq \lambda_{\max} \sum_{i=1}^{n} y_i^2 = \lambda_{\max}\, \|y\|^2 = \lambda_{\max} > 0.$$

□

2.1.5 Application

Proposition 2.1.16. Consider $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ defined by

$$f(X) = \frac{1}{2}X^T AX - X^T b, \quad \forall b, X \in \mathbb{R}^n$$

such that f is twice differentiable at X. Then for all $h \in \mathbb{R}^n$ we have

$$\nabla f(X) = AX - b = \frac{1}{2}(A^T + A)X - b \quad \text{and} \quad H_f(X) = \nabla^2 f(X) = \frac{1}{2}(A + A^T).$$

In particular, if A is symmetric, $H_f(X) = A$

Proof. We have,

$$f(X + h) - f(X) = \nabla f(X)^T(h) + \frac{1}{2}(h)^T \nabla^2 f(X)(h) + o(\|h\|^2) \tag{2.26}$$

in the other hand,

$$\begin{aligned} f(X + h) - f(X) &= \frac{1}{2}(X + h)^T A(X + h) - (X + h)^T b - \frac{1}{2}X^T AX + X^T b \\ &= \frac{1}{2}X^T AX + \frac{1}{2}X^T Ah + \frac{1}{2}h^T AX + \frac{1}{2}h^T Ah - X^T b - h^T b - \frac{1}{2}X^T AX + X^T b \\ &= \frac{1}{2}X^T Ah + \frac{1}{2}h^T AX - h^T b + \frac{1}{2}h^T Ah \end{aligned} \tag{2.27}$$

comparing (2.26) and (2.27), we get

$$\nabla^2 f(X) = A$$

and

$$\begin{aligned} \nabla f(X)^T(h) &= \frac{1}{2}X^T Ah + \frac{1}{2}h^T AX - h^T b \\ &= \frac{1}{2}X^T Ah + \frac{1}{2}X^T A^T h - b^T h \\ &= \big(\frac{1}{2}(A^T + A)X - b\big)^T(h) \end{aligned}$$

therefore,

$$\nabla f(X) = \frac{1}{2}(A^T + A)X - b = AX - b$$

□

1. Let

$$f(X) = \frac{1}{2}X^T AX - X^T b \tag{2.28}$$

which is equivalent to

$$f(X) = f(x, y, z) = 2x^2 - 2xy + y^2$$

where

$$X = \begin{pmatrix} x \\ y \end{pmatrix}, \quad A = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Using Proposition (2.1.16) we see that $H_f(x)$ is constant equal to A. The eigenvalues of A can be computed by solving the equation

$$\det(A - \lambda\, I_2) = 0,$$

which is equivalent to

$$(4 - \lambda)(2 - \lambda) = 4 \tag{2.29}$$

(2.29) has two solutions $3 - \sqrt{5} > 0$ and $3 + \sqrt{5} > 0$. This implies that f is strongly convex for $m = 3 - \sqrt{5} > 0$ and $M = 3 + \sqrt{5} > 0$.

2. Let

$$f(X) = \frac{1}{2}X^T AX - X^T b \tag{2.30}$$

which is equivalent to

$$f(X) = f(x, y, z) = 2x^2 + 4y^2 + 2z^2 + xy + 5yz + 2xz - 2x - 3y - 4z$$

where

$$X = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad A = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 8 & 5 \\ 2 & 5 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$$

$$\nabla f(X) = \frac{1}{2}(A^T + A)\, X - b = AX - b$$

and

$$\nabla^2 f(X) = A.$$

To prove that f is strictly convex, it is equivalent to prove that the symmetric matrix A is positive definite, i.e.,

$$X^T AX > 0, \quad \forall\, X \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

i.e., the eigenvalues of A are positive. The eigenvalues of A can be computed by solving the equation

$$\det(A - \lambda I_3) = 0.$$

This equation has three positive solutions $\lambda_1 \approx 11.87$, $\lambda_2 \approx 3.86$ and $\lambda_3 \approx 0.26$. We deduce that f is strictly convex.

3. Let

$$f(X) = \frac{1}{2}X^T AX - X^T b \tag{2.31}$$

which is equivalent to

$$f(X) = f(x, y, z) = \frac{1}{2}\big(2(x - y)^2 + 4z^2\big) - x - 2y$$

where

$$X = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad A = \begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$$

f is convex and not strictly convex, because the eigenvalues of the non-invertible matrix A are $\lambda_1 = 0$, $\lambda_2 = 4$ and $\lambda_3 = 4$, that means A is positive semi-definite. We can also see that f is convex because,

$$X^T AX = 2(x - y)^2 + 4z^2 \geq 0, \quad \forall\, X \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

2.2 Gradient Descent

Gradient descent is an optimization algorithm to find the minimum of a function. We start with a random point on the function and move in the negative direction of the gradient of the function to reach the local or global minima.

Consider the problem

$$x^* = \arg\min_{x\in\mathbb{R}^d} f(x), \tag{2.32}$$

and the following gradient method

$$x^{t+1} = x^t - \alpha\,\nabla f(x^t), \tag{2.33}$$

where f is L-smooth.

We will now prove that the iterates (2.33) converge. In Theorem 2.2.1 we will prove sublinear convergence under the assumption that f is convex. In Theorem 2.2.2 we will prove linear convergence (a stronger form of convergence) under the assumption that f is μ-strongly convex.

2.2.1 Convergence for convex and smooth functions

Theorem 2.2.1. Let f be a convex and L-smooth and let x^t for $t = 1, ..., n$ be the sequence of iterates generated by the gradient method (2.33). It follows that

$$f(x^n) - f(x^*) \leq \frac{2L\,\|x^1 - x^*\|_2^2}{n-1} \tag{2.34}$$

Proof. Let f be a convex and L-smooth. It follows that

$$\begin{aligned}\|x^{t+1} - x^*\|_2^2 &= \|x^t - x^* - \frac{1}{L}\nabla f(x^t)\|_2^2 \\ &= \|x^t - x^*\|_2^2 - \frac{2}{L}\langle x^t - x^*,\, \nabla f(x^t)\rangle + \frac{1}{L^2}\|\nabla f(x^t)\|_2^2 \\ &\overset{(2.16)}{\leq} \|x^t - x^*\|_2^2 - \frac{1}{L^2}\|\nabla f(x^t)\|_2^2.\end{aligned} \tag{2.35}$$

Thus $\|x^t - x^*\|_2^2$ is a decreasing sequence in t. Calling upon (2.12) and subtracting $f(x^*)$ from both sides gives

$$f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - \frac{1}{2L}\|\nabla f(x^t)\|_2^2 \tag{2.36}$$

Applying convexity we have that

$$\begin{aligned}f(x^t) - f(x^*) &\leq \langle \nabla f(x^t),\, x^t - x^*\rangle \\ &\leq \|\nabla f(x^t)\|_2\,\|x^t - x^*\|_2 \\ &\overset{(2.35)}{\leq} \|\nabla f(x^t)\|_2\,\|x^1 - x^*\|_2\end{aligned} \tag{2.37}$$

Isolating $\|\nabla f(x^t)\|_2$ in the above and inserting in (2.36) gives

$$f(x^{t+1}) - f(x^*) \overset{(2.36)+(2.37)}{\leq} f(x^t) - f(x^*) - \frac{1}{2L}\frac{1}{\|x^1 - x^*\|^2}\left(f(x^t) - f(x^*)\right)^2 \tag{2.38}$$

Let $\delta_t = f(x^t) - f(x^*)$ and $\beta = \dfrac{1}{2L}\dfrac{1}{\|x^1 - x^*\|^2}$.

Manipulating (2.38) we have that

$$\delta_{t+1} \leq \delta_t - \beta\, \delta_t^2$$

multiplying by $\dfrac{1}{\delta_t\, \delta_{t+1}}$ equivalently yields

$$\beta\, \frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}$$

as $\delta_{t+1} \leq \delta_t$ we equivalently get

$$\beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}.$$

Summing up both sides over $t = 1, ..., n-1$ and using telescopic cancellation we have that

$$(n-1)\beta \leq \frac{1}{\delta_n} - \frac{1}{\delta_1} \leq \frac{1}{\delta_n}.$$

□

2.2.2 Convergence for strongly convex and smooth convex functions

Now we prove some bounds that hold for strongly convex and smooth functions. In fact, if you observe, we will only use PL inequality (2.25) to establish the convergence result. Assuming a function satisfies the PL condition is a strictly weaker assumption than assuming strong convexity.

Theorem 2.2.2. Let f be L-smooth and μ-strongly convex function. From a given $x^0 \in \mathbb{R}^d$ and $\dfrac{1}{L} \geq \alpha > 0$, the iterates

$$x^{t+1} = x^t - \alpha\, \nabla f(x^t), \tag{2.39}$$

converge according to

$$\|x^{t+1} - x^*\|_2^2 \leq (1 - \alpha\, \mu)^{t+1}\, \|x^0 - x^*\|_2^2. \tag{2.40}$$

Proof. From (2.33) we have that

$$\begin{aligned}
\|x^{t+1} - x^*\|_2^2 &= \|x^t - x^* - \alpha\, \nabla f(x^t)\|_2^2 \\
&= \|x^t - x^*\|_2^2 - 2\alpha\, \langle x^t - x^*,\, \nabla f(x^t)\rangle + \alpha^2\, \|\nabla f(x^t)\|_2^2 \\
&\overset{(2.24)}{\leq} (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha\, (f(x^t) - f(x^*)) + \alpha^2\, \|\nabla f(x^t)\|_2^2 \\
&\overset{(2.13)}{\leq} (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha\, (f(x^t) - f(x^*)) + 2\alpha^2\, L(f(x^t) - f(x^*)) \\
&= (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha\, (1 - \alpha\, L)\, (f(x^t) - f(x^*)).
\end{aligned} \tag{2.41}$$

Since

$$\frac{1}{L} \geq \alpha$$

we have that

$$-2\alpha\, (1 - \alpha\, L) < 0,$$

and thus can be safely dropped to give

$$\|x^{t+1} - x^*\|_2^2 \leq (1 - \alpha\mu)\|x^t - x^*\|_2^2$$

It now remains to unroll the recurrence. □

Remark 2.2.3. Once we know the strong convexity constants m and M mentioned in (2.1.13) of a given function, we can use that to compute **the convergence rate of the gradient descent algorithm**. We know that from (2.2.2) that

$$f(x^{(k)}) - f(x^*) \leq \big(1 - \frac{m}{M}\big)^k \left(f(x^0) - f(x^*)\right)$$

where x^* is the global minimizer

$$f(x^{(k)}) \geq f(x^*).$$

The term $1 - \dfrac{m}{M}$ is the amount by which the error of current solution shrinks at every iteration of the gradient descent algorithm. When $\dfrac{m}{M}$ is very close to zero, the convergence will be slow, that means when the condition $1 - \dfrac{m}{M}$ is very close to one the convergence is very slow. For instance, when $1 - \dfrac{m}{M} = 0$ we get

$$f(x^{(k)}) - f(x^*) \leq 0$$

which means that we have reached the global minima that is the convergence is so fast, it is even instantaneous.

