UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

A CONTINUOUS-TIME MODEL OF STOCHASTIC GRADIENT DESCENT: CONVERGENCE RATES AND COMPLEXITIES UNDER ŁOJASIEWICZ INEQUALITY

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO

RODRIGO IGNACIO MAULÉN SOTO

PROFESOR GUÍA:
JUAN PEYPOUQUET URBANEJA
PROFESOR GUÍA 2:
GUILLAUME GARRIGOS

COMISIÓN:
JAIME SAN MARTÍN ARISTEGUI

Este trabajo ha sido parcialmente financiado por FONDECYT 1181179, ECOS 180039 Y CMM ANID PIA AFB170001

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS
POR: RODRIGO IGNACIO MAULÉN SOTO
FECHA: ABRIL 2021
PROFESORES GUÍA: JUAN PEYPOUQUET URBANEJA. GUILLAUME GARRIGOS.

A CONTINUOUS-TIME MODEL OF STOCHASTIC GRADIENT DESCENT: CONVERGENCE RATES AND COMPLEXITIES UNDER ŁOJASIEWICZ INEQUALITY

In this thesis we study the convergence rates and complexities of a continuous model of the Stochastic Gradient Descent (SGD) under convexity, strong convexity and Łojasiewicz assumptions, the latter being a way to generalize the concept of strong convexity. In the first chapter, the concept of Łojasiewicz inequality is introduced and results of the Gradient Descent method in its discrete and continuous version are shown for this case, together with the convex and strongly convex cases. In the second chapter, the SGD method is introduced, results are shown under convex and strongly convex assumptions, and new results are obtained under the Łojasiewicz Inequality. Then it is shown how to construct a continuous-time model of SGD under a Gaussian variance assumption and, at last, the necessary concepts of Stochastic Analysis are introduced to understand this model. Finally, in the third chapter this model is analyzed and its upper bounds and complexities are obtained, where it is deduced that if the variance is sufficiently small, then the complexity of the model matches that of the Gradient Descent for the cases: convex, strongly convex and Łojasiewicz.

En esta tesis se estudian las tasas de convergencia y complejidades de un modelo continuo del método del Gradiente Estocástico (SGD) bajo supuestos de convexidad, fuerte convexidad y Łojasiewicz, siendo este último una forma de generalizar el concepto de fuerte convexidad. En el primer capítulo se introduce el concepto de desigualdad de Łojasiewicz y se muestran resultados del método del Gradiente Descendente en su versión discreta y continua para este caso, junto con los casos convexo y fuertemente convexo. En el segundo capítulo se introduce el método de SGD, se muestran resultados bajo suposiciones de convexidad y fuerte convexidad, y se obtienen nuevos resultados bajo la desigualdad de Łojasiewicz. Luego se muestra como construir un modelo continuo de SGD bajo una suposición de varianza Gaussiana y por último se introducen los conceptos necesarios de Análisis Estocástico para comprender este modelo. Finalmente, en el tercer capítulo se analiza este modelo y se obtienen sus cotas superiores y complejidades, donde se deduce que si la varianza es suficientemente pequeña, entonces la complejidad del modelo coincide la del método del Gradiente Descendente para los casos: convexo, fuertemente convexo y Łojasiewicz.

A mi abuelita Lucy

Agradecimientos

Este trabajo ha sido producto de mucho esfuerzo, agradezco a la gente que amo por hacerme seguir adelante.

En primer lugar quiero agradecer a mi familia, a mis papás, Maribel y Rodrigo, por haber dado todo de sí para que llegue donde estoy hoy, a mi hermano Vicente por ser un pilar fundamental en mi vida y al Chavo, los amo con todo mi corazón. También a mi abuelita Paty y a mi tía Aly, sus juventudes de alma me alegran diariamente. Un agradecimiento especial a mi abuelita Lucy, que no alcanzó a presenciar este momento pero sé que sería la más feliz de todas.

Quiero dar gracias a mis amigos, soy un afortunado por verme rodeado de gente como ellos. A Pathetics, que me ha apoyado desde el primer día y me han acompañado por más de 10 años. A toda la gente linda del DIM, que hizo mi paso por la universidad mucho más ameno, en especial a la casa, donde salieron las ideas fundamentales para esta tesis. Al Nacho, que será el mejor artista del mundo. A Maira, por haber aprendido tanto juntos y animarme a comenzar en el veganismo.

Por último a Franny, que tiene un corazón hermoso y es la mejor compañía que podría tener, además fue mi principal colaboradora en pos de que el inglés de esta tesis tenga sentido.

Table of Contents

Notations

Let $n, m \in \mathbb{N}$.

- $[n] := \{1, \dots, n\}$.
- $\mathbb{R}^{n \times m}$ is the set of the real matrices of size $n \times m$.
- Let $M \in \mathbb{R}^{n \times n}$ and $M = (m_{ij})_{\mathrm{i},j \in [n]}$, then $tr(M) = \sum_{\mathrm{i}=1}^{n} m_{\mathrm{ii}}$.
- Let $M \in \mathbb{R}^{n \times m}$, $M^t \in \mathbb{R}^{m \times n}$ is the transpose of the matrix M.
- If $M \in \mathbb{R}^{n \times m}$, $\|M\|_F$ is the Frobenius norm defined as $\sqrt{tr(MM^t)}$.
- $\mathbb{R}^n_+$ is the positive orthant in $\mathbb{R}^n$.
- If $(x_k)_{k \in \mathbb{N}}$ is a sequence, then $\bar{x}_k$ is the average sequence $\frac{1}{k} \sum_{\mathrm{i}=1}^{k} x_{\mathrm{i}}$.

Let Ω be a probability space.

- Let $(\mathrm{i}_k)_{k \in \mathbb{N}}$ a sequence of random variables and $\mathcal{D}$ a distribution, $(\mathrm{i}_k)_{k \in \mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{D}$ if each i_k is drawn independently and identically according to the law of $\mathcal{D}$.
- If $X : \Omega \to \mathbb{R}^{\mathrm{d}}$ is a random variable, $\mathbb{E}[X]$ is the expectation of X and $\mathbb{V}[X]$ is the variance of X. If $\mathcal{D}$ is a distribution such that $X \sim \mathcal{D}$ and $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^p$ a function, then the expectation and variance of $f(X)$ are denoted by $\mathbb{E}_{\mathcal{D}}(f(X))$ and $\mathbb{V}_{\mathcal{D}}(f(X))$, respectively.
- If $X : \Omega \times \mathbb{R}_+ \to \mathbb{R}^{\mathrm{d}}$ is a stochastic process, then $\bar{X}(\omega, t)$ is the average process $t^{-1} \int_0^t X(\omega, s) ds$.

Let H be a real Hilbert space.

- Let $c \in \mathbb{R}$ and $f \in H \to \mathbb{R}$, $[f < c] = \{x \in H : f(x) < c\}$.
- For $x \in H$, $\|x\| = \sqrt{\langle x, x \rangle}$.
- For $x_0 \in H$, $r > 0$, then $B(x_0; r) := \{x \in H : \|x - x_0\| < r\}$.
- Let $(x_k)_{k \in N} \subset H, x \in H$, then $x_k \rightharpoonup x$ if $\langle x_k, y \rangle \to \langle x, y \rangle$ for all $y \in H$.

- If $A, B \subseteq H$, then $d(A, B)$ is the distance between A and B.
- If $A \subseteq H$, then $cl(A), int(A)$ is the closure and the interior of A, respectively.

Let B a finite set and $|B|$ its cardinality.

- For $b \in [|B|]$, $\binom{B}{b} := \{X \subseteq B : |X| = b\}$.
- $Unif(B)$ is the uniform law on the set B.

Let $f, g : \mathbb{R}_+ \to \mathbb{R}$ be functions.

- $f(x) = \mathcal{O}(g(x))$ if there exists $C > 0, x_0 > 0$ such that $f(x) \leq Cg(x)$ for all $x \geq x_0$.
- $f(x) = \Omega(g(x))$ if there exists $C > 0, x_0 > 0$ such that $f(x) \geq Cg(x)$ for all $x \geq x_0$.
- $f(x) = \Omega_0(g(x))$ if there exists $C > 0, x_0 > 0$ such that $f(x) \geq Cg(x)$ for all $x \in (0, x_0)$.
- $\tilde{\Omega}_0(g(x)) := \Omega_0\left(g(x) ln\left(\frac{1}{x}\right)\right)$.

Introduction

Optimization is the selection of the best element (under some criterion) from some set of available alternatives. The applications of this principle can be found in all types of disciplines, such as: Computer Science, Operations Research, Economics, Machine Learning, etc.. An optimization problem can be represented (without loss of generality) as:

Given some set $\mathcal{K}$, and a function $f : \mathcal{K} \to \mathbb{R}$, find a x^* such that $f(x^*) \leq f(x)$, for all $x \in \mathcal{K}$.

In order to find a solution, optimization algorithms (methods) work as follows: given an initial point $x_0 \in \mathcal{K}$, these algorithms will generate a sequence $(x_k)_{k\in\mathbb{N}}$, such that, while k grows, the sequence gets "closer" (in some sense) to a solution. In this Thesis we will distinguish two types of algorithms: Deterministic and Stochastic. For deterministic algorithms, the calculations to generate a new element of the sequence are exact, i.e. given $x_0 \in \mathcal{K}$, the same sequence will always be obtained each time the algorithm is run. Instead, for stochastic algorithms, the calculations depend on random variables, so obviously, given $x_0 \in \mathcal{K}$, different sequences will be obtained each time the algorithm is run. Nevertheless, we can often deduce results in expectation.

We will assume that $\mathcal{K} = H$ is a real Hilbert space and that $f : H \to \mathbb{R}$ is a differentiable and convex function. In this context, a very popular deterministic algorithm is the Gradient Descent (GD) method, which takes the following form: given any $x_0 \in H$,

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k), \quad k \in \mathbb{N}, \tag{1}$$

where $\gamma_k > 0$. If ∇f is Lipschitz continuous and γ_k is sufficiently small, then $f(x_k)$ converges to the minimum value of f.

The process of updating a sequence using an algorithm can be seen as a discrete dynamical system. However, there are also continuous dynamics that can let us achieve the minimum value of f. Although these two types of dynamics are a world unto themselves, they enjoy remarkable connections between them. As an instance of these connections, (1) can be seen as a first order discretization of a continuous dynamic, in effect, (1) can be rewritten as

$$\frac{x_{k+1} - x_k}{\gamma_k} = -\nabla f(x_k), \quad k \in \mathbb{N}.$$

If we impose $\gamma := \sup_{k\in\mathbb{N}} \gamma_k \to 0$, then we obtain its underlying continuous dynamic (CGD):

$$\begin{aligned} &\dot{x}(t) = -\nabla f(x(t)), \;\; t > 0 \\ &x(0) = x_0. \end{aligned} \tag{2}$$

It is easy to see that f decreases along the trajectory $x(t)$, since

$$\frac{d}{dt}(f(x(t))) = -\|\nabla f(x(t))\|^2 \leq 0,$$

this fact inspires the analysis of first-order descent methods (for more details about the convex case, see [1], for the non-convex case, see [2]). Besides, we obtain the same result as before in this dynamic, this is, $f(x(t))$ converges to the minimum value of f.

For the remainder of this Introduction, we will suppose that ∇f is Lipschitz continuous. A natural question is to wonder how "fast" a dynamic converges to the minimum, which is called "convergence rate". A classical result about this concept is that if the stepsize is constant (i.e. $\gamma_k \equiv \gamma$) and sufficiently small, then Gradient Descent (1) satisfies:

$$f(x_k) - min(f) = \mathcal{O}\left(\frac{1}{k}\right).$$

Interestingly, the continuous dynamic (2) satisfies:

$$f(x(t)) - min(f) = \mathcal{O}\left(\frac{1}{t}\right).$$

Those results can be improved if we have more information about the properties or the geometry of the function. For instance, if f is strongly convex, then (1) and (2) have linear convergence rates i.e., there exist $\rho, \tilde{\rho} \in (0,1)$ such that:

$$f(x_k) - min(f) = \mathcal{O}(\rho^k)$$

and

$$f(x(t)) - min(f) = \mathcal{O}\left(\tilde{\rho}^t\right).$$

respectively.

In practice, strong convexity is often too restrictive, so with the idea of relaxing it and retaining fast convergence rates, an interesting property to delve into is the Łojasiewicz Inequality with exponent $q \in [0,1)$. Roughly speaking, this property describes convex functions that behave like

$$x \mapsto d(x, argmin(f))^{\frac{1}{1-q}}$$

around its minimizers. An intuition of the functions that satisfy the Łojasiewicz Inequality is the following: the bigger q is, the "flatter" the function is (around its minimizers), therefore, a Gradient Descent algorithm will converge progressively slower as q grows. On the other hand, a strongly convex function ($\{x^*\} = argmin(f)$) behaves like

$$x \mapsto \|x - x^*\|^2$$

around x^*, which coincides with the behavior of functions that satisfy the Łojasiewicz Inequality with $q = \frac{1}{2}$.

As we have said, the convergence rates of a Gradient Descent algorithm should get progressively worse as q grows (we will focus on the $q \geq \frac{1}{2}$ case). More precissely, if f satisfies the Łojasiewicz Inequality with exponent $q \in [\frac{1}{2}, 1)$, then [3] ensures that a sequence $(x_k)_{k\in\mathbb{N}}$ generated by (1) satisfies:

$$\begin{cases} f(x_k) - min(f) = \mathcal{O}\left(\rho^k\right) \text{ for } \rho \in (0,1) & \text{if } q = \frac{1}{2}, \\ f(x_k) - min(f) = \mathcal{O}\left(k^{-\frac{1}{2q-1}}\right) & \text{if } q \in (\frac{1}{2}, 1). \end{cases}$$

And $x(t)$ following the dynamic of (2) satisfies:

$$\begin{cases} f(x(t)) - min(f) = \mathcal{O}(\rho^t) \text{ for } \rho \in (0,1) & \text{if } q = \frac{1}{2}, \\ f(x(t)) - min(f) = \mathcal{O}(t^{-\frac{1}{2q-1}}) & \text{if } q \in (\frac{1}{2}, 1). \end{cases}$$

Once again, we see that discrete and continuous dynamics share similar convergence rates under Łojasiewicz assumptions, and that these convergence rates get worse as q grows, as predicted.

Next, we will focus on the stochastic case and compare results.

If we want to implement the Gradient Descent Method (1) computationally, this algorithm relies on the fact that the gradient of f is "cheap" to compute, but in Machine Learning and Big Data optimization this is not the case. We will usually have a function of the form

$$f := \frac{1}{n}\sum_{\mathrm{i}=1}^{n} f_{\mathrm{i}}. \tag{3}$$

Where each f_{i} corresponds to one data point, so computing $\nabla f(x)$ cost n data points (which can be of the order of millions or bigger). In order to manage this structure, an idea would be to estimate the gradient instead of calculating it. In this sense, one of the most popular stochastic algorithms is the Stochastic Gradient Descent (SGD), which takes the following form: given any $x_0 \in H$,

$$x_{k+1} = x_k - \gamma_k \nabla f_{\mathrm{i}_k}(x_k), \tag{4}$$

where $(\mathrm{i}_k)_{k\in\mathbb{N}} \overset{\text{iid}}{\sim} Unif[n]$. The key idea behind this algorithm, as previously said, is that ∇f_j (with $j \sim Unif[n]$) is an unbiased estimator from the gradient of f, i.e. $\mathbb{E}[\nabla f_j(\cdot)] = \nabla f(\cdot)$. As an initial result on their convergence rates, the following can be observed: if every f_{i} has Lipschitz continuous gradient and $\gamma_k = \frac{C}{k^{2/3}}$ for some $C > 0$, then [4, Theorem 4] ensures that a sequence $(x_k)_{k\in\mathbb{N}}$ generated by (4) satisfies:

$$\mathbb{E}(f(x_k)) - min(f) = \mathcal{O}\left(\frac{1}{k^{1/3}}\right).$$

If instead, we choose a constant stepsize in (4), then the algorithm in general *does not converge*. Nevertheless, we can obtain upper bounds, for instance, if the constant stepsize γ is sufficently small, then

$$\mathbb{E}(f(\bar{x}_k)) - min(f) = \mathcal{O}\left(\frac{1}{k}\right) + \mathcal{O}(\sigma^2),$$

where $\sigma^2 := \mathbb{V}\left[\nabla f_j(x^*)\right]$ (with $j \sim Unif\,[n]$).

Moreover, if f is strongly convex, then there exists $\rho \in (0,1)$ such that:

$$\mathbb{E}(f(x_k)) - min(f) = \mathcal{O}(\rho^k) + \mathcal{O}(\sigma^2).$$

These upper bounds are interesting because if we had the "interpolation property" (see [5]), which implies that $\sigma^2 = 0$, we could recover the convergence rates of (1) in the convex and the strongly-convex case. This leads us to believe that the same thing (about upper bounds and convergence rates) will hold for (SGD) under another properties, such as Łojasiewicz Inequality. However, there are not many results in this case. Even so, we will attempt to get some. In general, σ^2 is strictly positive, so we cannot ensure a convergence rate if this variance is constant. Nevertheless, as a comment (since no work was done on this subject throughout this Thesis), there are stochastic algorithms, such as SVRG and SAGA (see [6]) that use variance reduction techniques and obtain the same convergence rates as in the deterministic case for convex and strongly convex functions.