2.2.3 Application

In this application we present three different examples and we will study the rate of convergence of our algorithm in order to achieve the minima when it exists. In each example we will have three cases to observe the consequences of the number of iterations and learning rate (to know how much we have to move, it controls the speed at which the weights are updated to reach the minimum point of loss function) on the speed of convergence. **The algorithm starts at** $X = (0, 0, 0)^T$ then, finds the gradient of the function and moves in the direction of the negative of the gradient. Let us assume the **learning rate** to be $10^{-1}, 10^{-2}$ **and** 10^{-3} **in each individual case of the three examples respectively**. We will observe the decreasing of the values of the function which should converge to the local/ global minima (for convex problem when the function is convex, GD will converge to the global minima). However, how many iterations should we perform? Let us set the **maximum number of iterations** $= 10^4$. Also, let us set a precision variable in our algorithm which calculates the difference between two consecutive values. Run a loop to perform gradient descent, if the difference between the values from two consecutive iterations is less than the **precision** $= 10^{-6}$ or the number of iterations exceeds 10^4 we set, **stop the algorithm, so this tells us when to stop the algorithm.**

Consider the function

$$f(X) = \frac{1}{2}X^T AX - X^T b$$

where A is symmetric matrix.

1. Let us consider the function

   $$f(X) = \frac{1}{2}X^T AX - X^T b$$

   with

   $$A = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 8 & 5 \\ 2 & 5 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}.$$

   We proved that f is strictly convex by calculating the eigenvalues of the matrix A. Set, $m = \lambda_3 \approx 0.26$ the smallest eigenvalue of A and $M = \lambda_1 \approx 11.87$ the largest eigenvalue of A.

We have

$$\nabla f(X) = AX - b = \frac{1}{2}(A^T + A)X - b.$$

In order to minimize f we have to solve

$$AX = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 8 & 5 \\ 2 & 5 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = b,$$

which is equivalent to

$$\nabla f(X) = 0$$

i.e.,

$$\begin{aligned} f_x &= 4x + y + 2z - 2 = 0 \\ f_y &= x + 8y + 5z - 3 = 0 \\ f_z &= 2x + 5y + 4z - 4 = 0 \end{aligned}$$

thus the function attains its minimum at

$$X = A^{-1}b = \begin{pmatrix} -1 \\ -2 \\ 4 \end{pmatrix},$$

where the global minima is

$$f(-1, -2, 4) = -4.$$

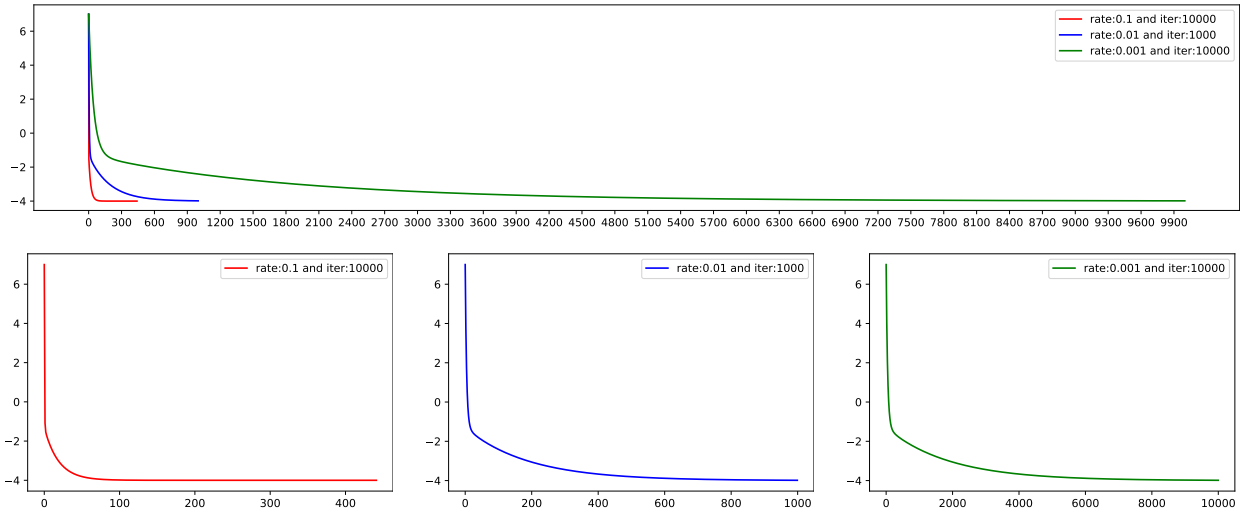

Figure 2.1: In the following figure we have the values of f at each iteration of the gradient descent algorithm with different learning rate. We can observe that the values is slowly decreasing and should converge to -4 (the global minima). In the red graph, the learning rate is 0.1 the algorithm starts at 0 and runs for 442 iterations before it terminates, while in the blue graph the algorithm runs for 10^3 iterations before it terminates. The global mimimum occurs approximately at $(-0.99, -1.99, 3.99)$. The convergence is very slow because the condition number $\frac{m}{M} = \frac{0.26}{11.86} \approx 0.022$ which is close to 0.

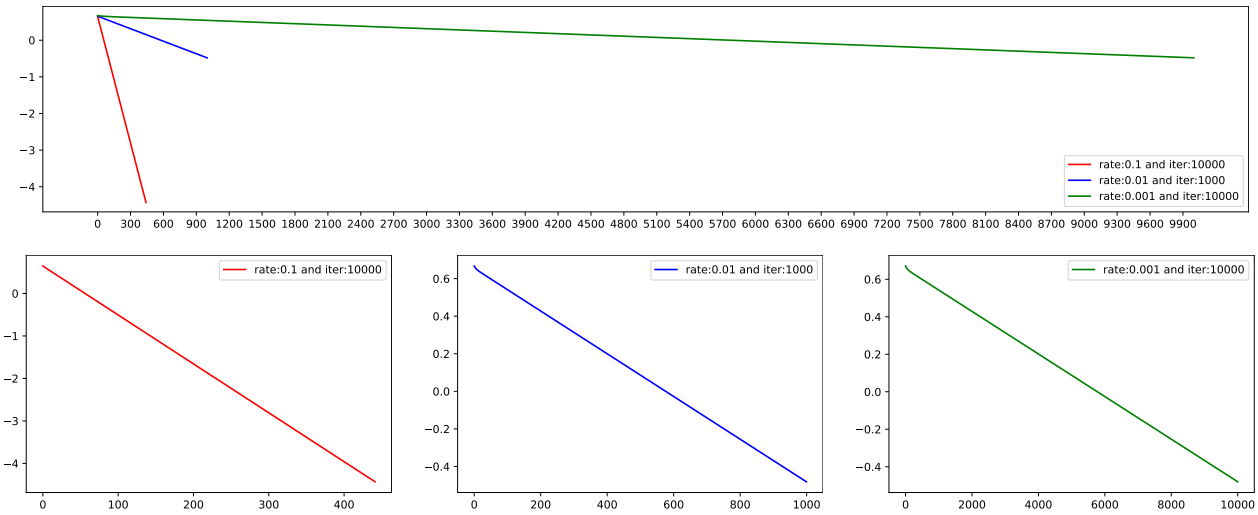

Figure 2.2: In the following figure we have to see the speed of convergence on the **logarithm** $\log_{10}$. The convergence is first order. The local minima is $x^* = (-1, -2, 4)$, we notice that in the red graph the gradient descent converges exponentially fast to the minimum, in which $\|x^k - x^*\| \approx 10^{-6}$, i.e., $x^k \to x^*$, while as the step size decreases we see the impact on the rate of convergence. Indeed, the slope is different, but the behaviour is the same (linear convergence in log scale). The rate of convergence is similar, what changes is the better choice of the learning rate value in the first plot.

2. In this case we will consider the same strictly convex function f with

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I_3 \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}.$$

The three eigenvalues are 1 and the function attains its minimum at $X = (2, 3, 4)$ where the global minima is $f(2, 3, 4) = -14.5$.

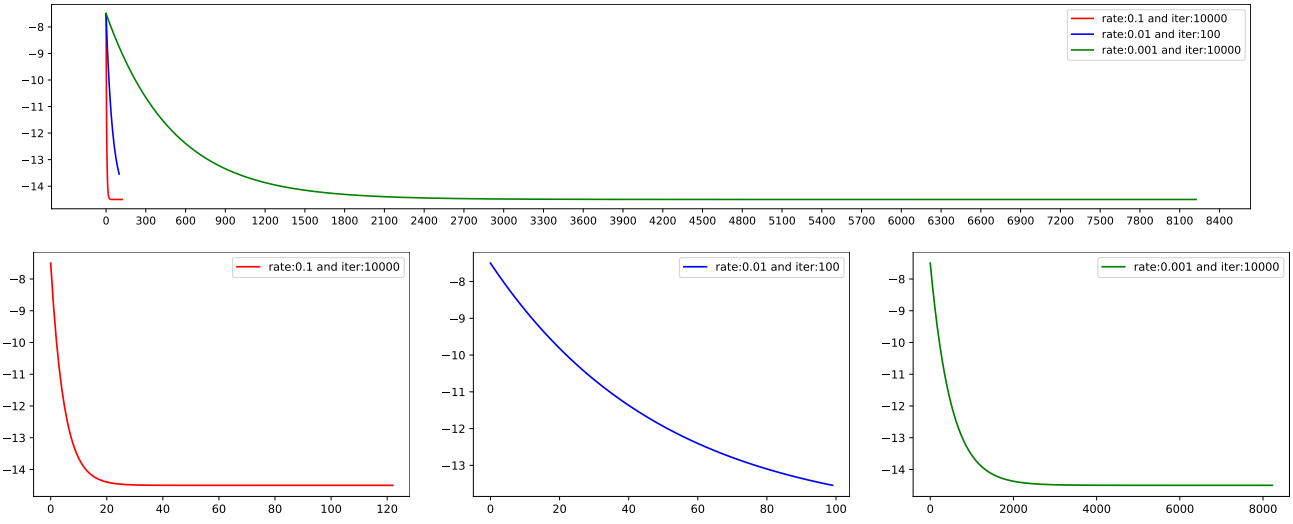

Figure 2.3: In the following figure we have the values of f at each iteration of the gradient descent algorithm with different learning rate and different number of iterations. We can observe that the values are fastly decreasing and should converge to -14.5 (the global minima). In the red graph, the learning rate is 0.1 the algorithm starts at 0 and runs for 123 iterations before it terminates, while in the blue graph the algorithm runs for 100 iterations before termination. In the green graph the algorithm runs for 8225 iterations before it terminates and it reaches the global minima even when the learning rate is 0.001. The convergence is very fast because the condition number is $\frac{m}{M} = \frac{1}{1} = 1$.

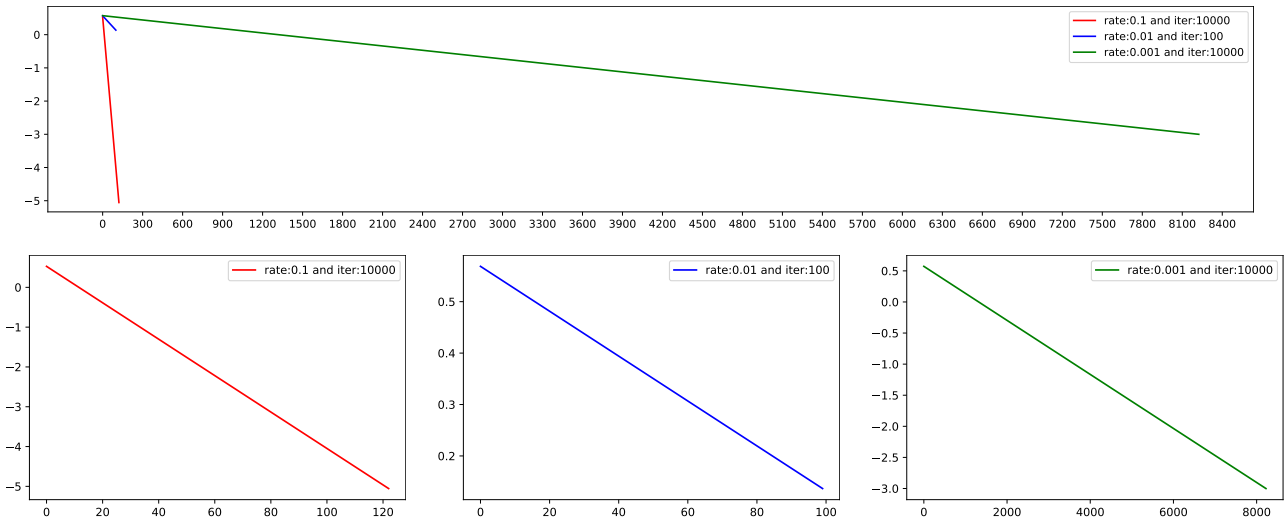

Figure 2.4: In this case the local/global minima is $x^* = (2, 3, 4)$, we notice that in the red graph the gradient descent converges exponentially faster to the minimum compared to the previous case, in which $\|x^k - x^*\| \approx 10^{-5}$, after 120 iterations i.e., $x^k \to x^*$, while in the blue graph as the step size and number of iterations decreases simultaneously where the number of iterations is really small 100, we realize x^k can't reach x^* compared to the green graph where $\|x^k - x^*\|$ becomes so close to zero. We realize that the behaviour is same and the rate of convergence is similar but it about the better choice of the learning rate and number of iterations.

3. In this case we consider f with

$$A = \begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}.$$

f is convex and not strictly convex because the eigenvalues are $\lambda_1 = 0$, $\lambda_2 = 4$ and $\lambda_3 = 4$. Since A is non-invertible matrix, there is no minimum here as $A^{-1}b$ can not be attained and the gradient descent goes to $-\infty$. Indeed, $\nabla f(X) = 0$ is equivalent to

$$f_x = 2x - 2y - 1 = 0, \quad f_y = -2x + 2y - 2 = 0 \quad \text{and} \quad f_z = 4z = 0$$

that gives

$$z = 0, \quad x - y = \frac{1}{2} \quad \text{and} \quad x - y = -1$$

i.e.,

$$f(x, x, 0) = -3x \to -\infty, \quad \text{as} \quad x \to \infty.$$

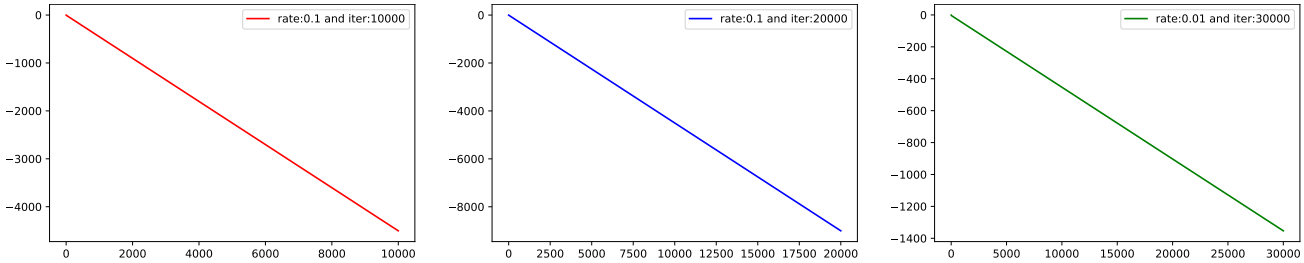

Figure 2.5: No convergence can be expected as $\frac{m}{M}$ is 0 and the minimum doesn't attained by f. we observe in the first graph we reach -4000 for 10^4 iterations and as we increase the number of iterations in the second graph to 2×10^4 the gradient descent (GD) reaches -8000 that means the gradient descent goes to $-\infty$. As well as, we observe in the third graph as we decrease the learning rate to 0.01 the GD needs more iterations, even if we increase the iteration we notice that GD reaches -1400.

2.3 Stochastic Gradient Descent For Convex and Strongly Convex Optimization Problems

SGD is an iterative method to optimize the loss function. It is a stochastic (system or a process that is linked with a random probability) approximation of GD optimization, since it replaces the actual gradient (calculated from the entire data set) by an estimate thereof (calculated from a randomly selected subset of the data where few samples are selected randomly instead of the whole data set for each iteration). In addition, SGD is a popular algorithm for training a wide range of models in machine learning (most importantly forms the basis of Neural Network), including (linear) support vector machines, logistic regression and graphical models.

Consider the problem

$$w^* = \arg\min_{w\in\mathbb{R}^d} \mathbf{E}_{\mathcal{D}}\big[F(w,x)\big] := f(w), \tag{2.42}$$

where $x \sim \mathcal{D}$. Here we will consider the stochastic subgradient method. Let $F(w,x)$ be convex in w, and consider the iterates

$$w^{t+1} = w^t - \alpha_t\, G(w^t, x^t), \tag{2.43}$$

where $G(w^t, x^t) \in \partial_w F(w^t, x^t)$ and

$$\mathbf{E}[G(w^t, x^t)\,|\,w^t] = g(w^t) \in \partial f(w), \tag{2.44}$$

and where $x^t \sim \mathcal{D}$ is independent and identically distributed sampled at each iteration.

2.3.1 Convex

Theorem 2.3.1. Let $r, B > 0$. Let $f(w,x)$ be convex in w. Assume that $\mathbf{E}_{\mathcal{D}}[\|G(w^t,x^t)\|_2^2] \le B^2$ where $B > 0$ for all t, and that the inputs and optimal point are in the ball $\|w\|_2 \le r$. From a given $w^0 \in \mathbb{R}^d$ and $\alpha_t = \dfrac{C}{B\sqrt{2t}}$, consider the iterates of the stochastic subgradient method

$$w^{t+1} = w^t - \alpha_t\, G(w^t, x^t), \tag{2.45}$$

The iterates satisfy

$$\mathbf{E}[f(\overline{x}_T)] - f(w^*) \le \frac{3B}{C\sqrt{T}}. \tag{2.46}$$

Proof.

$$\begin{aligned}\|w^{t+1} - w^*\|_2^2 &\overset{(2.45)}{=} \|w^t - w^* - \alpha_t\, G(w^t,\, x^t)\|_2^2 \\ &= \|w^t - w^*\|_2^2 - 2\alpha_t\,\langle G(w^t,\, x^t),\, w^t - w^*\rangle + \alpha_t^2\,\|G(w^t,\, x^t)\|_2^2\end{aligned} \tag{2.47}$$

Taking expectation conditioned on w^t we have that

$$\begin{aligned}\mathbf{E}\big[\|w^{t+1} - w^*\|_2^2\,|\,w^t\big] &\overset{(2.44)}{=} \|w^t - w^*\|_2^2 - 2\alpha_t\,\langle g(w^t),\, w^t - w^*\rangle + \alpha_t^2\,\mathbf{E}_{\mathcal{D}}[\|G(w^t,x^t)\|_2^2] \\ &\le \|w^t - w^*\|_2^2 - 2\alpha_t\,\langle g(w^t),\, w^t - w^*\rangle + \alpha_t^2\,B^2 \\ &\overset{(2.2)}{\le} \|w^t - w^*\|_2^2 - 2\alpha_t\,(f(w^t) - f(w^*)) + \alpha_t^2\,B^2\end{aligned} \tag{2.48}$$

Taking expectation and re-arranging we have that

$$\mathbf{E}\big[f(w^t)\big] - f(w^*) \le \frac{1}{2\alpha_t}\,\mathbf{E}\big[\|w^t - w^*\|_2^2\big] - \frac{1}{2\alpha_t}\,\mathbf{E}\big[\|w^{t+1} - w^*\|_2^2\big] + \frac{\alpha_t\,B^2}{2}. \tag{2.49}$$

Summing up from $t = 1, \dots, T$ and using that α_t is a non-increasing sequence, we have that

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{E}\big[f(w^t)\big] - f(w^*) &\leq \frac{1}{2\alpha_1}\,\|w^t - w^*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T-1}\big(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\big)\,\mathbf{E}\big[\|w^{t+1} - w^*\|_2^2\big] \\ &\quad - \frac{1}{2\alpha_{T+1}}\,\mathbf{E}\big[\|w^{T+1} - w^*\|_2^2\big] + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t \\ &\overset{\|w\|_2\leq r}{\leq} \frac{2}{\alpha_1}\,r^2 + \sum_{t=1}^{T-1}\big(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\big)\,2r^2 + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t \\ &\overset{\text{telescopic}}{=} \frac{2r^2}{\alpha_T} + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t. \end{aligned} \tag{2.50}$$