A concept related to the velocity of convergence of an algorithm is the number of steps we must take until we get a "good enough" solution, this will be called complexity. More precissely, the complexity is a function $u : \mathbb{R}_+ \to \mathbb{R}$ such that, if we take a particular $k = \Omega_0(u(\varepsilon))$, then

$$f(x_k) - min(f) \leq \varepsilon.$$

Remark. The definition of the complexity for continuous dynamics is analogous if we replace adequately k for t, it is also analogous in the stochastic case if we take a expectation over $f(x_k)$. Additionally, we will use the averaged sequence (or the averaged process) in the complexity for the convex case. Besides, in the stochastic case we will use the definition of complexity with $\tilde{\Omega}_0$ (see Notations).

Another interesting aspect about the upper bounds previously shown is that, although in general we will not have convergence rates if the stepsize is constant in the SGD case, we can deduce complexities if we impose the stepsize to be sufficiently small.

In the following table are gathered the complexities of the algorithms mentioned before. The complexities of SGD under Łojasiewicz Inequality assumptions are demonstrated in this Thesis (see Propositions 2.13 and 2.17), the rest of the complexities are shown throughout this document and can be found in the literature.

Property	Complexity GD	Complexity CGD	Complexity SGD
Convex	ε^{-1}	ε^{-1}	ε^{-2}
Strongly Convex	$ln(\varepsilon^{-1})$	$ln(\varepsilon^{-1})$	ε^{-1}
Łojasiewicz with $q = 1/2$	$ln(\varepsilon^{-1})$	$ln(\varepsilon^{-1})$	ε^{-1}
Łojasiewicz with $q \in (1/2, 1)$	$\varepsilon^{-(2q-1)}$	$\varepsilon^{-(2q-1)}$	$\varepsilon^{-(4q-1)}$

We started this Introduction deriving a continuous dynamic from an algorithm in the deterministic case. Now we want to do the same in the stochastic case, because by some aspects it is easier to work with continuous dynamics rather than with discrete algortihms. Although there is no unique way to do this, under some technical assumptions, we can model a continuous version of SGD in the case of $H = \mathbb{R}^{\mathrm{d}}$. The modeled dynamic is called a Stochastic Differential Equation (SDE) and has the following form: given any $X_0 \in \mathbb{R}^{\mathrm{d}}$,

$$\begin{aligned} dX(t) &= -\nabla f(X(t))\,dt + \sigma(t, X(t))\,dB(t), t \geq 0 \\ X(0) &= X_0, \end{aligned} \tag{5}$$

where $\sigma : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$ and $B(t)$ is a $\mathrm{d}-$dimensional Brownian motion. We are interested in finding upper bounds and complexities of (5) under the same previous properties of f, this is: convexity, strong convexity and Łojasiewicz Inequality. In order to do so, we will use Stochastic Analysis tools.

Indeed, let $X(t)$ satisfying (5), ∇f be $L-$Lipschitz continuous and

$$\sigma_*^2 := \sup_{t\geq 0, x\in\mathbb{R}^{\mathrm{d}}} \|\sigma(t,x)\|_F^2 < \infty.$$

Under some additional technical assumptions, the main results of this Thesis are:

- If $\bar{f}(t) = t^{-1}\int_0^t f(X(s))\,ds$, then

$$\mathbb{E}\left[\bar{f}(t)\right] - min(f) = \mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}(\sigma_*^2).$$

- If f is strongly convex ($\{x^*\} = argmin(f)$), then there exists $\rho \in (0,1)$ such that:

$$\frac{2}{L}\mathbb{E}[f(X(t) - min(f)] \leq \mathbb{E}\left[\|X(t) - x^*\|^2\right] = \mathcal{O}(\rho^t) + \mathcal{O}(\sigma_*^2).$$

- If f satisfies the Łojasiewicz Inequality with $q = \frac{1}{2}$, then there exists $\rho \in (0,1)$ such that:
$$\mathbb{E}\left[f(X(t))\right] - min(f) = \mathcal{O}(\rho^t) + \mathcal{O}(\sigma_*^2).$$

- If f satisfies the Łojasiewicz Inequality with $q \in (\frac{1}{2}, 1)$, then
$$\mathbb{E}\left[f(X(t))\right] - min(f) = \mathcal{O}\left(e^{-\mathcal{O}(\sigma_*^2)t}\right) + \mathcal{O}\left(\sigma_*^2\right).$$

These upper bounds are interesting because if we allow σ_*^2 to be arbitrarily small (simulating arbitrarily small stepsizes), then analogous to the SGD case, we can get complexities of (5) even though its convergence with positive probability is not guaranteed. Below we show the complexities of (5), just like in the previous table.

Property	Complexity CSGD (with $\sigma_*^2 = \mathcal{O}(\varepsilon)$)
Convex	ε^{-1}
Strongly Convex	$ln(\varepsilon^{-1})$
Łojasiewicz with $q = 1/2$	$ln(\varepsilon^{-1})$
Łojasiewicz with $q \in (1/2, 1)$	$\varepsilon^{-(2q-1)}$

Although that (5) is a continuous model of SGD and that the upper bounds between these two dynamics is similar, it is direct to notice that the complexities of (5) are the same as those in the previous table for GD and CGD. Therefore, we can see that the $\sigma_*^2 = \mathcal{O}(\varepsilon)$ hypothesis is strong, nevertheless, it is the assumption we use to deduce complexities of (5).

Chapter 1

Preliminaries

In this Chapter, we are going to show the concepts and results necessary to handle in order to understand the objectives of this Thesis.

Let H be a real Hilbert space and consider a function $f : H \to \mathbb{R}$. Throughout this Thesis, consider $S := argmin(f) = \{x \in H : f(x) = min(f)\}$ and assume that $S \neq \varnothing$.

1.1 Convexity and Strong Convexity

f is convex if,

$$\forall \lambda \in [0,1], \forall x, y \in H, \;\; f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{1.1}$$

And it will be denoted $f \in \Gamma_0(H)$.

Proposition 1.1 *Let $f : H \to \mathbb{R}$ be differentiable. The following are equivalent:*

i) f *is convex.*

ii) $\forall x, y \in H, \;\; f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

iii) $\forall x, y \in H, \;\; \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$.

If f is twice differentiable, then the previous conditions are equivalent to

iv) $\forall x, d \in H, \;\; \langle \nabla^2 f(x) d, d \rangle \geq 0$.

f is $\mu-$ strongly convex if,

$$\forall \lambda \in [0,1], \forall x, y \in H, \;\; f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2} \lambda (1-\lambda) \|x - y\|^2. \tag{1.2}$$

And it will be denoted $f \in \Gamma_\mu(H)$.

Proposition 1.2 *Let $f : H \to \mathbb{R}$ be differentiable. The following are equivalent:*

i) f is $\mu-$strongly convex.

ii) $\forall x, y \in H,\ \ f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2$.

iii) $\forall x, y \in H,\ \ \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$.

If f is twice differentiable, then the previous conditions are equivalent to

iv) $\forall x, d \in H,\ \ \langle \nabla^2 f(x)d, d\rangle \geq \mu\|d\|^2$.

Proposition 1.3 *If $f \in \Gamma_\mu(H)$, then the set of minimizers is a singleton, i.e. there exists $x^* \in H$ such that $S = \{x^*\}$.*

Proposition 1.4 *(see [7, Appendix A]) Assume that $f \in \Gamma_\mu(H)$ is differentiable, then f satisfies the Polyak-Łojasiewicz inequality, this is*

$$\forall x \in H,\ \ 2\mu(f(x) - min(f)) \leq \|\nabla f(x)\|^2.$$

1.2 Lipschitz-Continuity of the Gradient

A function $F : H \to H$ is Lipschitz with constant L or $L-$Lipschitz if:

$$\forall x, y \in H,\ \ \|F(x) - F(y)\| \leq L\|x - y\|. \tag{1.3}$$

A function $F : H \to H$ is cocoercive with constant β or $\beta-$cocoercive if:

$$\forall x, y \in H,\ \ \langle F(x) - F(y), x - y\rangle \geq \beta\|F(x) - F(y)\|^2. \tag{1.4}$$

Proposition 1.5 *If $F : H \to H$ is $\beta-$cocoercive, then F is $\frac{1}{\beta}$-Lipschitz.*

Proof. Direct from the Cauchy-Schwarz inequality. □

If f is differentiable and ∇f is $L-$Lipschitz Continuous, then it will be denoted $f \in C_L^{1,1}(H)$.

Proposition 1.6 *Assume that $f \in C_L^{1,1}(H)$, then*

$$\forall x, y \in H,\ \ f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|x - y\|^2. \tag{1.5}$$

Corollary 1.7 *Assume that $f \in C_L^{1,1}(H)$, then*

$$\forall x \in H,\ \ f\left(x - \frac{1}{L}\nabla f(x)\right) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|^2, \tag{1.6}$$

and

$$\forall x \in H, \quad \|\nabla f(x)\|^2 \leq 2L(f(x) - min(f)). \tag{1.7}$$

Corollary 1.8 *Assume that $f \in C_L^{1,1}(H) \cap \Gamma_\mu(H)$, then $\mu \leq L$.*

Proof. Direct from combining the results of Propositions 1.4 and 1.7. □

Proposition 1.9 *Assume that $f \in C_L^{1,1}(H) \cap \Gamma_0(H)$, then*

$$\forall x, y \in H, \quad f(y) - f(x) - \langle \nabla f(y), y - x \rangle \leq -\frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2.$$

We have seen in Proposition 1.5 that cocoercive implies Lipschitz, in order to get the equivalence in the case $F = \nabla f$, we will need f to be convex.

Theorem 1.10 *(Baillon-Haddad Theorem) [8, Corollary 10] Assume that $f \in \Gamma_0(H)$ and differentiable. Then $f \in C_L^{1,1}(H)$ if, and only if, ∇f is $\frac{1}{L}$-cocoercive.*

1.3 Łojasiewicz Inequality

Let $f : H \to \mathbb{R}$ be convex, $S \neq \varnothing$ and $q \in [0, 1)$. f satisfies the Łojasiewicz Inequality with exponent q on $A \subseteq H$ if there exists a coefficient $\mu_A > 0$ such that:

$$\forall x \in A, \ \ \mu_A(f(x) - min(f))^q \leq \|\nabla f(x)\|, \tag{1.8}$$

and it will be denoted $f \in \text{Ł}^q(A)$. We will refer to this notion as Global if $A = H$. By default its coefficient will be denoted by μ.

Definition. We say that f satisfies the Bounded Łojasiewicz Inequality with exponent q if:

$$\exists x^* \in S, \forall r > 0, \ \text{ such that } f \in \text{Ł}^q(B(x^*; r)),$$

and it will be denoted $f \in \text{Ł}_b^q(H)$.

Definition. We say that f satisfies the Local Łojasiewicz Inequality if:

$$\forall x^* \in S, \exists r > 0, \ \text{ such that } f \in \text{Ł}^q(B(x^*; r))$$

and it will be denoted $f \in \text{Ł}_{loc}^q(H)$.

It is direct to notice that the Global Łojasiewicz Inequality implies the Bounded and Local Łojasiewicz Inequalities.

Moreover, under some assumptions, the Bounded and Local Łojasiewicz Inequalities are equivalent.

Proposition 1.11 *Suppose that* $\dim(H) < \infty$ *and that* S *is compact. Then, the Local Łojasiewicz Inequality is equivalent to the Bounded Łojasiewicz Inequality, in other words* $Ł^q_{loc}(H) = Ł^q_b(H)$.

Proof. • $Ł^q_{loc}(H) \supseteq Ł^q_b(H)$.

Let $f \in Ł^q_b(H)$, by the definition of $Ł^q_b(H)$ there exists $x^* \in S$ such that for all $r > 0$, $f \in Ł^q(B(x^*; r))$. Let $y^* \in S, r_0 > 0$ arbitrary and $x \in B(y^*; r_0)$, then

$$\begin{aligned} \|x - x^*\| &\leq \|x - y^*\| + \|y^* - x^*\| \\ &\leq r_0 + R, \end{aligned}$$

where $R > 0$ exists because S is bounded. So we deduce that $B(y^*; r_0) \subseteq B(x^*; r_0 + R)$ and $f \in Ł^q(B(x^*; r_0 + R))$ because $f \in Ł^q_b(H)$, this implies $f \in Ł^q(B(y^*; r_0))$ for every $y^* \in S, r_0 > 0$, then $f \in Ł^q_{loc}(H)$.

- $Ł^q_{loc}(H) \subseteq Ł^q_b(H)$.

Let $f \in Ł^q_{loc}(H)$ and $x^* \in S$ arbitrary, there exists $r_{x^*} > 0$ such that $f \in Ł^q(B(x^*; r_{x^*}))$. We have $S \subset \bigcup_{x^* \in S} B(x^*; r_{x^*})$, since S is compact, there exists finite points $\{x^*_1, \ldots, x^*_n\} \subseteq S$ such that $S \subset \bigcup_{\mathrm{i}=1}^n B(x^*_\mathrm{i}; r_\mathrm{i}) =: B_0$. Let $\{\mu_\mathrm{i}\}_{\mathrm{i}=1}^n$ such that $\mu_\mathrm{i} > 0$ is the coefficient of the Łojasiewicz Inequality on $B(x^*_\mathrm{i}; r_\mathrm{i})$, consider $\mu_{B_0} = \min_{\mathrm{i} \in [n]} \mu_\mathrm{i} > 0$, so $f \in Ł^q(B_0)$ with coefficient μ_{B_0}.

Now we set $x^* \in S$ and let $R > 0$ arbitrary, we want to prove that $f \in Ł^q(B(x^*; R))$. If R is such that $B(x^*; R) \subseteq B_0$ then the proof is trivial, so let R such that $B(x^*; R) \backslash B_0 \neq \varnothing$.

Consider

$$\hat{\mu} := \inf_{x \in cl(B(x^*;R)) \backslash B_0} \frac{\|\nabla f(x)\|}{(f(x) - min(f))^q},$$

since B_0 is an open set (union of open sets), then $cl(B(x^*; R)) \setminus B_0$ is a closed set. Moreover, it is bounded, therefore, it is compact. On the other hand, since the denominator of the fraction is non-zero (because $S \subset B_0$), the function is continuous in $cl(B(x^*; R)) \setminus B_0$, then by the Weierstrass Theorem, the infimum becomes a minimum and $\hat{\mu} \geq 0$.

If $\hat{\mu} = 0$, implies that there exists $\tilde{x} \in cl(B(x^*; R)) \setminus B_0$, such that $\|\nabla f(\tilde{x})\| = 0$. Since f is convex, this implies that $\tilde{x} \in S \subset B_0$, a contradiction. We deduce that $\hat{\mu} > 0$,

so by letting $\mu := \min\{\mu_{B_0}, \hat{\mu}\}$, we satisfy the Łojasiewicz Inequality on $B(x^*; R)$ with coefficient μ. Since R is arbitrary, we conclude that $f \in \text{Ł}_b^q(H)$.

□

Proposition 1.12 *Let $A \subset H$ such that $A \neq \varnothing$ and assume that $f \in C_L^{1,1}(H) \cap \text{Ł}^{1/2}(A)$. Then $\mu_A^2 \leq 2L$.*

Proof. Direct from using (1.7) and the definition of $\text{Ł}^{1/2}(A)$. □

Proposition 1.13 *Every $\mu-$strongly convex function satisfies the Global Łojasiewicz Inequality with coefficient $\sqrt{2\mu}$ and exponent $\frac{1}{2}$. In other words:*

$$\Gamma_\mu(H) \subset \text{Ł}^{1/2}(H).$$

Proof. Direct from Proposition 1.4. □

Another interesting aspect of the Łojasiewicz Inequality is that implies a Hölderian error bound (see [1]), this is an inequality of the form

$$f(x) - min(f) \geq \mu d(x, S)^p.$$

This concept will be useful to find convergence rates and upper bounds of some algorithms under Łojasiewicz assumptions.

Proposition 1.14 *Let $x^* \in S, r > 0, f \in \text{Ł}^q(B(x^*; r))$ and $p := \frac{1}{1-q} \geq 1$, then there exists $\tilde{\mu}_r$ such that:*

$$f(x) - min(f) \geq \tilde{\mu}_r d(x, S)^p, \quad \forall x \in B(x^*; r). \tag{1.9}$$

Moreover, if $f \in \text{Ł}^q(H)$, then (1.9) holds for every $x \in H$ with a unique $\mu > 0$.