Finally let $\overline{x}_T = \dfrac{1}{T}\displaystyle\sum_{t=0}^{T-1} x_t$ and dividing by T and using Jensen's inequality we have

$$\mathbf{E}[f(\overline{x}_T)] - f(w^*) \overset{\text{Jensen's}}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\big[f(w^t)\big] - f(w^*) \tag{2.51}$$

$$\overset{(2.50)}{\leq} \frac{2r^2}{T\,\alpha_T} + \frac{B^2}{2T}\sum_{t=1}^{T}\alpha_t. \tag{2.52}$$

Now plugging in $\alpha_t = \dfrac{\alpha_0}{\sqrt{t}}$ we have that

$$\mathbf{E}[f(\overline{x}_T)] - f(w^*) \leq \frac{2r^2}{\alpha_0\sqrt{T}} + \frac{\alpha_0\,B^2}{2T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \frac{1}{\sqrt{T}}\big(\frac{2r^2}{\alpha_0} + \alpha_0\,B^2\big). \tag{2.53}$$

Minimizing the right-hand side in α_0 gives $\alpha_0 = \dfrac{r\sqrt{2}}{B}$ and the result since

$$\frac{2r^2}{\alpha_0} + \alpha_0\,B^2 = 2rB\sqrt{2} \leq 3rB. \tag{2.54}$$

□

2.3.2 Strongly convex

- **Constant stepsize**

Theorem 2.3.2. Let $F(w, x)$ be a μ-stongly convex function in w. From a given $w^0 \in \mathbb{R}^d$ and $\dfrac{1}{\mu} > \alpha \equiv \alpha_t > 0$, consider the iterates of the stochastic subgradient method (2.45). Assume that $\mathbf{E}_{\mathcal{D}}[\|G(w^t, x^t)\|_2^2] \leq B^2$ where $B > 0$ for all t. The iterates satisfy

$$\mathbf{E}\big[\|w^{t+1} - w^*\|_2^2\big] \leq (1 - \alpha\,\mu)^{t+1}\|w^0 - w^*\|_2^2 + \frac{\alpha}{\mu}B^2. \tag{2.55}$$

Proof. From (2.45) we have

$$\begin{aligned}\|w^{t+1} - w^*\|_2^2 &= \|w^t - w^* - \alpha\,G(w^t,\, x^t)\|_2^2 \\ &= \|w^t - w^*\|_2^2 - 2\alpha\,\langle G(w^t,\, x^t),\, w^t - w^*\rangle + \alpha^2\,\|G(w^t,\, x^t)\|_2^2 \end{aligned} \tag{2.56}$$

Taking expectation condition on w^t in the above gives

$$\begin{aligned}\mathbf{E}\big[\|w^{t+1}-w^*\|_2^2\,|\,w^t\big] &= \|w^t-w^*\|_2^2 - 2\alpha\,\langle g(w^t),\,w^t-w^*\rangle + \alpha^2\,\mathbf{E}_{\mathcal{D}}[\|G(w^t,x^t)\|_2^2]\\ &\le \|w^t-w^*\|_2^2 - 2\alpha\,\langle g(w^t),\,w^t-w^*\rangle + \alpha^2\,B^2\\ &\overset{(2.24)}{\le} (1-\alpha\,\mu)\|w^t-w^*\|_2^2 + \alpha^2\,B^2\\ &\overset{\text{recurrence}}{\le} (1-\alpha\,\mu)^{t+1}\|w^0-w^*\|_2^2 + \sum_{i=0}^{t}(1-\alpha\,\mu)^i\,\alpha^2\,B^2. \end{aligned} \tag{2.57}$$

Since

$$\sum_{i=0}^{t}(1-\alpha\,\mu)^i\,\alpha^2\,B^2 = \alpha^2\,B^2\,\frac{1-(1-\alpha\,\mu)^{t+1}}{\alpha\,\mu} \le \frac{\alpha^2\,B^2}{\mu}, \tag{2.58}$$

we have that

$$\mathbf{E}\big[\|w^{t+1}-w^*\|_2^2\,|\,w^t\big] \overset{(2.57)+(2.58)}{\le} (1-\alpha\,\mu)^{t+1}\|w^0-w^*\|_2^2 + \frac{\alpha\,B^2}{\mu}. \tag{2.59}$$

It now remains to taking expectation over the above. □

Remark 2.3.3. For any optimization problem, GD converges to a local minimizer if the learning rate is less than $\frac{1}{\mu}$ where μ is the Lipschitz smoothness of the loss function with respect to the parameters. The size of steps affects the performance of the model, small learning rates consume a lot of time to converge and large learning rates puts the model at risk of overshooting the minima so it will not be able to converge. Hence, our goal is to tune the learning rate so that the GD optimizer reaches the minima in the fewest number of steps.

2.3.3 Application: Gradient Noise

In this application we are going to do some experiments on SGD on the function f by adding a Gaussian Noise to every gradient with Mean Value of zero and certain Standard Deviation Value (SDV), we will add some Noise to the gradient and see the impact on the distance to the global minimizer w^* and on the convergent of the gradient descent. We consider (2.45) adding time-dependent Gaussian noise to the gradient at every training step t:

$$w^{t+1} = w^t - \alpha_t(\,\nabla f(w^t) + N(0,\sigma_t^2)) \tag{2.60}$$

where

$$\sigma_t^2 = \frac{\eta}{(1+t)^\gamma} \quad \text{and} \quad \alpha_t = \frac{0.1}{\sqrt{t}}$$

with η selected from $\{0.01, 0.3, 1.0\}$ and $\gamma = 0.55$.

Consider the function

$$\mathbf{f(X)} = \frac{\mathbf{1}}{\mathbf{2}}\mathbf{X^T A X} - \mathbf{X^T b}$$

Indeed, in this application we will consider the function f with two different matrices A_1 and $A_2 = I_3$. The first one where the speed of convergence is very slow as the condition number is close to 0 and the second one where the speed of convergence is very fast as the condition number is close to 1. We will present different cases in order to realize the impact of the constant gradient Noise, time-dependent gradient Noise, time-dependent rate and constant rate on the values of f and distance to w^* by changing η every time.

In each case, the first diagram will represent the function values $f(w^t)$ where the second diagram represents the distance to the $\arg\min(\|w^t-w^*\|)$ in $\log_{10}$-scale.

Let

$$A_1 = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 8 & 5 \\ 2 & 5 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}.$$

1. **We start by time-dependent rate without gradient Gaussian Noise:** $w^{t+1} = w^t - \alpha_t(\nabla f(w^t))$ with $\alpha_t = \dfrac{0.1}{\sqrt{t}}$

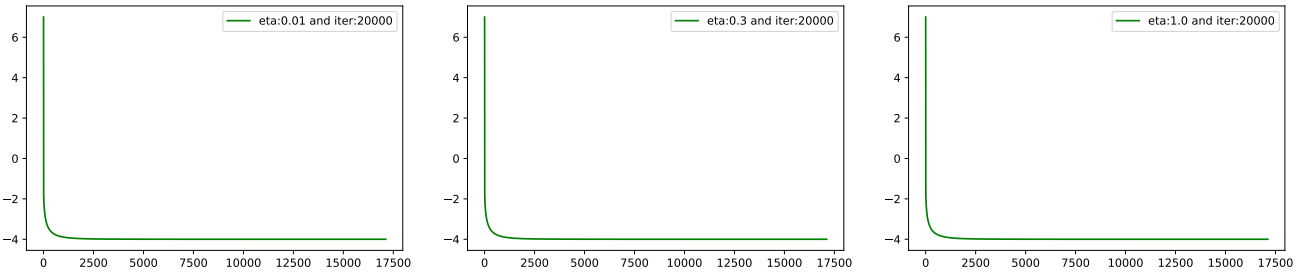

Figure 2.6: The diagram represents the function values $f(w^t)$. η has no impact on the values of f. As the condition number is close to 1 we see that the convergence is very slow, the minimum is attained in less than 2000 iterations as α_t decreases.

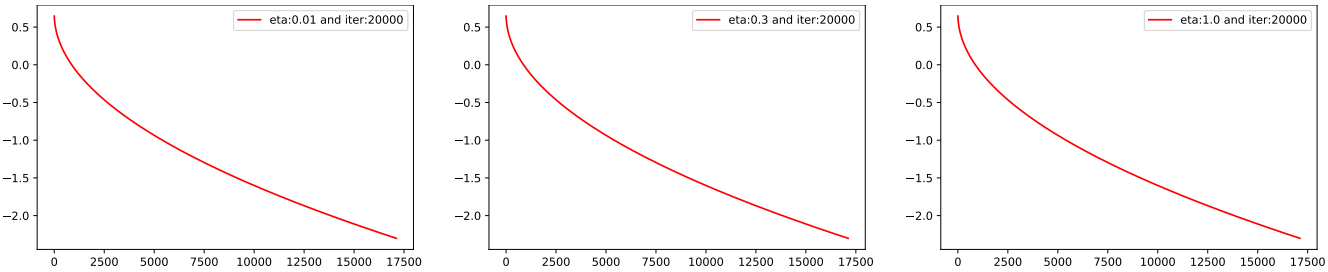

Figure 2.7: The diagram represents the distance to the $\arg\min(\|w^t - w^*\|)$. η does not has any impact on the speed of convergence as there is no Noise depending on it. At the beginning (up to 4000 iterations) the algorithm converges slowly to the minimum where $\|w^t - w^*\| \approx 10^{-0.7}$. Afterwards, as the step size decreases the speed of convergence becomes better as $\|w^t - w^*\| \approx 10^{-2.5}$.

2. **We add time-independent Gaussian Noise:** $w^{t+1} = w^t - \alpha_t(\nabla f(w^t) + N(0, \sigma_t^2))$ with $\alpha_t = \dfrac{0.1}{\sqrt{t}}$ and $\sigma_t^2 = 0.5$

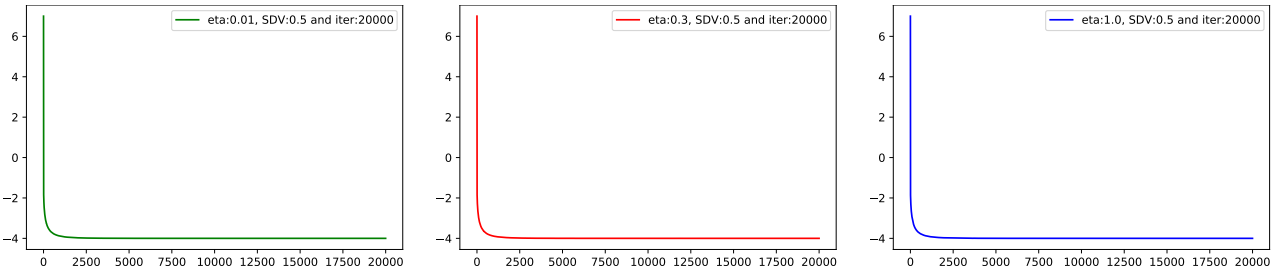

Figure 2.8: Here we have the same graphs as the previous case as the learning rate and the Noise are independent of η. We will see later, when we increase σ_t^2 and the speed of convergence, we will have a small perturbations on the values of f.

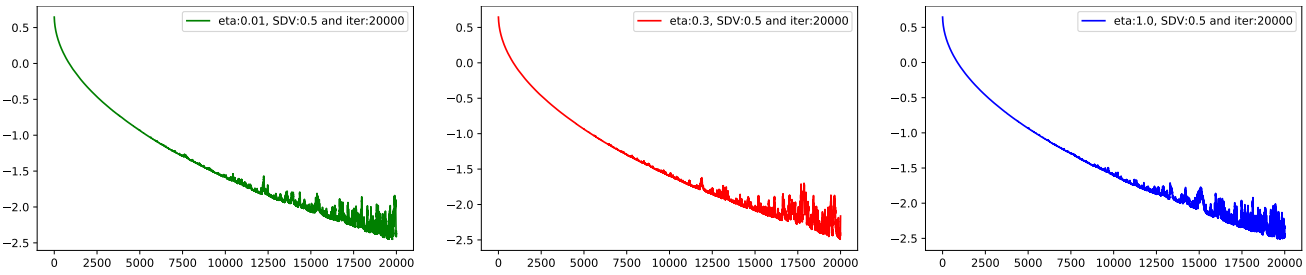

Figure 2.9: At the beginning (up to 8000 iterations) we observe that there is no Noise. Afterwards, we realize that as the step size decreases the Noise increases slowly. Note that, if we slightly increase SDV (σ_t^2) we will have more vibrations with higher amplitude the same if we decrease it to 0.1 we will get less vibrations and this is what we will see in the next graphs.

3. **In this case we have time-dependent Gaussian Noise and learning rate such that:** $\sigma_t^2 = \frac{\eta}{(1+t)^\gamma}$ and $\alpha_t = \frac{0.1}{\sqrt{t}}$

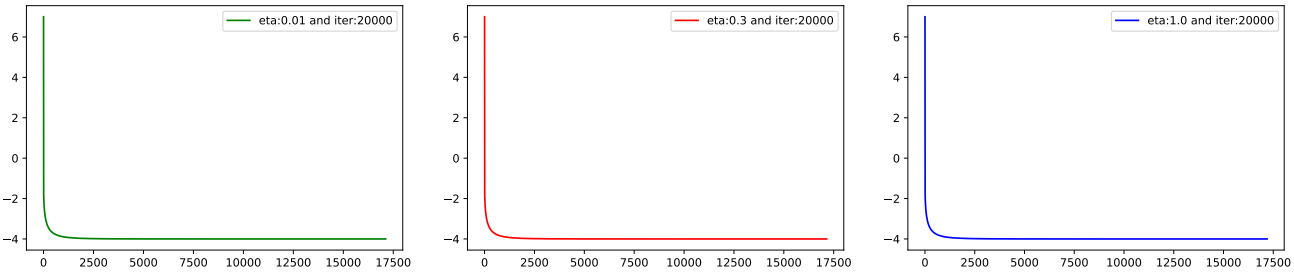

Figure 2.10: Similarly, as the learning rate set as a decreasing function along the iterations we will have the same as the previous cases, but regarding the Noise we know that σ_t^2 decreases and vanishes along the iterations. In order to see an impact on the values of f we need to increase or fix σ_t^2 and not only to increase the convergence.

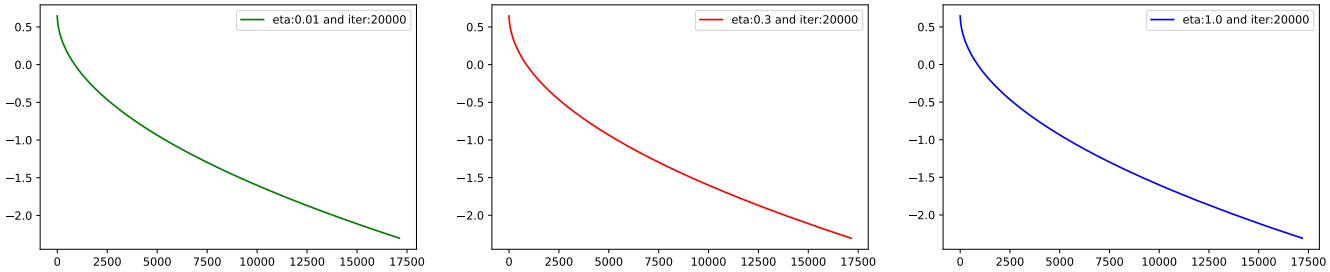

Figure 2.11: The vibrations disappear, because the step size and the STD of the gradient noise decrease (that is what I mentioned about decreasing the STD as the step size decreases). If we increase the speed of convergence we can see the impact of η with small vibrations too.

4. **In this case we have constant rate with time-independent Gaussian Noise such that:** $\alpha_t = 0.1$ and $\sigma_t^2 = 0.5$.

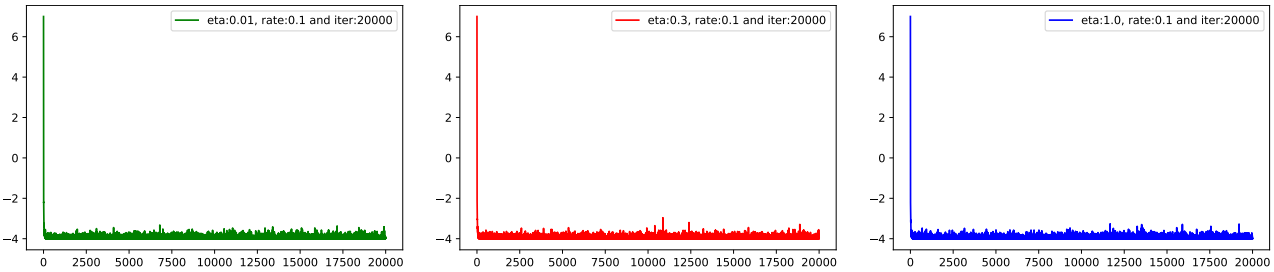

Figure 2.12: As the rate is constant and f is independent of η, f behaves the same in all cases with small perturbations.

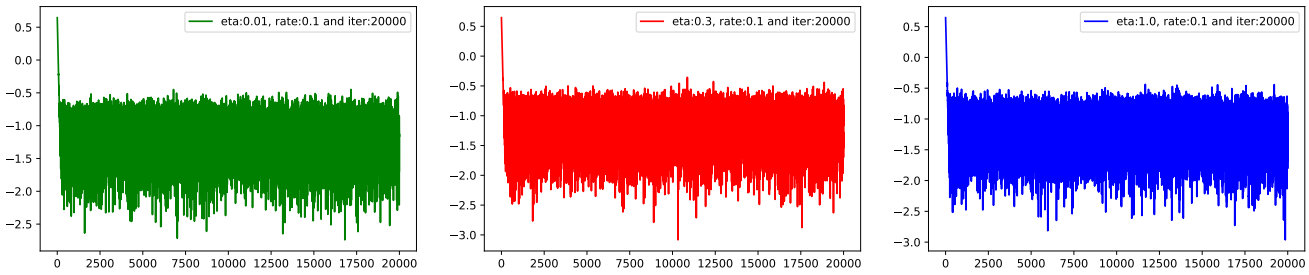

Figure 2.13: As the rate is constant, we can see better the impact of the Noise. Up to 500 iterations the speed of convergence is not so fast as $\|w^t - w^*\| \approx 10^{-1}$. Afterwards, the oscillations start and as the Noise is constant, the perturbations are the same.

5. **In this case we have constant rate with time-dependent Gaussian Noise such that:** $\sigma_t^2 = \dfrac{\eta}{(1+t)^\gamma}$ and $\alpha_t = 0.1$

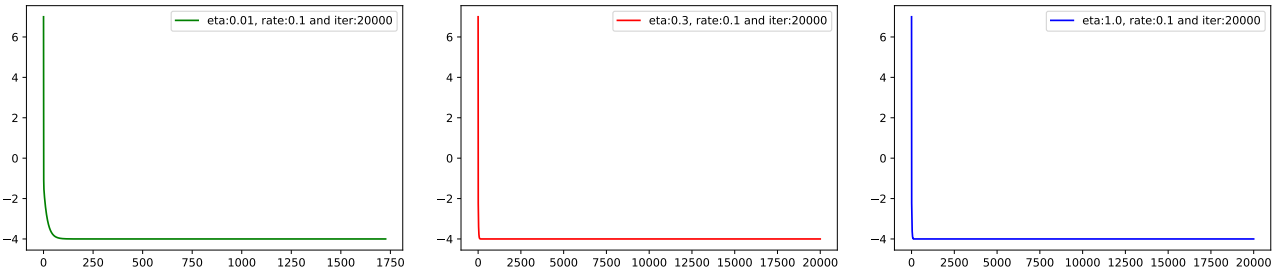

Figure 2.14: Here we realize the impact of η. The minimum is attained faster in all the graphs, but in the first graph when η is small the algorithm runs for 1700 iterations, thus it is clear to see how the function behaves at the beginning where the convergence (before 100 iterations) does not attain the minimum. As η increases the algorithm runs for 20000 iterations before it terminates.

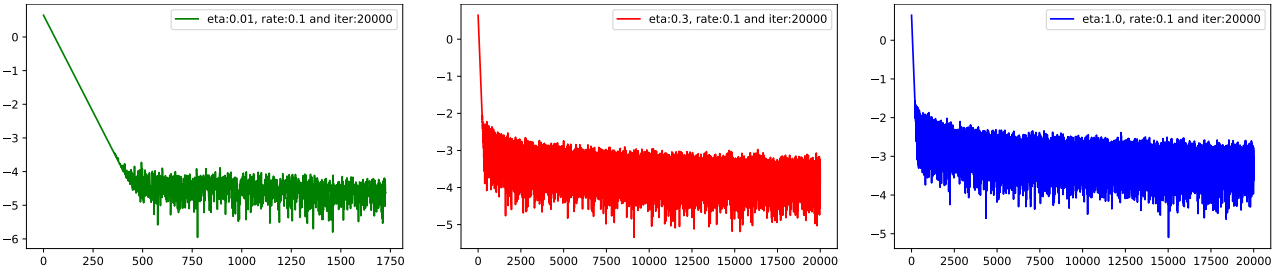

Figure 2.15: As the rate is constant, we can see better the impact of the Noise. We observe that up to 350-600 iterations the speed of convergence is fast as $\|w^t - w^*\| \approx 10^{-3}$ in all the graphs. Afterwards, the oscillations start, but in the 1st graph (Green) we observe that the algorithm runs for just 1700 iterations and it is clear to see the convergence at the beginning. As η increase the algorithm runs for 20000 iterations and complete with same vibrations in the red and blue graphs.