Proof. From Proposition 1.23 with $t = 0$ and $s \to \infty$, there exists $\mu_r > 0$ such that, if we define:

$$\varphi_r(y) = \frac{y^{1-q}}{(1-q)\mu_r},$$

then

$$\|x - y^*\| \leq \varphi_r(f(x) - min(f)), \ \forall x \in B(x^*; r),$$

for some $y^* \in S$. So

$$d(x, S) \leq \|x - y^*\| \leq \frac{(f(x) - min(f))^{1-q}}{(1-q)\mu_r}, \ \forall x \in B(x^*; r).$$

By letting $\tilde{\mu}_r = (\mu_r(1-q))^{\frac{1}{1-q}}$, we obtain

$$f(x) - min(f) \geq \tilde{\mu}_r d(x, S)^p, \ \forall x \in B(x^*; r),$$

and we conclude. □

We have seen that Łojasiewicz Inequality is a property about the geometry of the function, in this sense, gradient Lipschitz fulfills the same role, since it is a property that controls the growth of the function. We would like to know what is the resulting behavior if we combine these properties in a function.

Proposition 1.15 *Let $x^* \in S, r > 0$ such that $B(x^*; r) \setminus S \neq \varnothing$, $f \in C_L^{1,1}(H) \cap \text{Ł}^q(B(x^*; r))$, then $q \geq \frac{1}{2}$.*

Proof. By definition of $\text{Ł}^q(B(x^*; r))$, there exists $\mu_r > 0$ such that Łojasiewicz Inequality holds on $B(x^*; r)$. Squaring this inequality and using (1.7), we obtain

$$\mu_r^2(f(x)) - min(f))^{2q} \leq \|\nabla f(x)\|^2 \leq 2L(f(x) - min(f)), \ \ \forall x \in B(x^*; r).$$

Since we can divide by $f(x) - min(f)$ for $x \notin S$, we deduce that

$$\frac{\mu_r^2}{2L} \leq (f(x) - min(f))^{1-2q}, \ \ \forall x \in B(x^*; r) \setminus S. \tag{1.10}$$

If $1 - 2q > 0$, let $x_0 \in B(x^*; r) \setminus S$ and $\delta := \left(\frac{\mu_r^2}{2L}\right)^{\frac{1}{1-2q}} > 0$, by (1.10), we have that $f(x_0) \geq min(f) + \delta$.

Let $g : [0, 1] \to \mathbb{R}$ be defined by

$$g(t) = f(tx^* + (1 - t)x_0).$$

It is direct that g is continuous, on the other hand, $g(0) = f(x_0) > min(f) + \frac{\delta}{2}$ and $g(1) = f(x^*) = min(f) < min(f) + \frac{\delta}{2}$, so by Bolzano's Theorem, there exists $t^* \in (0, 1)$ such that

$$g(t^*) = f(t^*x^* + (1 - t^*)x_0) = min(f) + \frac{\delta}{2},$$

so by letting $\bar{x} := t^*x^* + (1-t^*)x_0 \in B(x^*; r) \setminus S$, we conclude that there exists $\bar{x} \in B(x^*; r) \setminus S$ such that $f(\bar{x}) < min(f) + \delta$, a contradiction with (1.10). Therefore, $q \geq \frac{1}{2}$.

□

Proposition 1.16 *If $f \in C_L^{1,1}(H) \cap \text{Ł}^q(H)$ is not constant, then $q = \frac{1}{2}$.*

Proof. Let $x^* \in S$ arbitrary, since f is non-constant, $S \neq H$, so consider $r > 0$ big enough such that $B(x^*; r) \setminus S \neq \varnothing$. Since $f \in \text{Ł}^q(H)$ we obtain that $f \in \text{Ł}^q(B(x^*; r))$, so by Proposition 1.15, we deduce $q \geq \frac{1}{2}$. Squaring the definition of $\text{Ł}^q(H)$ and using (1.7), then

$$\mu^2(f(x)) - min(f))^{2q} \leq \|\nabla f(x)\|^2 \leq 2L(f(x) - min(f)), \ \ \forall x \in H.$$

If $2q - 1 > 0$ and $x \notin S$, then

$$f(x) - min(f) \leq \left(\frac{2L}{\mu^2}\right)^{\frac{1}{2q-1}}, \ \ \forall x \in H \setminus S.$$

Thus f is bounded, so by Proposition 1.14, this implies that $d(x, S)$ is bounded for all $x \in H$. Since $S \neq H$, we have a contradicition. Hence $q = \frac{1}{2}$. □

Remark. The previous Proposition tells us that asking for Gradient Lipschitz and Global Łojasiewicz Inequality is very limiting and greatly reduces its applications. So, if we want both properties in a function, we will ask for the Bounded Łojasiewicz Inequality instead.

Proposition 1.17 *Let $A \subset H, q > \frac{1}{2}$ and assume that $f \in C^{1,1}_L(H) \cap Ł^q(A)$, then $A \subseteq [f \leq R]$ for some $R \in \mathbb{R}$.*

Proof. Squaring the definition of $Ł^q(A)$ and using (1.7), then there exists $\mu_A > 0$ such that

$$\mu_A^2(f(x)) - min(f))^{2q} \leq \|\nabla f(x)\|^2 \leq 2L(f(x) - min(f)), \ \ \forall x \in A.$$

If $x \notin S$, since $q > \frac{1}{2}$, we obtain

$$f(x) - min(f) \leq \left(\frac{2L}{\mu_A^2}\right)^{\frac{1}{2q-1}}, \ \ \forall x \in A \setminus S. \tag{1.11}$$

But obviously (1.11) is satisfied if $x \in S$, so by letting $R := min(f) + \left(\frac{2L}{\mu_A^2}\right)^{\frac{1}{2q-1}}$, we conclude that

$$f(x) \leq R, \ \ \forall x \in A.$$

□

1.4 Discrete Deterministic Dynamics: Gradient Descent

Let $f : H \to \mathbb{R}$ be convex and differentiable. We consider the general problem

$$\min_{x \in H} f(x).$$

The Gradient Descent method is defined as follows: given any $x_0 \in H$,

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k), \ \ k \in \mathbb{N}. \tag{GD}$$

1.4.1 Convergence of GD

Let us start by recalling classical results about Gradient Descent when f is convex and strongly convex.

Proposition 1.18 *(See [9, Theorem 2.2]) Assume that $f \in C^{1,1}_L(H) \cap \Gamma_0(H)$, let $\gamma_k \equiv \gamma < 2/L$ and $(x_k)_{k\in\mathbb{N}}$ generated by (GD).*

- *Since $f \in \Gamma_0(H)$, then*
  - *$x_k \rightharpoonup x^* \in S$.*
  - *For every $x^* \in S$, the sequence $(\|x_k - x^*\|)_{k\in\mathbb{N}}$ is non-increasing.*
  - *For every $k \in \mathbb{N}$,*

$$f(x_k) - min(f) \leq C\frac{d(x_0, S)^2}{2\gamma k}. \tag{1.12}$$

  with

$$C = \begin{cases} 1 & \text{if } \gamma \leq \frac{1}{L}, \\ 1 + 2(\gamma L - 1)(2 - \gamma L)^{-1} & \text{otherwise.} \end{cases}$$

- *If $f \in \Gamma_\mu(H)$ ($S = \{x^*\}$), then*

$$\|x_k - x^*\|^2 \leq \theta^k \|x_0 - x^*\|^2, \ \ \forall k \in \mathbb{N}, \tag{1.13}$$

  with

$$\theta = \begin{cases} 1 - \gamma\mu & \text{if } \gamma \leq \frac{2}{\mu+L}, \\ \gamma L - 1 & \text{otherwise.} \end{cases}$$

  The minimal θ is obtained when $\gamma = \frac{2}{\mu+L}$.

  Moreover,

$$f(x_k) - min(f) \leq \frac{L}{2}\theta^k \|x_0 - x^*\|^2, \ \ \forall k \in \mathbb{N}. \tag{1.14}$$

One way to show results of (GD) under Łojasiewicz Inequality is to assume that $f \in Ł^q(B(x^*; r))$ for some $x^* \in S, r > 0$ and ask $x_0 \in B(x^*; r)$, because due to the previous

Proposition, the entire sequence will be contained in $B(x^*; r)$, therefore the Łojasiewicz Inequality will be satisfied in the whole sequence. Nevertheless, for reasons that will be explained in section (2.1.2), in order to get results of (GD) in the Łojasiewicz case, we will ask for the Bounded Łojasiewicz Inequality i.e. $f \in Ł_b^q(H)$.

Proposition 1.19 *(see [3]) Assume that $f \in C_L^{1,1}(H) \cap Ł_b^q(H)$, let $\gamma_k \equiv \gamma < 2/L$ and $(x_k)_{k\in\mathbb{N}}$ generated by (GD), then:*

- $x_k \to x^* \in S$.
- *If $q = \frac{1}{2}$, then there exists $\rho \in (0,1)$ such that:*

$$f(x_k) - min(f) \leq [f(x_0) - min(f)]\rho^k, \ \forall k \in \mathbb{N}. \tag{1.15}$$

- *If $q \in (\frac{1}{2}, 1)$, then there exists $C > 0$ such that:*

$$f(x_k) - min(f) \leq Ck^{-\frac{1}{2q-1}}, \ \forall k \in \mathbb{N}. \tag{1.16}$$

Corollary 1.20 *[Complexities of Proposition 1.19] Let $\varepsilon > 0$ arbitrary.*

- *If $q = \frac{1}{2}$ and $k = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$. Then*

$$f(x_k) - min(f) \leq \varepsilon.$$

- *If $q \in (\frac{1}{2}, 1)$ and $k = \Omega_0\left(\frac{1}{\varepsilon^{2q-1}}\right)$. Then*

$$f(x_k) - min(f) \leq \varepsilon.$$

Corollary 1.21 *Assume that $f \in C_L^{1,1}(H) \cap Ł_b^q(H)$, let $\gamma_k \equiv \gamma < 2/L$, $(x_k)_{k\in\mathbb{N}}$ generated by (GD) and $\lim_k x_k =: x^* \in S$.*

- *If $q = \frac{1}{2}$, then there exists $C > 0, \rho \in (0,1)$ such that:*

$$\|x_k - x^*\| \leq C\rho^k, \ \forall k \in \mathbb{N}. \tag{1.17}$$

- *If $q \in (\frac{1}{2}, 1)$, then there exists $C > 0$ such that:*

$$\|x_k - x^*\| \leq Ck^{-\frac{1-q}{2q-1}}, \ \forall k \in \mathbb{N}. \tag{1.18}$$

Proof. Let $f \in C_L^{1,1}(H) \cap Ł_b^q(H)$, [3] assures us that there exists $C_1, C_2 > 0$ such that

$$\|x_{k+1} - x^*\| \leq C_1(f(x_k) - min(f))^{1/2} + C_2(f(x_k) - min(f))^{1-q}, \ \forall k \in \mathbb{N}.$$

The result is direct from combining the previous inequality with the results of Proposition 1.19. □

1.5 Continuous Deterministic Dynamics: Continuous Gradient Descent

Consider $f \in \Gamma_0(H)$. In order to minimize the function f over H, we can consider the underlying continuous dynamic of Gradient Descent algorithm, which is, given a $x_0 \in H$:

$$\begin{aligned} &\dot{x}(t) = -\nabla f(x(t)), \ \ t > 0, \\ &x(0) = x_0. \end{aligned} \tag{CGD}$$

1.5.1 Convergence of CGD

Let us recall classic results about Continuous Gradient Descent when f is convex and strongly convex.

Proposition 1.22 *Let $x(t)$ be a solution of (CGD) and $S \neq \varnothing$.*

- *Then $x(t) \rightharpoonup x^* \in S$.*
- *For every $x^* \in S$, the function $t \mapsto \|x(t) - x^*\|$ is non-increasing.*
- *Since $f \in \Gamma_0(H)$, then*

$$f(x(t)) - min(f) \leq \frac{d(x_0, S)^2}{2t}, \forall t > 0. \tag{1.19}$$

- *If $f \in \Gamma_\mu(H)$ $(S = \{x^*\})$, then*

$$\|x(t) - x^*\|^2 \leq \|x_0 - x^*\|^2 e^{-\mu t}, \ \ \forall t \geq 0. \tag{1.20}$$

Proof. The proofs of the first two items can be found in [10]. The proofs of the last two can be found in [11]. □

In the discrete case, Proposition 1.19 tells us that the hypothesis $f \in Ł_b^q(H)$ assures us that the sequence $(x_k)_{k\in\mathbb{N}}$ converges strongly to a point in S. The following Proposition will assure us the same in the continuous case.

Proposition 1.23 *(see [1, Theorem 27]) Let $x^* \in S, r > 0$ and assume that $f \in Ł^q(B(x^*; r))$. Consider $\chi_x(t)$ a solution of (CGD) such that $\chi_x(0) = x$ and*

$$\varphi(y) = \frac{y^{1-q}}{(1-q)}.$$

Then there exists $\mu_r > 0$, for every $0 \leq t < s$ such that:

$$\|\chi_x(t) - \chi_x(s)\| \leq \frac{\varphi\left(f(\chi_x(t)) - min(f)\right) - \varphi\left(f(\chi_x(s) - min(f)\right)}{\mu_r}, \quad \forall x \in B(x^*; r). \tag{1.21}$$

Moreover, if $x \in B(x^*; r)$, then $\chi_x(t)$ converges strongly to a minimizer of f.

PROOF. (Proof adapted from [1, Theorem 27]). Take $x \in B(x^*;r)$ and $0 \leq t < s$. Let $\mu_r > 0$ the coefficient of the Łojasiewicz Inequality on $B(x^*;r)$ and define $\varphi_r(y) = \frac{\varphi(y)}{\mu_r}$. Observe that

$$\begin{aligned}\varphi_r\left((f(\chi_x(t)) - min(f)\right) - \varphi_r\left((f(\chi_x(s)) - min(f)\right) &= \int_s^t \frac{d}{d\tau}\varphi_r\left((f(\chi_x(\tau) - min(f)\right))\right) d\tau \\ &= \int_t^s \varphi_r'\left(f(\chi_x(\tau) - min(f))\right) \|\dot{\chi_x}(\tau)\|^2 d\tau.\end{aligned}$$

Since $\chi_x(\tau) \in B(x^*;r)$ (see [10]), it follows that for all $\tau \geq 0$,

$$1 \leq \varphi_r'\left(f(\chi_x(\tau))\right) \|\nabla f(\chi_x(\tau))\| = \varphi_r'\left(f(\chi_x(\tau))\right) \|\dot{\chi_x}(\tau)\|,$$

where the first inequality is direct from the definition of $Ł^q(B(x^*;r))$. Thus, multiplying by $\|\dot{\chi_x}(\tau)\|$, integrating from t to s and using the fact that

$$\|\chi_x(t) - \chi_x(s)\| \leq \int_t^s \|\dot{\chi_x}(\tau)\| d\tau,$$

we deduce that

$$\|\chi_x(t) - \chi_x(s)\| \leq \varphi_r\left(f(\chi_x(t)) - min(f)\right) - \varphi_r\left(f(\chi_x(s) - min(f)\right), \quad \forall x \in B(x^*;r).$$

Moreover, since $f(\chi_x(t)) \to min(f)$, we conclude that for every $x \in B(x^*;r)$, the function $t \mapsto \chi_x(t)$ has the Cauchy property as $t \to \infty$. □

In addition, we will now show convergence rates for (CGD) in the Łojasiewicz case.

Proposition 1.24 *(see [3]) Assume that $f \in Ł_b^q(H)$ and let $x(t)$ be a solution of (CGD):*

- *If $q \in [0, \frac{1}{2})$, then there exists $\mu > 0$ such that $x(t)$ reaches S in at most*

$$t^* = \frac{(f(x_0) - min(f))^{1-2q}}{\mu^2(1-2q)}. \tag{1.22}$$

- *If $q = \frac{1}{2}$, then there exists $\mu > 0$ such that:*

$$f(x(t)) - min(f) \leq (f(x_0) - min(f))e^{-\mu^2 t}, \ \forall t \geq 0. \tag{1.23}$$

- *If $q \in (\frac{1}{2}, 1)$, then there exists $\mu > 0$ and $\eta = 2q - 1$ such that:*

$$f(x(t)) - min(f) \leq [(f(x_0) - min(f))^{-\eta} + \mu^2 \eta t]^{\frac{-1}{\eta}}, \ \forall t \geq 0. \tag{1.24}$$

Corollary 1.25 (Complexities of Proposition 1.24) *Let $\varepsilon > 0$ arbitrary.*

- *If $q \in (0, \frac{1}{2})$ and $t = \Omega_0(1)$. Then*

$$f(x(t)) - min(f) \leq \varepsilon.$$

- *If $q = \frac{1}{2}$ and $t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$. Then*

$$f(x(t)) - min(f) \leq \varepsilon.$$

- *If $q \in (\frac{1}{2}, 1)$ and $k = \Omega_0\left(\frac{1}{\varepsilon^{2q-1}}\right)$. Then*

$$f(x(t)) - min(f) \leq \varepsilon.$$

Chapter 2

Stochastic Dynamics

In this chapter we will focus on two main topics: First, we will examine discrete dynamics, studying different versions of Stochastic Gradient Descent. Then we will focus on continuous dynamics, where inspired by a continuous-time model of Mini-Batch SGD, we will lay out the necessary foundations of stochastic processes and Stochastic Differential Equations in order to study a continuous-time model of SGD in the next chapter.

2.1 Discrete Stochastic Dynamics: Stochastic Gradient Descent

We consider the following optimization problem

$$\min_{x\in H}\left[f(x)=\frac{1}{n}\sum_{\mathrm{i}=1}^{n}f_{\mathrm{i}}(x)\right], \tag{2.1}$$

where f is convex, each $f_{\mathrm{i}}:H\to\mathbb{R}$ is differentiable and ∇f_{i} is $L_{\mathrm{i}}-$Lipschitz. There are different algorithms to solve this problem, obviously we can use the Gradient Descent, but the computational cost of calculating the gradient at each iteration grows with n (which can be remarkably big). One of the most classical algorithm to manage this problem is the Vanilla Stochastic Gradient Descent (V-SGD) (see [12]), which is: given $x_0\in H$,

$$x_{k+1}=x_k-\gamma_k\nabla f_{\mathrm{i}_k}(x_k),\ \ k\in\mathbb{N}, \tag{V-SGD}$$

where $(\mathrm{i}_k)_{k\in\mathbb{N}}\overset{\text{iid}}{\sim}Unif[n]$. The idea behind this algorithm is that the estimator is unbiased from the gradient, i.e. $\mathbb{E}[\nabla f_j(\cdot)]=\nabla f(\cdot)$ (with $j\sim Unif[n]$), so the computational cost of calculating the gradient will not be a problem, naturally we omit a lot of information of the function at every iteration, so the convergence to a minimum will be slower than the Gradient Descent. For instance, if $f_{\mathrm{i}}\in C^{1,1}_{L_{\mathrm{i}}}(H)$ for each $\mathrm{i}\in[n]$, $\gamma_k=\frac{C}{k^{2/3}}$, for some $C>0$ (see [4, Theorem 4]) and let $(x_k)_{k\in\mathbb{N}}$ generated by (V-SGD), then

$$\mathbb{E}[f(x_k)]-min(f)=\mathcal{O}\left(\frac{1}{k^{1/3}}\right).$$

And we already know that if $f \in C_L^{1,1}(H)$ and $\gamma < \frac{2}{L}$, then by Proposition 1.18, $(x_k)_{k\in\mathbb{N}}$ generated by (GD) satisfies

$$f(x_k) - min(f) = \mathcal{O}\left(\frac{1}{k}\right).$$

So we want a trade-off between the convergence rate of Gradient Descent and the "cheap" computational cost of Stochastic Gradient Descent, this can be achieved with the Mini-Batch SGD, which is: given $x_0 \in H$,

$$x_{k+1} = x_k - \gamma_k \left(\frac{1}{b}\sum_{\mathrm{i}\in I_k} \nabla f_{\mathrm{i}}(x_k)\right), \ \ k \in \mathbb{N}. \tag{MB-SGD}$$

Where $b \in [n]$ is fixed and it is called the batch size, $(I_k)_{k\in\mathbb{N}} \overset{\text{iid}}{\sim} \binom{[n]}{b}$. If $b = 1$, we obtain (V-SGD) and if $b = n$, we obtain (GD), so the choice of b is very important and we must keep in mind the aforementioned trade-off. Usually the batch size satisfies $b << n$.

2.1.1 Stochastic Gradient Descent: Sampling Vectors

We can extend the previous ideas with the following approach (see [13]).

Definition. We say that a random vector $s \in \mathbb{R}^n_+$ drawn from some distribution $\mathcal{D}$ is a sampling vector if

$$\mathbb{E}_{\mathcal{D}}[s_{\mathrm{i}}] = 1, \quad \forall \mathrm{i} \in [n].$$

Let $f_s(x) := \frac{1}{n}\sum_{\mathrm{i}=1}^n s_{\mathrm{i}} f_{\mathrm{i}}(x)$. It is clear that $f(\cdot) = \mathbb{E}_{\mathcal{D}}[f_s(\cdot)]$ and $\mathbb{E}_{\mathcal{D}}[\nabla f_s(\cdot)] = \nabla f(\cdot)$. The proposed algorithm, which we will call (SGD) from now on, is as follows: given any $x_0 \in H$,

$$x_{k+1} = x_k - \gamma_k \nabla f_{s^k}(x_k), \ \ k \in \mathbb{N}, \tag{SGD}$$

where $s^k \overset{\text{iid}}{\sim} \mathcal{D}$.

Example: Let $(e_{\mathrm{i}})_{\mathrm{i}=1}^n$ be the canonical basis of $\mathbb{R}^n$. If we consider the random vector s such that:

- $\mathbb{P}(s = ne_{\mathrm{i}}) = \frac{1}{n}$ for every $\mathrm{i} \in [n]$, we can check that s is a sampling vector and $f_s \sim f_j$ (with $j \sim Unif[n]$), so we recover (V-SGD).
- $\mathbb{P}(s = \frac{n}{b}\sum_{\mathrm{i}\in I} e_{\mathrm{i}}) = \frac{1}{\binom{n}{b}}$ for every $I \in \binom{[n]}{b}$, we can check that s is a sampling vector and

$$f_s \sim \frac{1}{b}\sum_{\mathrm{i}\in J} f_{\mathrm{i}} \text{ (with } J \sim Unif\binom{[n]}{b}),$$

  so we recover (MB-SGD).

Remark. Do not confuse $\binom{[n]}{b}$ (subsets of $[n]$ of length b) with $\binom{n}{b}$ (a binomial coefficient).

In order to find some convergence rates or upper bounds of (SGD), we will need some structure over the new function f_s.

Definition. We say f is $\mathcal{L}-smooth$ (in expectation) with respect to (w.r.t.) distribution $\mathcal{D}$ if there exists $\mathcal{L} > 0$ such that for:

$$\forall x, y \in H, \quad \frac{1}{2\mathcal{L}}\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_s(x) - \nabla f_s(y)\|^2\right] \leq f(x) - f(y) - \langle \nabla f(y), x - y\rangle. \tag{2.2}$$

This structure is very related with the hypothesis ∇f is Lipschitz continuous, in fact, the following Proposition is analogous to Proposition 1.9.

Proposition 2.1 *Assume that f is $\mathcal{L}-smooth$ w.r.t. $\mathcal{D}$, then*

$$\forall x, y \in H, \quad f(y) - f(x) - \langle \nabla f(y), y - x\rangle \leq -\frac{1}{2\mathcal{L}}\|\nabla f(y) - \nabla f(x)\|^2.$$

Proof. Direct from Jensen's inequality and the definition of $\mathcal{L}-smooth$. □

Corollary 2.2 *Assume that f is $\mathcal{L}-smooth$ w.r.t. $\mathcal{D}$, then ∇f is $\mathcal{L}-$Lipschitz continuous.*

Proof. Consider the inequality of Proposition 2.1. Replacing x for y and vice versa, and adding the two inequalities, we obtain that ∇f is $\frac{1}{\mathcal{L}}$-cocoercive, and this in turn implies that ∇f is $\mathcal{L}$-Lipschitz continuous by Proposition 1.5 with $F = \nabla f$. □

Now we have proved that f is $\mathcal{L}-smooth$ w.r.t. $\mathcal{D}$ is a particular case of the condition ∇f is Lipschitz continuous.

Corollary 2.3 *Assume that $f \in C_L^{1,1}(H)$ is $\mathcal{L}-smooth$ w.r.t. $\mathcal{D}$, then $\mathcal{L} \geq L$.*

Assumption: Let $x^* \in S$. We assume that

$$\sigma^2 := \mathbb{E}_{\mathcal{D}}\left[\|\nabla f_s(x^*)\|^2\right] < \infty.$$

Remark.

$$\sigma^2 = \mathbb{V}_{\mathcal{D}}[\nabla f_s(x^*)],$$

since $\mathbb{E}_{\mathcal{D}}[\nabla f_s(x^*)] = \nabla f(x^*) = 0$.

This is a weak assumption (see [13]) and should be seen as an assumption on $\mathcal{D}$ rather than on f. For instance, if $\mathbb{P}(s \in \mathbb{R}^n_+) = 1$ and

$$\mathbb{E}_{\mathcal{D}}\left[s_{\mathrm{i}} \sum_{j=1}^{n} s_j\right] < \infty, \quad \forall \mathrm{i} \in [n].$$

Then σ^2 is finite.

Remark. In general $\sigma^2 > 0$, nevertheless, we could have the "interpolation property", i.e. $\sigma^2 = 0$ (see [5]), satisfied for instance, if each f_{i} attains its minimum at x^*.

Lemma 2.4 *Assume that f is $\mathcal{L}-$smooth w.r.t. $\mathcal{D}$. Then*

$$\forall x \in H, \quad \mathbb{E}_{\mathcal{D}}[\|\nabla f_s(x)\|^2] \leq 4\mathcal{L}(f(x) - min(f)) + 2\sigma^2.$$

Proof. Let $x \in H$ and $x^* \in S$ arbitrary, we have

$$\|\nabla f_s(x)\|^2 = \|\nabla f_s(x) - \nabla f_s(x^*) + \nabla f_s(x^*)\|^2 \leq 2\|\nabla f_s(x) - \nabla f_s(x^*)\|^2 + 2\|\nabla f_s(x^*)\|^2.$$

Then taking expectation with respect to $\mathcal{D}$ on both sides of the inequality, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_s(x)\|^2\right] &\leq 2\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_s(x) - \nabla f_s(x^*)\|^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_s(x^*)\|^2\right] \\ &\leq 4\mathcal{L}\left[f(x) - min(f)\right] + 2\sigma^2,\end{aligned}$$

where the last inequality was obtained using the definition of $\mathcal{L}-$smooth and the definition of σ^2. □

2.1.2 Discrete Stochastic Results

In order to state results of (SGD), there will be cases when the assumption $\sigma^2 < \infty$ will not be enough (see Propositions 2.5,2.13 and 2.17), in this cases we will assume a stronger assumption, this is

$$\bar{\sigma}^2 = \sup_k \mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2) < \infty.$$

In fact, this assumption is so strong that we will not need the function to be $\mathcal{L}-$smooth to obtain results. Although this quantity is not easy to estimate in practice, it will be useful theoretically.

Proposition 2.5 *(see [4, Theorem 7]) Let $\gamma_k = \frac{C}{k^\alpha}, \alpha \in (0,1]$, $(x_k)_{k\in\mathbb{N}}$ generated by (SGD) and assume that $\bar{\sigma}^2 < \infty$, then*

$$\mathbb{E}(f(\bar{x}_k)) - min(f) = \begin{cases} \mathcal{O}\left(\frac{1}{k^\alpha}\right) & \text{if} \quad \alpha \in (0, \frac{1}{2}), \\ \mathcal{O}\left(\frac{log(k)}{\sqrt{k}}\right) & \text{if} \quad \alpha = \frac{1}{2}, \\ \mathcal{O}\left(\frac{1}{k^{1-\alpha}}\right) & \text{if} \quad \alpha \in (\frac{1}{2}, 1). \end{cases}$$

Proposition 2.6 *(see [14],[15]) Assume that $f \in \Gamma_\mu(H)$ is $\mathcal{L}-$smooth, $x^* \in S$, let $(x_k)_{k\in\mathbb{N}}$ generated by (SGD) and $\gamma_k = \frac{C}{k}$, for some $C > \frac{1}{2\mu}$. Then*

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) = \mathcal{O}\left(\frac{1}{k}\right).$$

In order to know the quality of these results, we need to know what are the minimum convergence rates that (SGD) can get, for this we must take into account the following two Propositions.

Proposition 2.7 *(see [16]) Assume that $f \in C_L^{1,1}(H) \cap \Gamma_0(H)$, let $(x_k)_{k\in\mathbb{N}}$ generated by (SGD). Then*

$$\mathbb{E}(f(x_k)) - min(f) = \Omega\left(\frac{1}{\sqrt{k}}\right).$$

Proposition 2.8 *(see [17]) Assume that $f \in C_L^{1,1}(H) \cap \Gamma_\mu(H), S = \{x^*\}$, let $(x_k)_{k\in\mathbb{N}}$ generated by (SGD). Then*

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) = \Omega\left(\frac{1}{k}\right).$$

In conclusion, the convergence rate of Proposition 2.5 is optimal when $\alpha = \frac{1}{2}$ (up to a logarithmic term) and the result of Proposition 2.6 is optimal as well.

In a different direction, in general if we choose a constant stepsize, the algorithm (SGD) *does not converge*. Nevertheless, we can obtain remarkable upper bounds, this concept can be interpreted as a guarantee that the sequence will end up in a ball centered on a point of S and with some radius (which we will try to make small). Although an upper bound is not a convergence rate, one of the interesting aspects behind this concept is that we can get a complexity of the algorithm even though it *does not converge*.

Proposition 2.9 *Assume that $f \in \Gamma_0(H)$ is $\mathcal{L}-$smooth. Let $\gamma_k \equiv \gamma \in \left(0, \frac{1}{2\mathcal{L}}\right)$, $(x_k)_{k\in\mathbb{N}}$ generated by (SGD),*

- *Since $f \in \Gamma_0(H)$ then*

$$\mathbb{E}(f(\bar{x}_k)) - min(f) \leq \frac{d(x_0, S)^2}{2\gamma(1 - 2\mathcal{L}\gamma)k} + \frac{\sigma^2\gamma}{1 - 2\mathcal{L}\gamma}, \ \forall k \in \mathbb{N}. \tag{2.3}$$

- *If $f \in \Gamma_\mu(H)$. Then*

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) \leq (1 - \gamma\mu)^k\|x_0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}, \ \forall k \in \mathbb{N}. \tag{2.4}$$

We can see that we obtain upper bounds which are of the form of the convergence rates in the deterministic case (see Proposition 1.18) plus a constant depending on the variance σ^2 and the stepsize γ. So, with the purpose of making the constant term arbitrarily small in order to recover convergence rates, one idea would be to decrease σ^2 to zero. In this sense, there are stochastic algorithms, such as SVRG and SAGA (see [6]) that use variance reduction techniques and recover the same convergence rates as in Proposition 1.18. Other

idea would be to make γ arbitrarily small, although this does not ensure us we will recover the original convergence rates, it leads us to get complexities of Proposition 2.9.

In order to get their complexity rates, we will need a technical lemma.

Lemma 2.10 *Let $\alpha > 0$, then*

$$\frac{ln(x)}{ln(1-x^\alpha)} \leq \frac{ln\left(\frac{1}{x}\right)}{x^\alpha}, \ \ \forall x \in (0,1).$$

Proof. Consider the function $h : (0,1) \to \mathbb{R}$ such that $h(x) = \left(1 + \frac{x}{1-x}\right)^{1/x}$. Taking derivative, it is direct to verify that h is increasing, moreover $\lim_{x\to 0} h(x) = e$ due to the definition of e. Therefore

$$h(x) \geq e, \ \ \forall x \in (0,1).$$

Let $x \in (0,1)$ arbitrary, then $h(x^\alpha) \geq e$, i.e.

$$e \leq \left(1 + \frac{x^\alpha}{1-x^\alpha}\right)^{1/x^\alpha}.$$

Taking natural logarithm at both sides, we observe that

$$\begin{aligned} 1 &\leq \frac{1}{x^\alpha} ln\left(\frac{1}{1-x^\alpha}\right) \\ \iff 1 &\leq -\frac{1}{x^\alpha} ln(1-x^\alpha) \\ \iff \frac{1}{ln(1-x^\alpha)} &\geq -\frac{1}{x^\alpha}. \end{aligned}$$

Then multiplying at both sides by $ln(x)$, we obtain

$$\frac{ln(x)}{ln(1-x^\alpha)} \leq \frac{ln\left(\frac{1}{x}\right)}{x^\alpha},$$

and we conclude.