Remark 2.3.4. In the deterministic method, according to Figure 2.2 we realize that when there is no Noise on the gradient and when we use a decreasing learning rate, the algorithm converge more slowly as $\|w^t - w^*\| \approx 10^{-0.6}$ compared to the stochastic method in Figure 2.7 as $\|w^t - w^*\| \approx 10^{-2.5}$.

Here we consider the second case where:

$$\mathbf{f(X)} = \frac{\mathbf{1}}{\mathbf{2}}\mathbf{X^T A X} - \mathbf{X^T b}$$

with

$$A_2 = I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}.$$

Here the speed of convergence is very fast as the condition number is 1. We will analyze and see some figures in order to observe the difference compared to the previous case when the convergence was very slow with matrix A_1.

1. **We have time-dependent rate with time-independent Gaussian Noise:** $\alpha_t = \dfrac{0.1}{\sqrt{t}}$ and $\sigma_t^2 = 0.5$

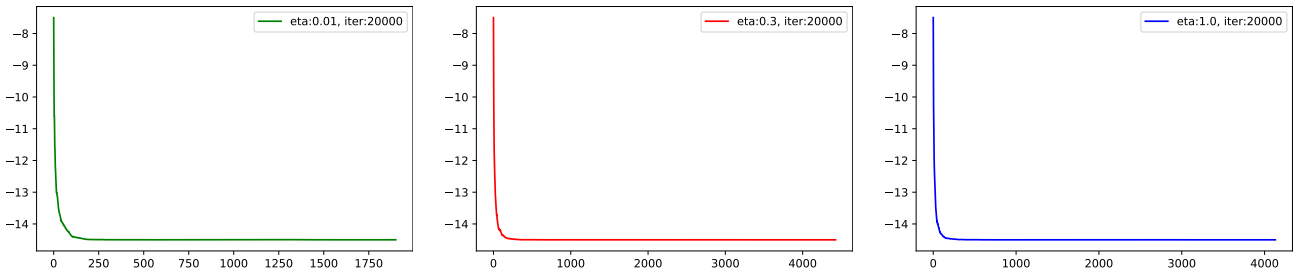

Figure 2.16: As f and σ_t^2 are independent of η, f behaves the same in all the cases without perturbations as σ_t^2 is small, but as the rate decreases we see the algorithm runs for 1781, 4428 and 4132 iterations in the Green, Red and Blue graphs respectively.

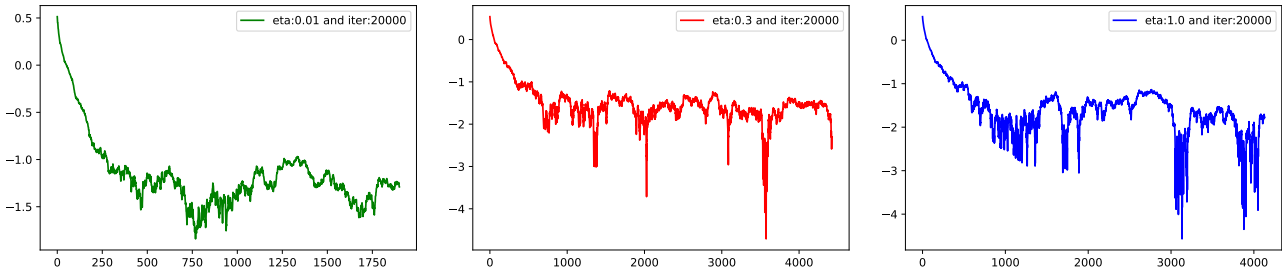

Figure 2.17: Here we see the distance to the minima becomes better as the rate decreases, until it reaches $\|w^t - w^*\| \approx 10^{-4}$. The oscillations are almost around 10^{-2} and this is referred to the small Noise given $\sigma_t^2 = 0.5$.

2. **We have time-dependent rate with time-independent Gaussian Noise:** $\alpha_t = \dfrac{0.1}{\sqrt{t}}$ and $\sigma_t^2 = 4.0$.

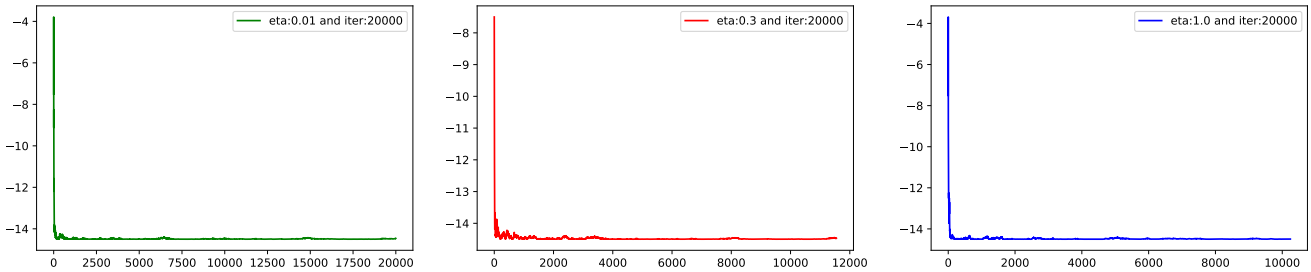

Figure 2.18: Here we just increase σ_t^2. We realise that f still behaves the same in all the cases but with small perturbations specially at the beginning as the rate is slightly bigger.

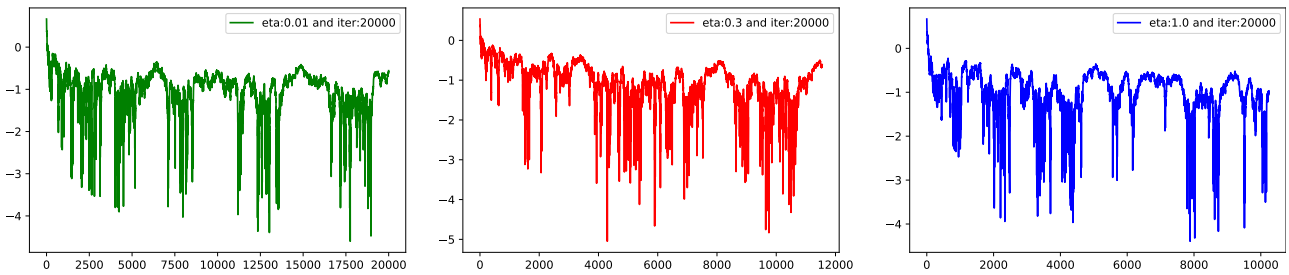

Figure 2.19: Similarly, as the previous case, but as σ_t^2 increases we realize that the perturbations are mostly around 10^{-4} and the distance to the minima becomes better as $\|w^t - w^*\| \approx 10^{-4}$.

3. **Here we have time-dependent rate with time-dependent Gaussian Noise such that:**

$$\sigma_t^2 = \frac{\eta}{(1+t)^\gamma} \quad \text{and} \quad \alpha_t = \frac{0.1}{\sqrt{t}}$$

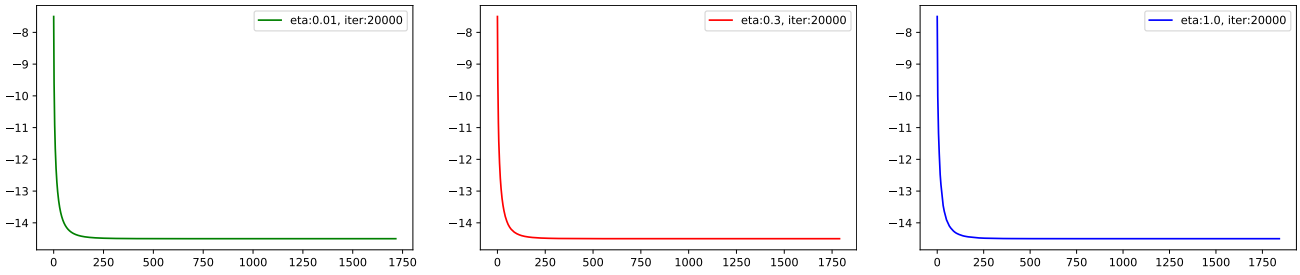

Figure 2.20: As the learning rate set as a decreasing function along the iterations and as f is independent of η we see that f behaves the same in all cases, but regarding the Noise we know that σ_t^2 decreases and vanishes along the iterations. In our case where the convergence is so fast, in order to see an impact on the values of f we need to increase or fix σ_t^2.

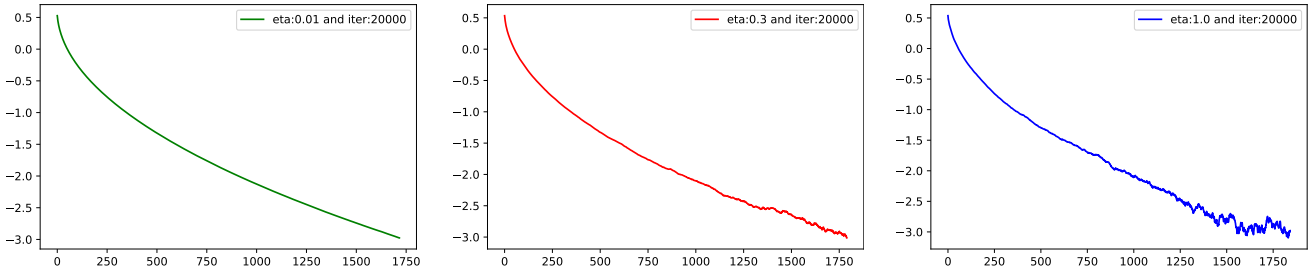

Figure 2.21: Here we see the impact of η. The convergence is so fast, we realize that as η increases more perturbations occurred.