□

Corollary 2.11 *[Complexities of Proposition 2.9] Let $\varepsilon \in (0,1)$ arbitrary.*

- *Since $f \in \Gamma_0(H)$, if $\gamma = \mathcal{O}(\varepsilon)$ and $k = \Omega_0\left(\frac{1}{\varepsilon^2}\right)$. Then*

$$\mathbb{E}(f(\bar{x}_k)) - min(f) \leq \varepsilon. \tag{2.5}$$

- *If $f \in \Gamma_\mu(H)$, $\gamma = \mathcal{O}(\varepsilon)$ and $k = \tilde{\Omega}_0\left(\frac{1}{\varepsilon}\right)$. Then*

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) \leq \varepsilon. \tag{2.6}$$

PROOF. • In order to guarantee $\mathbb{E}(f(\bar{x}_k)) - min(f) \leq \varepsilon$, we will require that

$$\frac{d(x_0, S)^2}{2\gamma(1 - 2\mathcal{L}\gamma)k} \leq \frac{\varepsilon}{2} \text{ and } \frac{\sigma^2\gamma}{1 - 2\mathcal{L}\gamma} = \frac{\varepsilon}{2}.$$

The second condition imposes $\gamma = \frac{\varepsilon}{2(\sigma^2 + \mathcal{L}\varepsilon)}$, moreover since $\sigma^2 > 0$, then $\frac{\varepsilon}{2(\sigma^2+\mathcal{L}\varepsilon)} < \frac{1}{2\mathcal{L}}$. Therefore we have

$$\gamma(1 - 2\mathcal{L}\gamma) = \frac{\varepsilon\sigma^2}{2(\sigma^2 + \mathcal{L}\varepsilon)^2}.$$

The first condition and the previous equation are equivalent to

$$\begin{aligned} \frac{d(x_0, S)^2}{2\gamma(1 - 2\mathcal{L}\gamma)k} &\leq \frac{\varepsilon}{2} \\ \iff k &\geq \frac{d(x_0, S)^2}{\gamma(1 - 2\mathcal{L}\gamma)\varepsilon} \\ \iff k &\geq \frac{2(\sigma^2 + \mathcal{L}\varepsilon)^2 d(x_0, S)^2}{\varepsilon^2\sigma^2}. \end{aligned}$$

So, if we take a particular $k = \Omega_0\left(\frac{1}{\varepsilon^2}\right)$, we can satisfy the last inequality and we conclude.

• In order to guarantee $\mathbb{E}\left(\|x_k - x^*\|^2\right) \leq \varepsilon$, we will require that

$$(1 - \gamma\mu)^k\|x_0 - x^*\|^2 \leq \frac{\varepsilon}{2} \text{ and } \frac{2\gamma\sigma^2}{\mu} = \frac{\varepsilon}{2}.$$

The second condition imposes $\gamma = \frac{\mu\varepsilon}{4\sigma^2}$, and we will assume that ε is small enough such that $\frac{\mu\varepsilon}{4\sigma^2} < \frac{1}{2\mathcal{L}}$. The first condition and the condition on γ are equivalent to

$$\begin{aligned} (1 - \gamma\mu)^k\|x_0 - x^*\|^2 &\leq \frac{\varepsilon}{2} \\ \iff k\, ln(1 - \gamma\mu) &\leq ln\left(\frac{\varepsilon}{2\|x_0 - x^*\|^2}\right) \\ \iff k &\geq \frac{ln\left(\frac{\varepsilon}{2\|x_0 - x^*\|^2}\right)}{ln(1 - \gamma\mu)} \\ \iff k &\geq \frac{ln\left(\frac{\varepsilon}{2\|x_0 - x^*\|^2}\right)}{ln(1 - \frac{\mu^2}{4\sigma^2}\varepsilon)}. \end{aligned}$$

And taking a particular $k = \Omega_0\left(\frac{ln(\varepsilon)}{ln(1-\varepsilon)}\right)$, we can satisfy the last inequality and we could conclude a complexity rate. Nevertheless, since $\varepsilon \in (0, 1)$, then considering Lemma 2.10 with $\alpha = 1$, we have that

$$\frac{ln(\varepsilon)}{ln(1 - \varepsilon)} \leq \frac{ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon}, \quad \forall\varepsilon \in (0, 1).$$

So by taking a particular $k = \Omega_0\left(\frac{ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon}\right) =: \tilde{\Omega}_0\left(\frac{1}{\varepsilon}\right)$, we can conclude that

$$f(x_k) - min(f) \leq \varepsilon.$$

□

Before we start getting results from (SGD) in the Łojasiewicz case, we will need the following Lemma.

Lemma 2.12 *Assume that $f \in C_L^{1,1}(H) \cap \Gamma_0(H)$, let $\gamma_k \equiv \gamma$ and $(x_k)_{k\in\mathbb{N}}$ generated by (SGD). Then*

$$\mathbb{E}(f(x_{k+1})) - \mathbb{E}(f(x_k)) \leq -\gamma \mathbb{E}(\|\nabla f(x_k)\|^2) + \frac{L}{2}\gamma^2 \mathbb{E}(\mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2)), \ \ \forall k \in \mathbb{N}.$$

Proof. Let $k \in \mathbb{N}$ arbitrary, using the Proposition 1.6 with $y = x_{k+1}$ and $x = x_k$, we obtain

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Recalling that $x_{k+1} - x_k = -\gamma \nabla f_{s^k}(x_k)$, we have

$$f(x_{k+1}) - f(x_k) \leq -\gamma \langle \nabla f(x_k), \nabla f_{s^k}(x_k) \rangle + \frac{L\gamma^2}{2}\|\nabla f_{s^k}(x_k)\|^2.$$

Now we take conditional expectation at both sides, conditioning on x_k (formally, on the $\sigma-$algebra of the sequence until x_k) which will be denoted $\mathbb{E}[\cdot|x_k]$. Then

$$\mathbb{E}[f(x_{k+1}) - f(x_k)|x_k] \leq -\gamma \mathbb{E}[\langle \nabla f(x_k), \nabla f_{s^k}(x_k) \rangle | x_k] + \frac{L\gamma^2}{2}\mathbb{E}[\|\nabla f_{s^k}(x_k)\|^2 | x_k].$$

Since $\nabla f(x_k)$ is constant under the conditional expectation

$$\mathbb{E}[f(x_{k+1}) - f(x_k)|x_k] \leq -\gamma \langle \nabla f(x_k), \mathbb{E}[\nabla f_{s^k}(x_k)|x_k] \rangle + \frac{L\gamma^2}{2}\mathbb{E}[\|\nabla f_{s^k}(x_k)\|^2 | x_k].$$

And the conditional expectations of the right hand side coincide with $\mathbb{E}_{\mathcal{D}}$ since $s^k \sim \mathcal{D}$ is the only stochastic term, therefore we can omit the dependence on k of this term, then using that $\mathbb{E}_{\mathcal{D}}(\nabla f_s(x_k)) = \nabla f(x_k)$, we deduce that

$$\mathbb{E}[f(x_{k+1}) - f(x_k)|x_k] \leq -\gamma \|\nabla f(x_k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2).$$

Then, taking expectation at both sides, by the law of total expectation, we conclude that

$$\mathbb{E}(f(x_{k+1})) - \mathbb{E}(f(x_k)) \leq -\gamma \mathbb{E}(\|\nabla f(x_k)\|^2) + \frac{L}{2}\gamma^2 \mathbb{E}(\mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2)).$$

□

In the discrete deterministic case, one way to show results under Łojasiewicz Inequality (as said before) is to assume that $f \in \text{Ł}^q(B(x^*; r))$ for some $x^* \in S, r > 0$ and $x_0 \in B(x^*; r)$, because due to Proposition 1.18, the entire sequence will be contained in $B(x^*; r)$, therefore the Łojasiewicz Inequality will be satisfied in the whole sequence. Nevertheless, in the discrete stochastic case, the initial point does not necessarily localize the sequence. So we are going to

assume that the sequence is bounded a.s. along with the Bounded Łojasiewicz Inequality in order to get the Łojasiewicz Inequality satisfied in the whole sequence. In an attempt to unify the hypotheses over the objective function between the Deterministic and Stochastic Case we decided to opt for the approach of asking for the Bounded Łojasiewicz Inequality property.

Proposition 2.13 *Assume that* $f \in C_L^{1,1}(H) \cap \text{Ł}_b^{1/2}(H)$, *let* $\gamma_k \equiv \gamma \in \left(0, \frac{1}{2L}\right)$, $(x_k)_{k\in\mathbb{N}}$ *generated by (SGD), assume that:*

- $(x_k)_{k\in\mathbb{N}}$ *is bounded a.s..*
- $\bar{\sigma}^2 = \sup_k \mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2) < \infty$.

Then there exists $\mu > 0$ *such that:*

$$\mathbb{E}(f(x_k)) - min(f) \leq (1 - \gamma\mu^2)^k [f(x_0) - min(f)] + \frac{L\bar{\sigma}^2\gamma}{2\mu^2}, \ \forall k \in \mathbb{N}. \tag{2.7}$$

Proof. Let $a_k := \mathbb{E}(f(x_k)) - min(f)$ and $\mathcal{K}$ a bounded set such that $(x_k)_{k\in\mathbb{N}} \subseteq \mathcal{K}$ a.s., let $\mu > 0$ be the coefficient that exists because f satisfies the Łojasiewicz Inequality on $\mathcal{K}$. Using Lemma 2.12 and $f \in \text{Ł}_b^{1/2}(H)$, then

$$a_{k+1} - a_k \leq -\gamma\mu^2 a_k + \frac{L}{2}\gamma^2 \mathbb{E}(\mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2)).$$

Using the hypothesis of $\bar{\sigma}^2$, we obtain

$$\begin{aligned} a_{k+1} - a_k &\leq -\gamma\mu^2 a_k + \frac{L}{2}\gamma^2\bar{\sigma}^2 \\ \iff a_{k+1} &\leq (1 - \gamma\mu^2)a_k + \frac{L}{2}\gamma^2\bar{\sigma}^2. \end{aligned}$$

Then, unrolling this recurrence, we deduce that

$$a_{k+1} \leq (1 - \gamma\mu^2)^{k+1} a_0 + \frac{L\bar{\sigma}^2\gamma}{2\mu^2}.$$

□

Corollary 2.14 (Complexity of Proposition 2.13) *Let* $\varepsilon_0 > 0$ *and* $\varepsilon \in (0, \varepsilon_0)$ *arbitrary. If* $\gamma = \mathcal{O}(\varepsilon)$ *and* $k = \tilde{\Omega}_0\left(\frac{1}{\varepsilon}\right)$. *Then*

$$\mathbb{E}\left(f(x_k)\right) - min(f) \leq \varepsilon. \tag{2.8}$$

Proof. Analogous to the proof of the strongly convex case in Corollary 2.11. □

The following Proposition shows a result of (SGD) in the Łojasiewicz case that just require the assumption $\sigma^2 < \infty$ and not $\bar{\sigma}^2 < \infty$.

Proposition 2.15 *Assume that $f \in C_L^{1,1}(H) \cap \text{Ł}_b^{1/2}(H)$ and $\mathcal{L}-$smooth, let $\gamma_k \equiv \gamma$, $(x_k)_{k\in\mathbb{N}}$ generated by (SGD) and assume that:*

- *$(x_k)_{k\in\mathbb{N}}$ is bounded a.s..*
- *γ is small enough.*

Then there exists $\mu > 0$ such that, if we let $r := \gamma(\mu^2 - 2\gamma L\mathcal{L}) \in (0, \frac{1}{2}]$, then

$$\mathbb{E}(f(x_k)) - min(f) \leq (1-r)^k[f(x_0) - min(f)] + \frac{L\gamma\sigma^2}{\mu^2 - 2\gamma L\mathcal{L}}, \ \forall k \in \mathbb{N}. \tag{2.9}$$

Remark. The condition of "γ is small enough" is related to the coefficient of Łojasiewicz of the set where the sequence is contained, explicitly if $\mu > 0$ is the mentioned coefficient, then $\gamma < \frac{\mu^2}{2L\mathcal{L}}$. Since the coefficient depends on the sequence and this in turn depends on the stepsize, we cannot put the condition of the stepsize directly in the statement.

Proof. Let $a_k := \mathbb{E}(f(x_k)) - min(f)$ and $\mathcal{K}$ a bounded set such that $(x_k)_{k\in\mathbb{N}} \subseteq \mathcal{K}$ a.s., let $\mu > 0$ be the coefficient that exists because f satisfies the Łojasiewicz Inequality on $\mathcal{K}$. Using Lemma 2.12 and $f \in \text{Ł}_b^{1/2}(H)$, then

$$a_{k+1} - a_k \leq -\gamma\mu^2 a_k + \frac{L}{2}\gamma^2\mathbb{E}(\mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2)).$$

Moreover, using Lemma 2.4, we obtain

$$\begin{aligned} &a_{k+1} - a_k \leq -\gamma\mu^2 a_k + \frac{L}{2}\gamma^2(4\mathcal{L}a_k + 2\sigma^2) \\ &\iff a_{k+1} \leq (1-r)a_k + L\gamma^2\sigma^2. \end{aligned}$$

Then, unrolling this recurrence, we deduce that

$$a_{k+1} \leq (1-r)^{k+1}a_0 + \frac{L\gamma\sigma^2}{\mu^2 - 2\gamma L\mathcal{L}}.$$

Where $r > 0$ because γ is small enough and $r \leq \frac{1}{2}$ comes from the fact that $\mu^4 \leq 4L\mathcal{L}$ (by Proposition 1.12) and viewing r as a quadratic equation over γ. □

We are going to consider (SGD) when the objetive function is in $C_L^{1,1}(H) \cap \text{Ł}_b^q(H)$ for $q \in (\frac{1}{2}, 1)$, in order to get results in this case we will need the following Lemma.

Lemma 2.16 *Consider the recurrence, given $y_0 > 0$:*

$$y_{k+1} \leq y_k - ay_k^b + c,$$

where $a, c > 0$, $b > 1$ and $\Delta := ab\left(\frac{c}{a}\right)^{1-\frac{1}{b}} < 1$. Then

$$y_k \leq (1-\Delta)^k y_0 + \left(\frac{c}{a}\right)^{\frac{1}{b}}, \ \forall k \in \mathbb{N}.$$

PROOF. Let $\phi(x) = ax^b - c$, which is convex (since $b > 1$) and has one root $\bar{x} = \left(\frac{c}{a}\right)^{\frac{1}{b}}$. The recurrence could be rewritten as

$$y_{k+1} - y_k \leq -\phi(y_k).$$

On the other hand, the convexity of ϕ implies that

$$-\phi(x) \leq -\phi'(\bar{x})(x - \bar{x}).$$

Then

$$y_{k+1} - y_k \leq -\phi'(\bar{x})(y_k - \bar{x}).$$

Using the change of variable $z_k = y_k - \bar{x}$, we obtain

$$z_{k+1} \leq (1 - \phi'(\bar{x}))z_k.$$

Since $\phi'(\bar{x}) = \Delta < 1$, we can conclude that

$$z_k \leq (1 - \phi'(\bar{x}))^k z_0 \leq (1 - \phi'(\bar{x}))^k y_0.$$

Thus

$$y_k \leq (1 - \phi'(\bar{x}))^k y_0 + \bar{x}.$$

□

We now have the tools to be able to deduce results of (SGD) under functions satisfying $Ł_b^q(H)$ with $q \in (\frac{1}{2}, 1)$. These types of functions are flatter around its minimizers as q grows, so we expect that as q grows, the complexity will get worse. Moreover, this complexity should be worse than that obtained for the Gradient Descent (and its continuous version) with this same class of functions (see Corollaries 1.20 and 1.25). The following Proposition and its corresponding Corollary are a result of this Thesis.

Proposition 2.17 *Let $q \in (\frac{1}{2}, 1)$, assume $f \in C_L^{1,1}(H) \cap Ł_b^q(H)$, let $\gamma_k \equiv \gamma$, $(x_k)_{k\in\mathbb{N}}$ generated by (SGD), assume that:*

- *$(x_k)_{k\in\mathbb{N}}$ is bounded a.s..*
- *$\bar{\sigma}^2 = \sup_k \mathbb{E}_{\mathcal{D}}(\|\nabla f_s(x_k)\|^2) < \infty$.*

Then there exists $\mu > 0$ such that $\Delta := 2\gamma\mu^2 q \left(\frac{L\gamma\bar{\sigma}^2}{2\mu^2}\right)^{1-\frac{1}{2q}} \in (0, 1)$ and

$$\mathbb{E}(f(x_k)) - min(f) \leq (1 - \Delta)^k \left(f(x_0) - min(f)\right) + \left(\frac{L\gamma\bar{\sigma}^2}{2\mu^2}\right)^{\frac{1}{2q}}, \quad \forall k \in \mathbb{N}. \tag{2.10}$$

PROOF. Let $a_k := \mathbb{E}(f(x_k)) - min(f)$ and $\mathcal{K}$ a bounded set such that $(x_k)_{k\in\mathbb{N}} \subseteq \mathcal{K}$ a.s., let $\mu_0 > 0$ be the coefficient that exists because f satisfies the Łojasiewicz Inequality on $\mathcal{K}$. Using Lemma 2.12 and $f \in Ł_b^q(H)$, then

$$a_{k+1} - a_k \leq -\gamma\mu_0^2 \mathbb{E}\left[(f(x_k) - min(f))^{2q}\right] + \frac{L}{2}\gamma^2 \mathbb{E}(\mathbb{E}_{\mathcal{D}}(\|\nabla f_{s^k}(x_k)\|^2)).$$

By Jensen's inequality and the definition of $\bar{\sigma}^2$, we obtain

$$a_{k+1} - a_k \leq -\gamma\mu_0^2 a_k^{2q} + \frac{L}{2}\gamma^2\bar{\sigma}^2.$$

Let $\mu_1 > 0$ such that:

$$2\gamma\mu_1^2 q \left(\frac{L\gamma\bar{\sigma}^2}{2\mu_1^2}\right)^{1-\frac{1}{2q}} < 1.$$

And let $\mu = min\{\mu_0, \mu_1\}$, we deduce that

$$a_{k+1} - a_k \leq -\gamma\mu^2 a_k^{2q} + \frac{L}{2}\gamma^2\bar{\sigma}^2.$$

So by Lemma 2.16 with $a = \gamma\mu^2, b = 2q, c = \frac{L}{2}\gamma^2\bar{\sigma}^2$ and $y_0 = f(x_0) - min(f)$, $\Delta = 2\gamma\mu^2 q \left(\frac{L\gamma\bar{\sigma}^2}{2\mu^2}\right)^{1-\frac{1}{2q}} \in (0,1)$ by construction. Then a_k satisfies:

$$a_k \leq (1-\Delta)^k (f(x_0) - min(f)) + \left(\frac{L\gamma\bar{\sigma}^2}{2\mu^2}\right)^{\frac{1}{2q}}, \ \forall k \in \mathbb{N}.$$

□

Corollary 2.18 (Complexity of Proposition 2.17) *Let $\varepsilon \in (0,1)$ arbitrary, if $\gamma = \mathcal{O}(\varepsilon^{2q})$ and $k = \tilde{\Omega}_0\left(\frac{1}{\varepsilon^{4q-1}}\right)$. Then*

$$\mathbb{E}(f(x_k) - min(f)) \leq \varepsilon.$$

Proof. Let us consider (2.10), let $M > 0$ and define $\hat{\mu} = min\{\mu, M\}$, then (2.10) still holds if we change μ for $\hat{\mu}$, i.e., defining $\hat{\Delta} := 2\gamma\hat{\mu}^2 q \left(\frac{L\gamma\bar{\sigma}^2}{2\hat{\mu}^2}\right)^{1-\frac{1}{2q}} \in (0,1)$, we have that

$$\mathbb{E}(f(x_k)) - min(f) \leq \left(1-\hat{\Delta}\right)^k (f(x_0) - min(f)) + \left(\frac{L\gamma\bar{\sigma}^2}{2\hat{\mu}^2}\right)^{\frac{1}{2q}}, \ \forall k \in \mathbb{N}. \tag{2.11}$$

In order to guarantee $\mathbb{E}(f(x_k) - min(f)) \leq \varepsilon$, we will require that

$$(1-\hat{\Delta})^k (f(x_0) - min(f)) \leq \frac{\varepsilon}{2} \text{ and } \left(\frac{L\gamma\bar{\sigma}^2}{2\hat{\mu}^2}\right)^{\frac{1}{2q}} = \frac{\varepsilon}{2}.$$

This implies that $\gamma = \frac{\hat{\mu}^2}{2^{2q-1}L\bar{\sigma}^2}\varepsilon^{2q} = \mathcal{O}(\varepsilon^{2q})$ and

$$\begin{aligned}
(1-\hat{\Delta})^k(f(x_0)-min(f)) &\leq \frac{\varepsilon}{2} \\
\iff (1-\hat{\Delta})^k &\leq \frac{\varepsilon}{2(f(x_0)-min(f))} \\
\iff k\, ln(1-\hat{\Delta}) &\leq ln\left(\frac{\varepsilon}{2(f(x_0)-min(f))}\right) \\
\iff k &\geq \frac{ln\left(\frac{\varepsilon}{2(f(x_0)-min(f))}\right)}{ln(1-\hat{\Delta})} \\
\iff k &\geq \frac{ln\left(\frac{\varepsilon}{2(f(x_0)-min(f))}\right)}{ln\left(1-2\hat{\mu}^4 q\left(\frac{L\bar{\sigma}^2}{2}\right)^{1-\frac{1}{2q}}\left(\frac{1}{2^{(2q-1)}L\bar{\sigma}^2}\right)^{2-\frac{1}{2q}}\varepsilon^{4q-1}\right)}.
\end{aligned}$$

The coefficient $\hat{\mu}$ depends on γ (which depends on ε), but since $\hat{\mu} \leq M$, then taking a particular $k = \Omega_0\left(\frac{ln(\varepsilon)}{ln(1-\varepsilon^{4q-1})}\right)$, we can satisfy the last inequality and we could conclude a complexity rate. Nevertheless, since $\varepsilon \in (0,1)$, then considering Lemma 2.10 with $\alpha = 4q-1$, we have that

$$\frac{ln(\varepsilon)}{ln(1-\varepsilon^{4q-1})} \leq \frac{ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon^{4q-1}}, \ \ \forall \varepsilon \in (0,1).$$

So by taking a particular $k = \Omega_0\left(\frac{ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon^{4q-1}}\right) =: \tilde{\Omega}_0\left(\frac{1}{\varepsilon^{4q-1}}\right)$, we can conclude that

$$f(x_k) - min(f) \leq \varepsilon.$$

□

As we can notice, the closer q is to $\frac{1}{2}$, the better the complexity is, reaching the one obtained for the strongly convex case when $q = \frac{1}{2}$ ($\mathcal{O}(\varepsilon^{-1})$). On the other hand, the closer q is to 1, the worse the complexity is, reaching the worst case ($\mathcal{O}(\varepsilon^{-3})$) in the limit ($q = 1$).

Besides, comparing the complexity of the Gradient Descent (and its continuous version) with that of SGD, we conclude that in all the cases studied, the complexity of the Gradient Descent is better than that of SGD, more precissely:

- In the convex case, the comparison is between ε^{-1} and ε^{-2}, respectively.
- In the strongly convex and Łojasiewicz with exponent $q = \frac{1}{2}$ case, the comparison is between $ln(\varepsilon^{-1})$ and ε^{-1}, respectively.
- Now we can deduce that in the Łojasiewicz with exponent $q \in (\frac{1}{2}, 1)$ case, the comparison is between $\varepsilon^{-(2q-1)}$ and $\varepsilon^{-(4q-1)}$, respectively.

2.1.3 Mini-Batch SGD revisited

This section is going to be the link between the discrete and continuous stochastic dynamics, where we will model the Mini-Batch SGD (MB-SGD) as a continuous process and we will discuss the assumptions required to do so.

Consider the Mini-Batch SGD algorithm (MB-SGD) in $H = \mathbb{R}^{\mathrm{d}}$. In the sampling vectors examples we commented that the estimator of the Mini-Batch is unbiased from the gradient of f, i.e.

$$\mathbb{E}\left[\frac{1}{b}\sum_{\mathrm{i}\in J}\nabla f_{\mathrm{i}}(\cdot)\right] = \nabla f(\cdot), \ \ (\text{with } J \sim \mathit{Unif}\binom{[n]}{b}).$$

Let us assume a constant stepsize $\gamma_k \equiv \gamma$, from this structure we can build a continuous-time model of (MB-SGD). To do so, first we can rewrite the recurrence as

$$x_{k+1} = x_k - \gamma(\nabla f(x_k) + V(x_k)), \ \ k \in \mathbb{N},$$

where $V(x) = \frac{1}{b}\sum_{\mathrm{i}\in I}\nabla f_{\mathrm{i}}(x) - \nabla f(x)$ (with $I \sim \mathit{Unif}\binom{[n]}{b}$) is a random variable with zero mean and covariance $\Sigma(x) = \frac{\Sigma_{MB}(x)}{b}$, where

$$\Sigma_{MB}(x) := \frac{1}{n}\sum_{\mathrm{i}=1}^{n}(\nabla f(x) - \nabla f_{\mathrm{i}}(x))(\nabla f(x) - \nabla f_{\mathrm{i}}(x))^t.$$

Let σ_{MB} be the positive square root of Σ_{MB} and set $(Z_k)_{k\in\mathbb{N}}$ a sequence of random variables with zero mean and unit covariance such that $\frac{\sigma_{MB}(x_k)}{\sqrt{b}}Z_k$ has the same distribution as $V(x_k)$ (conditioned on x_k), then we can write (MB-SGD) as

$$x_{k+1} = x_k - \gamma\nabla f(x_k) - \sqrt{\frac{\gamma}{b}}\sigma_{MB}(x_k)\sqrt{\gamma}Z_k, \ \ k \in \mathbb{N}. \tag{2.12}$$

In order to build a continuous-time model of (MB-SGD), we will assume that each Z_k is Gaussian distributed i.e. $Z_k \sim \mathcal{N}(0_{\mathrm{d}}, I_{\mathrm{d}})$. This assumption relies on the fact that if the size of the mini-batch is large enough, then we can invoke the Central Limit Theorem and conclude that the distribution of Z_k is approximately Gaussian (see the Berry-Esseen Theorem, [18]). If we further assume that γ is small enough, then we can notice that (2.12) is the first-order discretization with stepsize γ of a SDE. Thus, its correspondent continuous-time model is

$$\begin{aligned} dX(t) &= -\nabla f(X(t))dt + \sqrt{\frac{\gamma}{b}}\sigma_{MB}(X(t))dB(t), \ \ t > 0, \\ X(0) &= x_0, \end{aligned} \tag{2.13}$$

where $\sigma_{MB} : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$, and $B(t)$ is a d$-$dimensional Brownian motion.

The Gaussian assumption relies on the fact that σ_{MB} has finite variance. Even though this assumption could be seen intuitive, it is not always true (see [19, Chapter 1.2]). If we do not suppose the finite variance hypothesis, then we will not have a Brownian motion necessarily but instead an $\alpha - \mathit{stable}$ Lévy motion (see [19]). Assuming the Gaussian assumption, we will return to analyze the SDE (2.13) in Chapter 3: A Continuous-Time Model of Stochastic Gradient Descent.

2.2 Continuous Stochastic Dynamics: Ito processes and SDE

In order to understand what it means equation (2.13) and more general continuous-time models for stochastic optimization algorithms, we are going to consider $F : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}}$, $G : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}\times m}$ and $B(t)$ a $m-$dimensional Brownian motion, we want to formalize and make sense of the following differential equation:

$$dX(t) = F(t, X(t))dt + G(t, X(t))dB(t).$$

To achieve this, we will need the following definitions, concepts and properties about Stochastic Analysis:

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_t | t \geq 0\}$ be a filtration of the $\sigma-$algebra $\mathcal{F}$. An event $E \in \mathcal{F}$ happens almost surely if $\mathbb{P}(E) = 1$, and it will be denoted as "$E,\ \mathbb{P} - a.s.$".

A stochastic process ($\mathbb{R}^{\mathrm{d}} - value$d) is a function $X : \Omega \times \mathbb{R}_+ \to \mathbb{R}^{\mathrm{d}}$. It is said to be continuous if $X(\omega, \cdot) \in C(\mathbb{R}_+; \mathbb{R}^{\mathrm{d}})$ for almost all $\omega \in \Omega$, and it is said to be bounded (a.s.) if there exists $r > 0$ such that $X(\omega, t) \subseteq B(0; r)$ for all $t \geq 0$ and for almost all $\omega \in \Omega$. We will denote $X(t) := X(\cdot, t)$. We will study dynamics whose solutions are stochastic processes (for instance (2.13)). In order to ensure uniqueness of a solution, we will introduce a relation over these processes. Two stochastic processes $X, Y : \Omega \times [0, T] \to \mathbb{R}^{\mathrm{d}}$ are said to be equivalent if

$$X(t) = Y(t),\ \ \forall t \in [0, T],\ \ \mathbb{P} - a.s..$$

This definition leads us to define the equivalence relation $\mathcal{R}$, which associates the equivalent stochastic processes in the same class (the definition of equivalent stochastic processes).

Furthermore, we will need some properties about the measurability of these processes. A stochastic process $X : \Omega \times \mathbb{R}_+ \to \mathbb{R}^{\mathrm{d}}$ is progressively measurable if for every $t \geq 0$, the map $\Omega \times [0, t] \to \mathbb{R}^{\mathrm{d}}$ defined by $(\omega, s) \to X(\omega, s)$ is $\mathcal{F}_t \bigotimes \mathcal{B}([0, t]) - measurable$ (where $\bigotimes$ is the product $\sigma-$algebra and $\mathcal{B}$ is the Borel $\sigma-$algebra). On the other hand, X is $\mathcal{F}_t - adapted$ if $X(t)$ is $\mathcal{F}_t$ measurable for every $t \geq 0$. It is direct from the definition that if X is progressively measurable then X is $\mathcal{F}_t - adapted$.

With these concepts, we can introduce some interesting spaces. We define the quotient space:

$$S_{\mathrm{d}}^0[0, T] := \left\{X : \Omega \times [0, T] \to \mathbb{R}^{\mathrm{d}} | X \text{ is a progressively measurable continuous stochastic process}\right\} \Big/ \mathcal{R}.$$

Furthermore, $S_{\mathrm{d}}^0 := \bigcap_{T \geq 0} S_{\mathrm{d}}^0[0, T]$.

For $p > 0$, we define $S_{\mathrm{d}}^p[0,T]$ as the subset of the processes $X(t)$ in $S_{\mathrm{d}}^0[0,T]$ such that $\sup_{t\in[0,T]} \|X(t)\|^p$ has finite first moment. In other words

$$S_{\mathrm{d}}^p[0,T] := \left\{ X \in S_{\mathrm{d}}^0[0,T] \Big| \mathbb{E}\left(\sup_{t\in[0,T]} \|X_t\|^p \right) < \infty \right\}.$$

Furthermore, $S_{\mathrm{d}}^p := \bigcap_{T\geq 0} S_{\mathrm{d}}^p[0,T]$.

A useful definition will be the usual L^p definition, but in the stochastic case. We denote $\mathcal{L}^p([0,T];\mathbb{R}^{\mathrm{d}})$, with $p > 0$, to the family of $\mathbb{R}^{\mathrm{d}}-$valued, $\mathcal{F}_t-$adapted processes $h : \Omega\times[0,T] \to \mathbb{R}^{\mathrm{d}}$ such that:

$$\int_0^T \|h(t)\|^p dt < \infty, \ \ \mathbb{P}-a.s..$$

We will write $h \in \mathcal{L}^p(\mathbb{R}_+;\mathbb{R}^{\mathrm{d}})$, with $p > 0$, if $h \in \mathcal{L}^p([0,T];R^{\mathrm{d}})$ for every $T > 0$.

We will focus on a special case of Stochastic Processes, called Ito processes, because the dynamics of the Continuous-Time Model of SGD (in which we are interested) will be this type of process (see for instance (2.13)).

Definition. Assume that $f \in \mathcal{L}^1(\mathbb{R}_+;\mathbb{R}^{\mathrm{d}})$ and $g \in \mathcal{L}^2(\mathbb{R}_+;\mathbb{R}^{\mathrm{d}\times m})$. $X \in S_{\mathrm{d}}^0$ is an Ito process if it takes the following form

$$X(t) = X_0 + \int_0^t f(s)ds + \int_0^t g(s)dB(s), \ \ t \geq 0, \mathbb{P}-a.s.. \tag{2.14}$$

or its differential form (which is an alternative way to express (2.14))

$$\begin{aligned} dX(t) &= f(t)dt + g(t)dB(t), \ t \geq 0, \mathbb{P}-a.s. \\ X(0) &= X_0, \end{aligned} \tag{2.15}$$

where B is a $\mathcal{F}_t$-adapted $m-$dimensional Brownian motion.

Now we are ready to describe and analyze the following Stochastic Differential Equation (SDE).

Let $F : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}}$ and $G : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}\times m}$ be measurable functions. Consider the SDE:

$$\begin{aligned} dX(t) &= F(t,X(t))dt + G(t,X(t))dB(t), \ \ t > 0, \mathbb{P}-a.s. \\ X(0) &= X_0. \end{aligned} \tag{2.16}$$

This SDE is diferent than the one presented in (2.15), since X appears on the right hand side of the equation. So we will formalize what it means to solve (2.16).

X is called a solution of (2.16) if for every $T > 0$, $F \in \mathcal{L}^1([0,T];\mathbb{R}^{\mathrm{d}}), G \in \mathcal{L}^2([0,T];\mathbb{R}^{\mathrm{d}\times m})$ and

$$X(t) = X_0 + \int_0^t F(s, X(s))\,ds + \int_0^t G(s, X(s))\,dB(s), \;\; t \in [0,T], \mathbb{P} - a.s.. \tag{2.17}$$

Every solution of (2.16) is an Ito Process.

The following Theorem will ensure sufficient conditions for (2.16) to have a unique solution.