4. **We have constant rate with time-dependent Gaussian Noise:** $\sigma_t^2 = \frac{\eta}{(1+t)^\gamma}$ and $\alpha_t = 0.1$

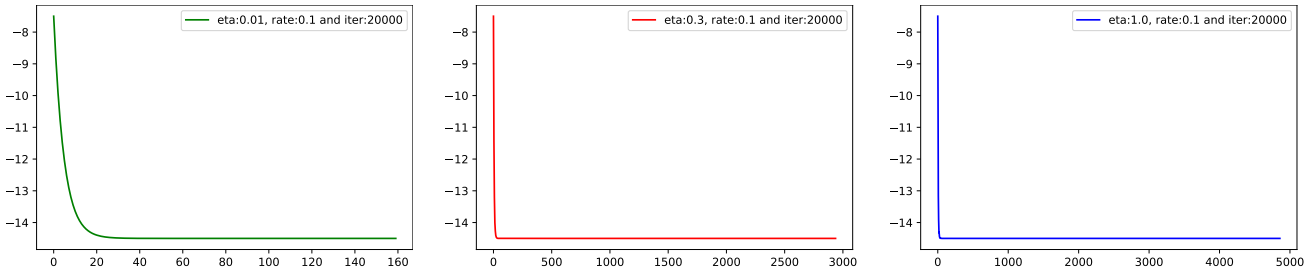

Figure 2.22: In the Green graph as η is small, the minimum occurred after 20 iterations and the algorithm runs for just 160 iterations, and as η increases the algorithm runs up to 5000 iterations without any perturbations as σ_t^2 vanishes along the iterations.

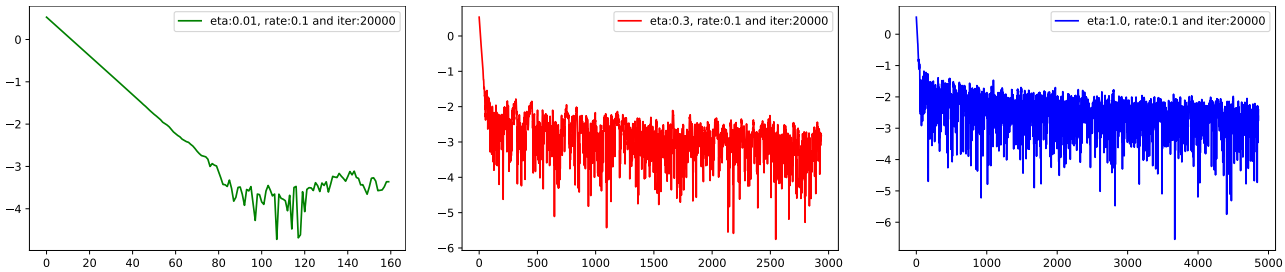

Figure 2.23: We see in the Green graph that the perturbations approximately start after 60 iterations. As η increases the algorithm runs up to 5000 iterations to see that the convergence around the minima with almost same amplitude as $\|w^t - w^*\| \approx 10^{-4}$.

5. **We have time-independent rate with time-independent Gaussian Noise:** $\alpha_t = 0.1$ and $\sigma_t^2 = 0.5$.

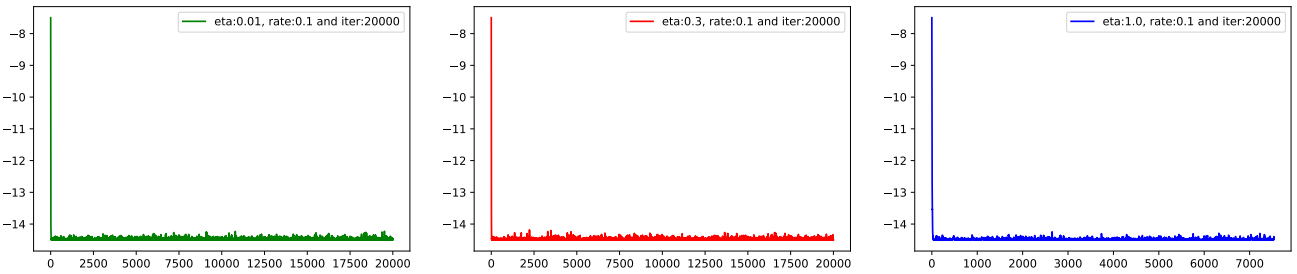

Figure 2.24: As the rate is constant and f is independent of η, f behaves the same in all cases with small perturbations..

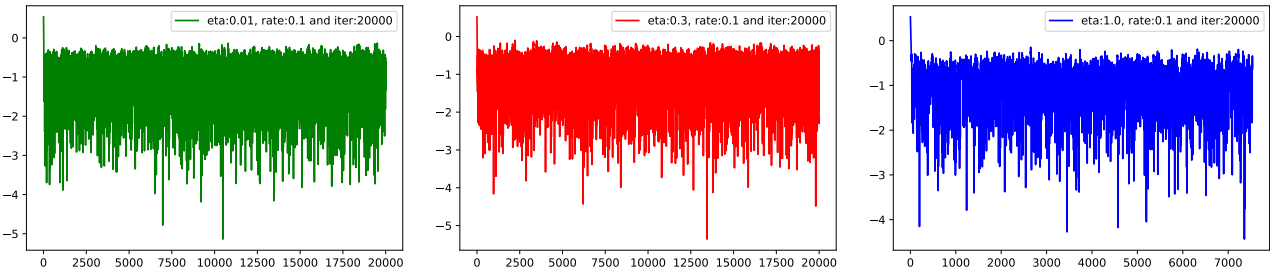

Figure 2.25: As the rate is constant, we can see better the impact of the Noise. We observe that after 500 iterations the perturbations is almost with same amplitude as $\|w^t - w^*\| \approx 10^{-4}$ and σ_t^2 is constant.

6. **Here we have time-independent rate with time-independent Gaussian Noise:** $\alpha_t = 0.1$ and $\sigma_t^2 = 2.0$.

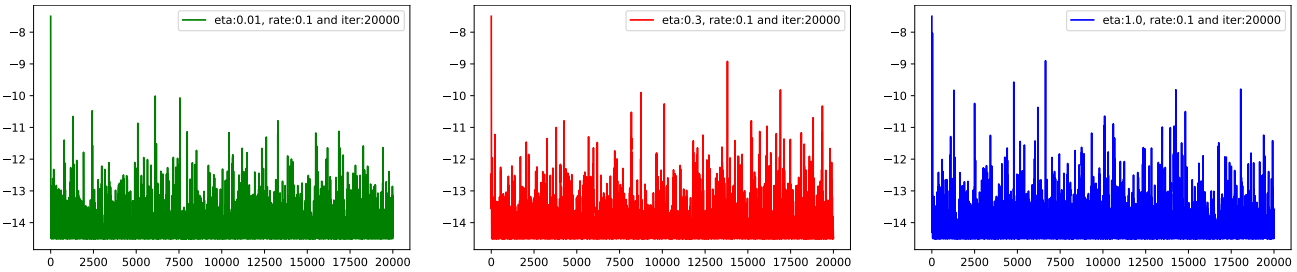

Figure 2.26: We just increase σ_t^2. It is clear to get more perturbations with higher amplitude compared to the previous case.

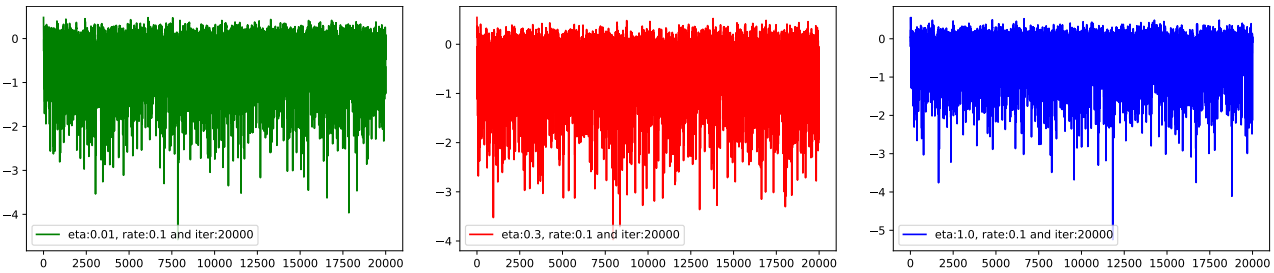

Figure 2.27: Compared to the previous case, as we increase σ_t^2, we see slightly more oscillations with almost same amplitude.

Remark 2.3.5. In gradient descent, if there is Noise, using constant step size does not guarantee convergence to the optimum, as the noise will give iterates that move around too much near the optimum. It much better to use a step size that is proportional to distance to the optimal solution as the step size is often depends upon the gradient of various weights at a particular point (if the gradient is large, typically a large step size is used. As the gradient decreases, the step size is lowered).

Chapter 3

Stochastic Semi-discrete Optimization.

In this chapter we will consider a stochastic approach to solve the semi-discrete optimal transport problem for large-scale problems where the dimension is high. However, evaluating the sum-gradient may require expensive evaluations of the gradients from all summand functions. To economize on the computational cost at every iteration, stochastic gradient descent samples a subset of summand functions at every step. This is very effective in the case of large-scale machine learning problems.

3.1 Average stochastic gradient descent-ASGD

In order to be optimal, the weights defining the transport maps should solve a $\mathcal{C}^1$-concave maximization problem

$$(KP) = \max_{w \in \mathbb{R}^d} \mathcal{K}(w), \tag{3.1}$$

where $\mathcal{K}$ is the Kantorovich functional given by

$$\mathcal{K}(w) = \int_{\mathbb{R}^d} w^c \, d\mu - \int_{\mathbb{R}^d} w \, d\nu = \int_{\mathbb{R}^d} w^c(x) \, d\mu(x) - \sum_{i=1}^{N} w_i \, \nu_i \tag{3.2}$$

Here, we will exploit and study a stochastic optimization algorithm for semi-discrete OT that was recently proposed in [17].

In a high-dimensional setting [16], one may turn to using Monte-Carlo estimates for the gradient instead of exact computations. In other words, the maximization of $\mathcal{K}$ can be addressed with stochastic gradient ascent, which is made possible by writing

$$\mathcal{K}(w) = \mathbf{E}[h(X, w)], \quad \text{where} \quad h(X, w) = w^c(x) - \sum_{i=1}^{N} w_i \, \nu_i \tag{3.3}$$

and where X is a random variable of distribution μ. Notice that for $x \in L_i^w$, $w \mapsto w^c(x)$ is smooth with gradient $-e_i$ (where (e_i) is the canonical basis of $\mathbb{R}^d$). Therefore, for any $v \in \mathbb{R}^d$, for almost all $x \in \mathbb{R}^d$, $w \mapsto h(x, w)$ is differentiable at v and

$$\nabla_w \, h(x, v) = -e_{T_v(x)} + \nu. \tag{3.4}$$

In order to minimize $-\mathcal{K}$, Genevay and al. [17] recently proposed the following averaged stochastic gradient descent (ASGD) initialized with $\tilde{w}^1 = 0$

$$\begin{cases} \tilde{w}^k = \tilde{w}^{k-1} + \dfrac{C}{\sqrt{k}} \nabla_w \, h(x^k, \tilde{w}^{k-1}), \quad \text{where } x^k \sim \mu \\ w^k = \dfrac{1}{k}(\tilde{w}^1 + \cdots + \tilde{w}^k). \end{cases} \tag{3.5}$$

Since $\nabla_w \, h(x^k, \tilde{w}^{k-1})$ exists $x - a.s.$ and is bounded, the convergence of this algorithm is ensured by Theorem 2.3.1 in the sense $\max(\mathcal{K}) - \mathbf{E}[\mathcal{K}(w^k)] = \mathcal{O}\left(\dfrac{\log k}{\sqrt{k}}\right)$.