Theorem 2.19 *(See [20, Theorem 5.2.1]) Let $F : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}}$ and $G : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}\times m}$ be measurable functions satisfying for every $T > 0$*

$$\|F(t,x)\| + \|G(t,x)\|_F \leq C_1(1 + \|x\|), \quad \forall x \in \mathbb{R}^{\mathrm{d}}, \forall t \in [0,T], \tag{2.18}$$

for some constant C_1, and such that for every $T > 0$:

$$\|F(t,x) - F(t,y)\| + \|G(t,x) - G(t,y)\|_F \leq C_2\|x - y\|, \quad \forall x, y \in \mathbb{R}^{\mathrm{d}}, \forall t \in [0,T], \tag{2.19}$$

for some constant C_2. Then (2.16) has a unique solution $X \in S^2_{\mathrm{d}}$.

A difficulty we face in order to be able to analyze the behavior of a solution X of (2.16), is that X itself usually does not provide us with much information. For instance, if we want to know if $X(t)$ is approaching to a set $A \subseteq \mathbb{R}^{\mathrm{d}}$, we need to know how $d(X(t), A)$ behaves rather than $X(t)$. This is one of the reasons why we want to inquire into what happens with the Ito process X when it is mapped via a function $\phi : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ such that $(t, X(t)) \mapsto \phi(t, X(t))$.

Theorem 2.20 (Ito's Formula) *Let $X \in S^0_{\mathrm{d}}$ be a solution of (2.16) i.e. an Ito process and let $\phi \in C^2(\mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}}; \mathbb{R})$. Then the process*

$$Y(t) = \phi(t, X(t)), \quad \forall t \geq 0, \mathbb{P} - a.s.,$$

is again an Ito process, such that

$$dY(t) = \frac{\partial \phi}{\partial t}(t, X(t))\,dt + \sum_{\mathrm{i}} \frac{\partial \phi}{\partial x_{\mathrm{i}}}(t, X(t))\,dX_{\mathrm{i}}(t) + \frac{1}{2}\sum_{\mathrm{i},j} \frac{\partial^2 \phi}{\partial x_{\mathrm{i}} \partial x_j}(t, X(t))\,dX_{\mathrm{i}}(t)\,dX_j(t), \tag{2.20}$$

where $dB_{\mathrm{i}}\,dB_j = \delta_{\mathrm{i}j}\,dt$, $dB_{\mathrm{i}}\,dt = dt\,dB_{\mathrm{i}} = 0$ ($\delta_{\mathrm{i}j}$ is the Kronecker Delta).

Or equivalently

$$\begin{aligned} Y(t) =& Y(0) + \int_0^t \frac{\partial \phi}{\partial t}(s, X(s))\,ds + \int_0^t \langle \nabla \phi(s, X(s)), F(s, X(s))\rangle\,ds \\ &+ \frac{1}{2}\int_0^t tr[G(s, X(s))G^t(s, X(s))Hess(\phi(s, X(s)))]\,ds + \int_0^t \langle \nabla \phi(s, X(s)), G(s, X(s))\,dB(s)\rangle. \end{aligned} \tag{2.21}$$

The following Corollary is crucial in this Thesis and describes the behavior in expectation of the map $(t, X(t)) \mapsto \phi(t, X(t))$ (with $X(t)$ ând ϕ as before).

Corollary 2.21 *Let $X \in S_{\mathrm{d}}^0$ be a solution of (2.16) i.e. an Ito Process, and $\phi \in C^2(\mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}}; \mathbb{R})$. Then the process*

$$Y(t) = \phi(t, X(t)),$$

is an Ito Process, such that

$$\begin{aligned} \mathbb{E}[Y(t)] = Y(0) + \mathbb{E}\left[\int_0^t \frac{\partial \phi}{\partial t}(s, X(s))ds + \int_0^t \langle \nabla \phi(s, X(s)), F(s, X(s))\rangle ds\right] \\ + \frac{1}{2}\mathbb{E}\left[\int_0^t tr[G(s, X(s))G^t(s, X(s))Hess(\phi(s, X(s)))]ds\right]. \end{aligned} \tag{2.22}$$

Proof. We just have to prove that

$$\mathbb{E}\left[\int_0^t \langle \nabla \phi(s, X(s)), G(s, X(s))dB(s)\rangle\right] = 0.$$

And this is true since

$$\mathbb{E}\left[\int_0^T |\langle \nabla \phi(s, X(s)), G(s, X(s))\rangle|^2 ds\right] < \infty, \ \ \forall T > 0 \text{ (see [21, Theorem 1.5.8])}.$$

□

Equipped with the results shown in this section, we are in a position to define the continuous-time model of *SGD* on which we will work.

Chapter 3

A Continuous-Time Model of Stochastic Gradient Descent

Inspired by (2.13), we are going to define and analyze a Continuous-Time Model of *SGD*.

Consider $f \in C^{1,1}(\mathbb{R}^{\mathrm{d}}) \cap \Gamma_0(\mathbb{R}^{\mathrm{d}})$ and the following dynamic:

$$\begin{aligned} dX &= -\nabla f(X)dt + \sigma(t,X)dB \\ X(0) &= X_0, \end{aligned} \tag{CSGD}$$

where:

1. B is a $\mathcal{F}_t$-adapted $m-$dimensional Brownian motion.
2. The $\mathrm{d} \times m$ volatility matrix $\sigma_{\mathrm{i}k} : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is measurable and

$$\begin{aligned} \sup_{t,x} |\sigma_{\mathrm{i}k}(t,x)| &< \infty \\ |\sigma_{\mathrm{i}k}(t,x') - \sigma_{\mathrm{i}k}(t,x)| &\leq l\|x' - x\|, \end{aligned} \tag{H}$$

for some $l > 0$ and for all $t \geq 0, x, x' \in \mathbb{R}^{\mathrm{d}}$

Also (H) implies the existence of $\sigma_* > 0$ such that:

$$tr[\Sigma(t,x)] \leq \sigma_*^2,$$

for all $t \geq 0, x \in \mathbb{R}^{\mathrm{d}}$, where $\Sigma = \sigma\sigma^t$.

This is the Continuous-Time Model of Stochastic Gradient Descent on which we will focus in order to analyze its behavior and properties.

Before stating the main propositions related to (CSGD), we will need some previous definitions and technical results.

3.1 Technical results

Let $A, B \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$ symmetric matrices, we write $A \preccurlyeq B$ if $B - A$ is a positive semi-definite matrix, i.e.

$$x^t(B-A)x \geq 0, \ \ \forall x \in \mathbb{R}^{\mathrm{d}}.$$

Proposition 3.1 *Let $\phi \in C_L^{1,1}(\mathbb{R}^{\mathrm{d}})$, then ϕ is almost everywhere twice differentiable and its Hessian satisfies*

$$Hess(\phi(x)) \preccurlyeq LI, \ \ \textit{for (Lebesgue) almost all } x \in \mathbb{R}^{\mathrm{d}}.$$

Proof. Direct from Rademacher's Theorem (see [22, Theorem 3.1.6]) and Proposition 1.6. □

Corollary 3.2 *Let $\phi \in C_L^{1,1}(\mathbb{R}^{\mathrm{d}})$, $\sigma : \mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}\times m}$ and $\Sigma = \sigma\sigma^t$. Then*

$$tr(\Sigma(t,x) \cdot Hess(\phi(x))) \leq L \cdot tr(\Sigma(t,x)), \ \ \forall t \geq 0, \textit{ for (Lebesgue) almost all } x \in \mathbb{R}^{\mathrm{d}}.$$

Proof. Let $t \geq 0, x \in \mathbb{R}^{\mathrm{d}}$ such that $H_x := Hess(\phi(x))$ exists and $\sigma := \sigma(t,x)$. Denote

$$H_x = (h_{\mathrm{i}j})_{\mathrm{d}\times\mathrm{d}}, \text{ and } \sigma = (\sigma_{\mathrm{i}j})_{\mathrm{d}\times m}.$$

$$tr[H_x\Sigma] = \sum_{\mathrm{i}=1}^{\mathrm{d}}\sum_{k=1}^{\mathrm{d}}\sum_{j=1}^{m}\sigma_{kj}h_{\mathrm{i}k}\sigma_{\mathrm{i}j}. \tag{3.1}$$

On the other hand, by Proposition 3.1

$$u^tH_xu \leq Lu^tu, \ \ \forall u \in \mathbb{R}^{\mathrm{d}}.$$

Choosing $u = \sigma_{:j}$ (column j of σ) we have

$$\sum_{k=1}^{\mathrm{d}}\sum_{\mathrm{i}=1}^{\mathrm{d}}\sigma_{kj}h_{\mathrm{i}k}\sigma_{\mathrm{i}j} \leq L\sum_{\mathrm{i}=1}^{\mathrm{d}}\sigma_{\mathrm{i}j}^2.$$

Adding up the previous equation over $j \in [m]$

$$\sum_{j=1}^{m}\sum_{k=1}^{\mathrm{d}}\sum_{\mathrm{i}=1}^{\mathrm{d}}\sigma_{kj}h_{\mathrm{i}k}\sigma_{\mathrm{i}j} \leq L\sum_{j=1}^{m}\sum_{\mathrm{i}=1}^{\mathrm{d}}\sigma_{\mathrm{i}j}^2.$$

Rearranging terms

$$\sum_{\mathrm{i}=1}^{\mathrm{d}}\sum_{k=1}^{\mathrm{d}}\sum_{j=1}^{m}\sigma_{kj}h_{\mathrm{i}k}\sigma_{\mathrm{i}j} \leq L\sum_{\mathrm{i}=1}^{\mathrm{d}}\sum_{j=1}^{m}\sigma_{\mathrm{i}j}^2.$$

Plugging the previous inequality in (3.1), we obtain

$$tr[\Sigma \cdot H_x] \leq L \sum_{\mathrm{i}=1}^{\mathrm{d}} \sum_{j=1}^{m} \sigma_{\mathrm{ij}}^2 = L \cdot tr[\Sigma],$$

and we conclude. □

In order to find upper bounds of (CSGD), at some point we will need to exchange the derivative with the expectation, the following Proposition will give us sufficient conditions to ensure this exchange.

Proposition 3.3 *Let D be an open subset of $\mathbb{R}$. Suppose $g : \Omega \times D \to \mathbb{R}$ satisfies the following conditions:*

1. *$\mathbb{E}[g(\omega, t)] < \infty$ for each $t \in D$.*
2. *For almost all $\omega \in \Omega$, $\frac{\partial g}{\partial t}$ exists for all $t \in D$.*
3. *There is a function $Z : \Omega \to \mathbb{R}$ such that $\mathbb{E}[Z] < \infty$ and*

$$\left| \frac{\partial}{\partial t} g(\omega, t) \right| \leq Z(\omega), \quad \forall t \in D \text{ and almost every } \omega \in \Omega.$$

Then

$$\frac{d}{dt}\mathbb{E}[g(\omega, t)] = \mathbb{E}\left[\frac{\partial g(\omega, t)}{\partial t} \right], \quad \forall t \in D.$$

Proof.

$$\begin{aligned} \frac{d}{dt}\mathbb{E}[g(\omega, t)] &= \lim_{h \to 0} \frac{\mathbb{E}[g(\omega, t+h)] - \mathbb{E}[g(\omega, t)]}{h} \\ &= \lim_{h \to 0} \mathbb{E}\left[\frac{g(\omega, t+h) - g(\omega, t)}{h} \right] \\ &= \lim_{h \to 0} \mathbb{E}\left[\frac{\partial}{\partial t} g(\omega, \tau(h)) \right]. \end{aligned}$$

where $\tau(h) \in (t, t+h)$ exists by the mean value Theorem. On the other hand, we suppose

$$\left| \frac{\partial}{\partial t} g(\omega, \tau(h)) \right| \leq Z(\omega).$$

By the Dominated Convergence Theorem, we conclude that

$$\frac{d}{dt}\mathbb{E}[g(\omega, t)] = \mathbb{E}\left[\lim_{h \to 0} \frac{\partial}{\partial t} g(\omega, \tau(h)) \right] = \mathbb{E}\left[\frac{\partial g(\omega, t)}{\partial t} \right].$$

□

To get results of (CSGD) when the objetive function is in $C_L^{1,1}(\mathbb{R}^{\mathrm{d}}) \cap \text{Ł}_b^q(\mathbb{R}^{\mathrm{d}})$ for $q \in (\frac{1}{2}, 1)$, we will need a continuous version of Lemma 2.16.

Proposition 3.4 *Consider the differential inequation*

$$y'(t) + ay(t)^b \leq c, \quad y(0) = y_0, \tag{3.2}$$

where $a, b, c > 0$ *and* $b > 1, y_0 > 0$. *Then*

$$y(t) \leq \left(\frac{c}{a}\right)^{\frac{1}{b}} + y_0 e^{-\Delta t}, \forall t \geq 0,$$

where $\Delta = ab\left(\frac{c}{a}\right)^{1-\frac{1}{b}}$.

Proof. Let $\phi(y) = ay^b - c$ and $\bar{y} := \left(\frac{c}{a}\right)^{\frac{1}{b}}$ be the root of ϕ. Also

- $\phi'(y) = aby^{b-1}$, which is positive when $y > 0$.
- $\phi''(y) = ab(b-1)y^{b-2}$, which is positive when $y > 0$.

Then ϕ is increasing and convex on $(0, \infty)$; it is negative before $\bar{y}$, and it is positive after. We can rewrite (3.2) as

$$y'(t) + \phi(y(t)) \leq 0.$$

By the convexity of ϕ we have

$$\phi(y(t)) - \phi(\bar{y}) - \phi'(\bar{y})(y(t) - \bar{y}) \geq 0.$$

Then

$$y'(t) \leq -\phi(y(t)) \leq -\phi'(\bar{y})(y(t) - \bar{y}).$$

Introducing the change of variable $z(t) := y(t) - \bar{y}$ we obtain

$$z'(t) \leq -\phi'(\bar{y})z(t).$$

This differential inequality implies that

$$y(t) - \bar{y} = z(t) \leq z(0)e^{-\phi'(\bar{y})t} \leq y_0 e^{-\phi'(\bar{y})t},$$

and since $\phi'(\bar{y}) = \Delta$, we conclude. □

3.2 Main Results: Upper bounds and complexity rates for (CSGD)

In this section we will discuss the existence and uniqueness of a solution for (CSGD), show upper bounds of this dynamic and obtain remarkable complexity rates.

Proposition 3.5 *Assume that $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$. If we consider the dynamic (CSGD) under the hypothesis (H), then (CSGD) has a unique solution $X \in S_d^2$.*

Proof. Direct from checking the conditions of Theorem 2.19. □

Remark. There are other hypotheses like [23, Section 2.2] that allow us to deduce the existence and uniqueness of a solution for (2.13), which is a particular case of (CSGD).

Assume that $f \in C^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ and set $x^* \in S$. In the following Propositions we will show upper bounds for different terms, in particular of quantities like:

i) $\mathbb{E}\left[\|x - x^*\|^2\right]$ (Squared distance to a particular solution).

ii) $\mathbb{E}[f(x)] - min(f)$.

iii) $\mathbb{E}\left[d(x, S)^2\right]$.

In order to relate these quantities, we are going to recall Lemma 1.6 and Proposition 1.14.

To switch from an upper bound of the term:

- i) to ii), we will use Lemma 1.6 if $f \in C_L^{1,1}(\mathbb{R}^d)$.
- ii) to iii), we will use Proposition 1.14 if f satisfies the Łojasiewicz Inequality on a particular set.
- i) to iii), we will use the definition of distance to a set.