3.1.1 Numerical study and results of the stochastic methods for OT

In order to maximize the loss function $\mathcal{K}$ (minimize $-\mathcal{K}$), we propose the ASGD to approximate the optimal solution w of the semi-discrete problem. Here we have μ is the uniform distribution on $[0,1]^2$ and ν is a discrete target distribution on the red points y (manually chosen) drawn in each figure. We have 1000 points sampled according to the source distribution μ in order to approximate or average the SGD. The green segments are here to represent the transport map $x \mapsto T_w(x)$, that transport the source points to the target point in each Laguerre cell L_i^w such that $\mu(L_i^w) = \nu(y_i)$.

Figure 3.1: We have a distribution of 7 points with respective masses $\nu(y_i) = 0.24, 0.2, 0.22, 0.15, 0.1, 0.02$ and 0.04. We notice that between zero and 18 iterations of SGD, it is clear that an estimate of the cost function $-\mathcal{K}$ tries to go towards the optimum, it decreases rapidly from -0.035 to -0.05, this is due to the stability that the loss function tends to, the initial part still seems to make pretty more stable progress and as it is close to solution it fluctuate more. Afterwards as the number of iterations increases the estimate of $-\mathcal{K}$ oscillates around the optimal value, and this oscillation is due to the random variations and the estimation of the loss function $-\mathcal{K}$.

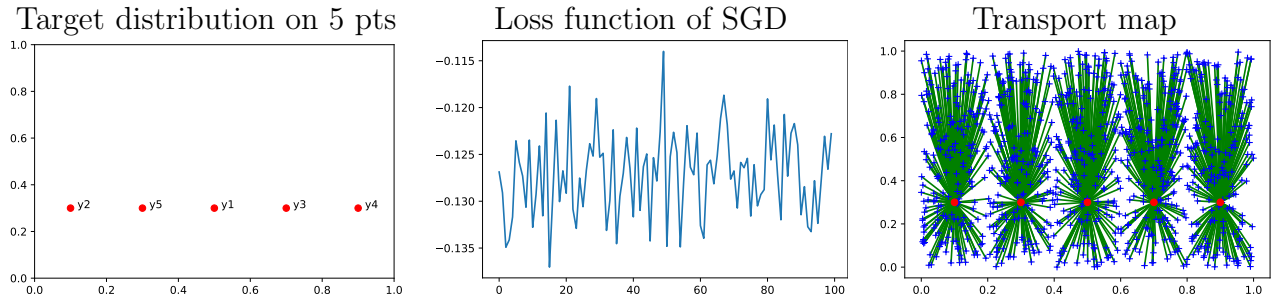

Figure 3.2: In this case we changed both the positions and masses of the given points, thus we have a uniform distribution of 5 points with equal masses such that $\nu(y_i) = 0.2$. We notice that there is a small oscillation of the estimate of the cost function $-\mathcal{K}$ around the optimal value, which is the case of Voronoi cells for w close to 0.

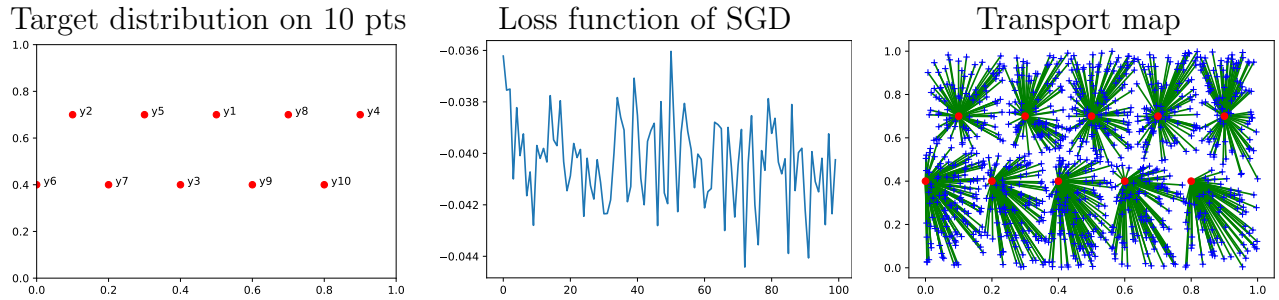

Figure 3.3: In this case we kept the equality of masses where $\nu(y_i) = 0.1$. and we only changed the positions of the given 10 points to realise an estimate of $-\mathcal{K}$ behaves in the same way of previous case.

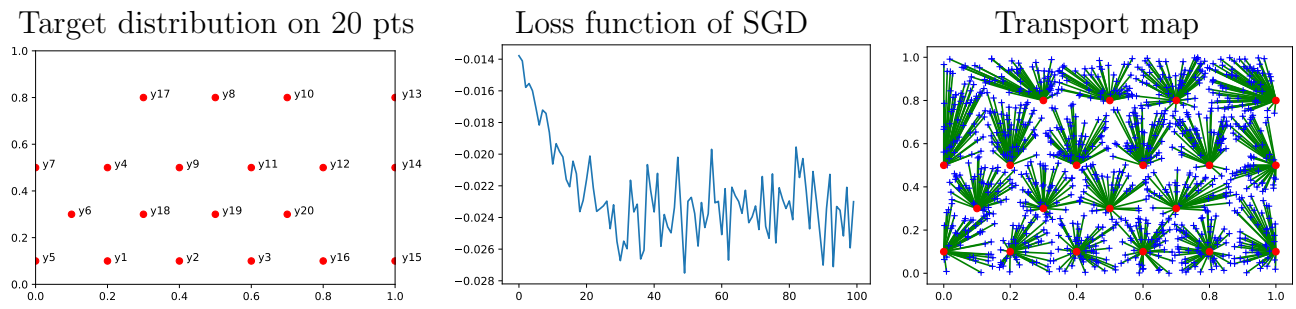

Figure 3.4: More interesting than the previous, we have a distribution of 20 points with equal masses such that $\nu(y_i) = 0.05$. We notice that between zero and 20 iterations the estimate of the cost function $-\mathcal{K}$ decreases rapidly from -0.014 to -0.024, more stable progress in the beginning. In the end as $-\mathcal{K}$ comes close to the optimal solution it fluctuates more, where the oscillations around the optimal value are only due to the random variations and due to the estimation of the loss function $-\mathcal{K}$.

Figure 3.5: We have a distribution of 20 points such that $0.03 \leq \nu(y_i) \leq 0.08$. We notice that between zero and 30 iterations the estimate of the cost function $-\mathcal{K}$ decreases rapidly from -0.015 to -0.04, until the loss function tends to stabilize and oscillates around the optimal value.

Conclusion

Computing the c-transform is essentially a nearest neighbor search. The regularity properties of the optimal dual variables will allow to propose new approximation classes for ϕ and ψ, and thus design new scalable numerical OT solvers that rely on a principled parameterization of the dual variables ϕ and ψ.

The developed numerical solutions could be integrated to several applications that require large-scale optimal transport, for example domain adaptation [10], generative networks [11], texture synthesis [12], or shape analysis [13].

By looking at this list of applications, we can see that the field of large-scale optimal transport still important research effort, where many tasks could be study such as: Analyze the regularity of optimal dual variables ϕ and ψ, quantize the impact of restricting to a sub-class of dual variables, compare with other techniques based on entropic regularization [14] and regularization of Brenier potentials [15] and incorporate the proposed numerical solver to address OT problems in machine learning and image processing. However, the stochastic gradient descent proposed in [8] does not scale well when the cardinal of Y is large. In contrast, the stochastic approach of [9] based on a parameterization with neural networks should scale better, but remains to be analyzed precisely.

Bibliography

[1] F. Santambrogio, Optimal Transport for Applied Mathematicians, book, Birkhauser, dans Progress in Nonlinear Differential Equations and Their Applications 87, Birkhauser Basel (2015).

[2] L. Ambrosio, and G. Nicola. "A user's guide to optimal transport." In Modelling and optimisation of flows on networks, pp. 1-155. Springer, Berlin, Heidelberg, 2013.

[3] Q. Merigot, B. Thibert. Optimal transport: discretization and algorithms. 2020.

[4] M. Thorpe, Introduction to Optimal Transport, F2.08, Centre for Mathematical Sciences University of Cambridge Lent 2018 Current Version: Thursday 8th March, 2018.

[5] C. Villani. Topics in Optimal Transportation. American Math. Society, 2003.

[6] J. Kitagawa, Q. Mérigot, and B. Thibert. A Newton algorithm for semi-discrete optimal transport. Journal of the European Math Society, 2017.

[7] B. Bercu and J. Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. preprint arXiv:1812.09150, 2018.

[8] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In Advances in neural information processing systems, pages 3440–3448, 2016.

[9] V. Seguy, B. Bhushan Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-Scale Optimal Transport and Mapping Estimation. preprint arXiv:1711.02283, 2017.

[10] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(9):1853–1865, 2015.

[11] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. preprint arXiv:1701.07875, 2017.

[12] A. Leclaire and J. Rabin. A fast multi-layer approximation to semi-discrete optimal transport. In Proceedings of SSVM, pages 341–353. Springer, 2019.

[13] J. Feydy, P. Roussillon, A. Trouvé, and P. Gori. Fast and scalable optimal transport for brain tractograms. In Proceedings of MICCAI, pages 636–644. Springer, 2019.

[14] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Conference on Neural Information Processing Systems (NIPS'13), pages 2292–2300, 2013.

[15] F.-P. Paty, A. d'Aspremont, and M. Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. preprint arXiv:1905.10812, 2019.

[16] B. Galerne, A. Leclaire, J. Rabin. A Texture Synthesis Model Based on Semi-discrete Optimal Transport in Patch Space. SIAM Journal on Imaging Sciences, Society for Industrial and Applied Mathematics, 2018.

[17] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, 2016. Stochastic optimization for large-scale optimal transport. In Advances in neural information processing systems (pp. 3440-3448).

[18] Robert M. Gower. Convergence Theorems for Gradient Descent, October 5, 2018.