The main results of this Thesis related to the upper bounds and complexities of (CSGD) are presented in the following Propositions:

Proposition 3.6 *Assume that $f \in C_L^{1,1}(\mathbb{R}^{\mathrm{d}}) \cap \Gamma_0(\mathbb{R}^{\mathrm{d}})$, let $x^* \in S$. If we consider the dynamic (CSGD) under the hypothesis (H) and $X \in S_{\mathrm{d}}^2$ its unique solution, then the following statements holds:*

(i) Let $f_{min}(t) := \min_{s\in[0,t]} f(X(s))$, $\bar{X}(t) = t^{-1}\int_0^t X(s)ds$ and $\bar{f}(t) := t^{-1}\int_0^t f(X(s))ds$. Since $f \in \Gamma_0(\mathbb{R}^{\mathrm{d}})$ then

$$\mathbb{E}\left[f_{min}(t)\right] - min(f) \leq \mathbb{E}\left[f(\bar{X}(t))\right] - min(f) \leq \mathbb{E}\left[\bar{f}(t)\right] - min(f) \leq \frac{d(X_0,S)^2}{2t} + \frac{\sigma_*^2}{2}, \ \ \forall t > 0. \tag{3.3}$$

(ii) If $f \in \Gamma_\mu(\mathbb{R}^{\mathrm{d}})$, then

$$\mathbb{E}\left(\frac{\|X(t) - x^*\|^2}{2}\right) \leq \frac{\|X_0 - x^*\|^2}{2} e^{-\mu t} + \frac{\sigma_*^2}{2\mu}, \ \ \forall t \geq 0. \tag{3.4}$$

Proof. (i) Let $g(t) = \phi(t, X(t)) = \frac{\|X(t)-x^*\|^2}{2}$ and $G(t) = \mathbb{E}(g(t))$. Using Corollary 2.21, we obtain

$$\begin{aligned} G(t) - G(0) &= \mathbb{E}\left[\int_0^t \langle \nabla f(X(s)), x^* - X(s)\rangle ds\right] + \frac{1}{2}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s))]ds\right] \\ &\leq -\mathbb{E}\left[\int_0^t f(X(s)) - min(f)ds\right] + \frac{1}{2}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s))]ds\right] \\ &\leq -\mathbb{E}\left[\int_0^t f(X(s)) - min(f)ds\right] + \frac{\sigma_*^2 t}{2}, \end{aligned} \tag{3.5}$$

where the first inequality is given by the fact that $f \in \Gamma_0(\mathbb{R}^{\mathrm{d}})$. Then rearranging terms, using $G(t) \geq 0$ and dividing by t, we obtain

$$\frac{1}{t}\mathbb{E}\left[\int_0^t f(X(s)) - min(f)ds\right] \leq \frac{\|X_0 - x^*\|^2}{2t} + \frac{\sigma_*^2}{2}, \ \ \forall t > 0. \tag{3.6}$$

And using that x^* is arbitrary, we can make the inequality of (3.7) tighter and get

$$\frac{1}{t}\mathbb{E}\left[\int_0^t f(X(s)) - min(f)ds\right] \leq \frac{d(X_0,S)^2}{2t} + \frac{\sigma_*^2}{2}, \ \ \forall t > 0. \tag{3.7}$$

(ii) Let $g(t) = \phi(t, X(t)) = \frac{\|X(t)-x^*\|^2}{2}, G(t) = \mathbb{E}(g(t))$. Using Corollary (2.21), we obtain

$$G(t) - g(0) = \mathbb{E}\left[\int_0^t \langle -\nabla f(X(s)), X(s) - x^*\rangle ds\right] + \frac{1}{2}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s))]ds\right]. \tag{3.8}$$

Let $T > 0, t \in [0, T]$. Analyzing the right hand side of (3.8), we conclude that is differentiable in $(0, T)$, to see this, we must look at the following term

$$\frac{d}{dt}\mathbb{E}\left[\int_0^t \langle -\nabla f(X(s)), X(s) - x^*\rangle ds\right] + \frac{1}{2}\frac{d}{dt}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s))]ds\right]. \tag{3.9}$$

We have

$$\left|\frac{d}{dt}\int_0^t \langle -\nabla f(X(s)), X(s)-x^*\rangle ds\right| = |\langle -\nabla f(X(t)), X(t)-x^*\rangle| \leq L \sup_{t\in[0,T]} \|X(t)-x^*\|^2.$$

Where the equality is given by [24, Theorem 4.10] and the inequality is given by Cauchy-Schwarz Inequality. Also,

$$\mathbb{E}\left(L \sup_{t\in[0,T]} \|X(t)-x^*\|^2\right) = L\mathbb{E}\left(\sup_{t\in[0,T]} \|X(t)-x^*\|^2\right) < \infty,$$

because $X \in S_{\mathrm{d}}^2$.

Similarly,

$$\left|\frac{d}{dt}\int_0^t tr[\Sigma(s,X(s))]ds\right| = tr[\Sigma(t,X(t))] \leq \sigma_*^2$$

and

$$\mathbb{E}(\sigma_*^2) = \sigma_*^2 < \infty.$$

Thus, applying Proposition 3.3 twice, with

$$D=(0,T), g_1(\omega,t) = \int_0^t \langle -\nabla f(X(\omega,s)), X(\omega,s)-x^*\rangle ds \text{ and } g_2(\omega,t) = \int_0^t tr[\Sigma(s,X(\omega,s))]ds,$$

we obtain (3.9) is equal to

$$\mathbb{E}(\langle -\nabla f(X(t)), X(t)-x^*\rangle) + \frac{1}{2}\mathbb{E}(tr[\Sigma(t,X(t))]).$$

So the right hand side of (3.8) is differentiable in (0,T), then we can derivate both sides

$$\dot{G}(t) = \mathbb{E}(\langle -\nabla f(X(t)), X(t)-x^*\rangle) + \frac{1}{2}\mathbb{E}(tr[\Sigma(t,X(t))]), \ \ \forall t \in (0,T). \tag{3.10}$$

Using $f \in \Gamma_\mu(\mathbb{R}^{\mathrm{d}})$, then

$$\dot{G}(t) \leq -\mu G(t) + \frac{\sigma_*^2}{2}, \ \ \forall t \in (0,T).$$

Now we can solve the ordinary differential inequality by using the integrating factor method and obtain

$$G(t) \leq \frac{\|X_0-x^*\|^2}{2}e^{-\mu t} + \frac{\sigma_*^2}{2\mu}, \quad \forall t \in [0,T].$$

Using that $T>0$ is arbitrary, we conclude the inequality holds for all $t \geq 0$.

Moreover, by Proposition 1.6

$$\mathbb{E}[f(X(t)) - min(f)] \leq L\left(\frac{\|X_0-x^*\|^2}{2}e^{-\mu t} + \frac{\sigma_*^2}{2\mu}\right), \ \ \forall t \geq 0, \tag{3.11}$$

and we get another upper bound.

□

Proposition 3.7 *Consider the dynamic (CSGD), assume that there exists $\sigma_*^2 > 0$ such that*

$$\sup_{t\geq 0, x\in\mathbb{R}^{\mathrm{d}}} \|\sigma(t,x)\|_F^2 \leq \sigma_*^2.$$

Let X be a solution of (CSGD), $x^* \in S$ and suppose that X is bounded, then the following statements holds:

(i) If $f \in Ł_b^{1/2}(\mathbb{R}^{\mathrm{d}})$, then there exists $\mu > 0$ such that:

$$\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \frac{d(X_0,S)^2}{2} e^{-\frac{\mu^2}{2}t} + \frac{2\sigma_*^2}{\mu^2}, \ \forall t \geq 0. \tag{3.12}$$

(ii) If $f \in Ł_b^q(\mathbb{R}^{\mathrm{d}})$, $q \in (\frac{1}{2}, 1)$, there exists $\tilde{\mu} > 0$ such that

$$\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \frac{d(X_0,S)^2}{2} e^{-\left(\frac{\tilde{\mu}^{2-2q}}{1-q}\right)\left(\sigma_*^2\right)^{2q-1}t} + \frac{1}{2}\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{2(1-q)}, \ \forall t \geq 0. \tag{3.13}$$

Moreover, if $f \in C_L^{1,1}(\mathbb{R}^{\mathrm{d}})$ then

iii) For $f \in Ł_b^{1/2}(\mathbb{R}^{\mathrm{d}})$, there exists $\mu > 0$ such that:

$$\mathbb{E}\left[f(X(t))\right] - min(f) \leq (f(X_0) - min(f))e^{-\mu^2 t} + \frac{\sigma_*^2 L}{2\mu^2}, \ \forall t \geq 0. \tag{3.14}$$

iv) For $f \in Ł_b^q(\mathbb{R}^{\mathrm{d}})$, $q \in (\frac{1}{2}, 1)$, there exists $\mu > 0$ such that:

$$\mathbb{E}\left[f(X(t))\right] - min(f) \leq (f(X_0) - min(f))e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-1/2q}t} + \left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1/2q}, \ \forall t \geq 0. \tag{3.15}$$

Proof. (i) Let $\hat{g}(t) = \hat{\phi}(t, X(t)) = \frac{d(X(t),S)^2}{2}$, $\hat{G}(t) = \mathbb{E}(\hat{g}(t))$, $\mathcal{K}$ a bounded set such that $X \subseteq \mathcal{K}$ a.s. and $\mu > 0$ be the coefficient that exists because f satisfies the Łojasiewicz Inequality on $\mathcal{K}$. We have

$$\nabla\hat{\phi}(t, X(t)) = X(t) - P_S(X(t)),$$

where $P_S(x)$ is the projection of x in S and $Hess(\hat{\phi}(t, X(t))) \preccurlyeq 2I$. Using Corollary 2.21, we obtain

$$\begin{aligned}\hat{G}(t) - \hat{g}(0) =& \mathbb{E}\left[\int_0^t \langle -\nabla f(X(s), X(s) - P_S(X(s))\rangle ds\right] \\ &+ \frac{1}{2}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s))Hess(\hat{\phi}(s, X(s)))]ds\right].\end{aligned} \tag{3.16}$$

The right hand side of (3.16) is differentiable, to see this, we must look at the following term

$$\frac{d}{dt}\mathbb{E}\left[\int_0^t \langle -\nabla f(X(s), X(s) - P_S(X(s))\rangle ds\right] + \frac{1}{2}\frac{d}{dt}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s))Hess(\hat{\phi}(s, X(s)))]ds\right]. \tag{3.17}$$

We have for almost all $\omega \in \Omega$ that

$$\begin{aligned}\left|\frac{d}{dt}\int_0^t \langle -\nabla f(X(s), X(s) - P_S(X(s))\rangle ds\right| &= |\langle -\nabla f(X(t), X(t) - P_S(X(t))\rangle| \\ &\leq \|\nabla f(X(t))\|\|X(t) - x^*\| \leq R,\end{aligned}$$

for some $R > 0$ since X is bounded and ∇f is continuous. The first equality is given by [24, Theorem 4.10]. Similarly, for almost all $\omega \in \Omega$ we have

$$\left|\frac{d}{dt}\int_0^t tr[\Sigma(s, X(s))Hess(\hat{\phi}(s, X(s)))]ds\right| = tr[\Sigma(t, X(t))Hess(\hat{\phi}(t, X(t)))] \leq 2\sigma_*^2.$$

Where the inequality is given by Corollary 3.2. Thus, applying Proposition 3.3 twice, with

$$D = \mathbb{R}_+, g_1(\omega, t) = \int_0^t \langle -\nabla f(X(s), X(s) - P_S(X(s))\rangle ds \text{ and } g_2(\omega, t) = \int_0^t tr[\Sigma(s, X(\omega, s))Hess(\hat{\phi}(s, X(\omega, s)))]ds,$$

we obtain that (3.17) is equal to

$$\mathbb{E}\left[\langle -\nabla f(X(t), X(t) - P_S(X(t))\rangle\right] + \frac{1}{2}\mathbb{E}\left[tr[\Sigma(t, X(t))Hess(\hat{\phi}(t, X(t)))]\right].$$

Now we can derivate both sides of (3.16) and get

$$\begin{aligned}\dot{\hat{G}}(t) &= \mathbb{E}\left[\langle -\nabla f(X(t), X(t) - P_S(X(t))\rangle\right] + \frac{1}{2}\mathbb{E}\left[tr[\Sigma(t, X(t))Hess(\hat{\phi}(t, X(t)))]\right] \\ &\leq -\mathbb{E}(f(X(t)) - min(f)) + \sigma_*^2, \ \ \forall t \geq 0.\end{aligned} \tag{3.18}$$

Where it has been used $f \in \Gamma_0(\mathbb{R}^{\mathrm{d}})$ and Corollary 3.2. Using $f \in \text{Ł}_b^{1/2}(\mathbb{R}^{\mathrm{d}})$ and plugging the inequality of Proposition 1.14 in (3.18), we obtain

$$\dot{\hat{G}}(t) \leq -\frac{\mu^2}{2}\hat{G}(t) + \sigma_*^2, \ \ \forall t \geq 0.$$

Now we can solve the ordinary differential inequality by using the integrating factor method and obtain

$$\hat{G}(t) \leq \frac{d(X_0, S)}{2}e^{-\frac{\mu^2}{2}t} + \frac{2\sigma_*^2}{\mu^2}, \quad \forall t \geq 0.$$

(ii) Let $p = \frac{1}{1-q}$. Inequality (3.18) still holds in this case because $f \in \text{Ł}_b^{1/2}(\mathbb{R}^{\mathrm{d}})$ it is not used yet, using that $f \in \text{Ł}_b^q(\mathbb{R}^{\mathrm{d}}), q \in (\frac{1}{2}, 1)$ (μ as in i)) and plugging the inequality of Proposition 1.14 in (3.18), we obtain that there exists $\tilde{\mu} > 0$ such that

$$\begin{aligned}\dot{\hat{G}}(t) &\leq -\tilde{\mu}\mathbb{E}\left(d(X(t), S)^p\right) + \sigma_*^2 \\ &\leq -2^{p/2}\tilde{\mu}\hat{G}(t)^{p/2} + \sigma_*^2, \ \ \forall t \geq 0.\end{aligned} \tag{3.19}$$

Where the second inequality comes from $p > 2$ and Jensen's inequality.

Then by Proposition 3.4 with $a = 2^{p/2}\tilde{\mu}, b = \frac{p}{2}, c = \sigma_*^2, y(0) = \frac{d(X_0,S)^2}{2}$, we have that $\hat{G}$ satisfies

$$\hat{G}(t) \leq \frac{1}{2}\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{2/p} + \frac{d(X_0,S)^2}{2} e^{-\tilde{\mu}p\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{1-2/p} t}, \ \ \forall t \geq 0.$$

This is

$$\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \frac{d(X_0,S)^2}{2} e^{-\tilde{\mu}p\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{1-2/p} t} + \frac{1}{2}\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{2/p}, \ \ \forall t \geq 0. \tag{3.20}$$

(iii) Let $\tilde{g}(t) = \tilde{\phi}(t, X(t)) = f(X(t)) - min(f), \tilde{G}(t) = \mathbb{E}(\tilde{g}(t))$, $\mathcal{K}$ a bounded set such that $X \subseteq \mathcal{K}$ a.s. and $\mu > 0$ be the coefficient that exists because f satisfies the Łojasiewicz Inequality on $\mathcal{K}$. Using Corollary 2.21, we obtain

$$\tilde{G}(t) - \tilde{g}(0) = -\mathbb{E}\left[\int_0^t \|\nabla f(X(s))\|^2 ds\right] + \frac{1}{2}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s)) Hess(f(X(s)))] ds\right]. \tag{3.21}$$

Analyzing the right hand side of (3.21), we conclude that is differentiable, to see this, we must look at the following term

$$-\frac{d}{dt}\mathbb{E}\left[\int_0^t \|\nabla f(X(s))\|^2 ds\right] + \frac{1}{2}\frac{d}{dt}\mathbb{E}\left[\int_0^t tr[\Sigma(s, X(s)) Hess(f(X(s)))] ds\right]. \tag{3.22}$$

We have for almost all $\omega \in \Omega$ that

for some $R^2 > 0$ since X is bounded. The equality is given by [24, Theorem 4.10] and the inequality is given by the fact that ∇f is $L - Lipschitz$. Similarly, for almost all $\omega \in \Omega$ we have

$$\left|\frac{d}{dt}\int_0^t tr[\Sigma(s, X(s)) Hess(f(X(s)))] ds\right| = tr[\Sigma(t, X(t)) Hess(f(X(t)))] \leq \sigma_*^2 L.$$

Where the inequality is given by Corollary 3.2. Thus, applying Proposition 3.3 twice, with

$$D = \mathbb{R}_+, g_1(\omega, t) = \int_0^t \|\nabla f(X(\omega, s))\|^2 ds \text{ and } g_2(\omega, t) = \int_0^t tr[\Sigma(s, X(\omega, s)) Hess(f(X(\omega, s)))] ds,$$

we obtain that (3.22) is equal to

$$-\mathbb{E}\left[\|\nabla f(X(t))\|^2\right] + \frac{1}{2}\mathbb{E}\left[tr[\Sigma(t, X(t)) Hess(f(X(t)))]\right].$$

So the right hand side of (3.21) is differentiable, then we can derivate both sides

$$\begin{aligned} \dot{\tilde{G}}(t) &= -\mathbb{E}\left[\|\nabla f(X(t))\|^2\right] + \frac{1}{2}\mathbb{E}\left[tr[\Sigma(t, X(t)) Hess(f(X(s)))]\right] \\ &\leq -\mathbb{E}\left[\|\nabla f(X(t))\|^2\right] + \frac{\sigma_*^2 L}{2}, \ \ \forall t \geq 0. \end{aligned} \tag{3.23}$$

Where the inequality comes from Corollary 3.2. Then using that $f \in Ł_b^{1/2}(\mathbb{R}^d)$ we have

$$\dot{\tilde{G}}(t) \leq -\mu^2 \tilde{G}(t) + \frac{\sigma_*^2 L}{2}, \ \ \forall t \geq 0.$$

Now we can solve the ordinary differential inequality by using the integrating factor method and obtain

$$\tilde{G}(t) \leq (f(X_0) - min(f))e^{-\mu^2 t} + \frac{\sigma_*^2 L}{2\mu^2}, \ \ \forall t \geq 0. \tag{3.24}$$

Moreover, since $f \in Ł_b^{1/2}(\mathbb{R}^d)$ we can use the inequality of Proposition 1.14 in (3.24) and obtain

$$\mathbb{E}\left(\frac{d(X(t), S)^2}{2}\right) \leq \frac{2(f(X_0) - min(f))e^{-\mu^2 t}}{\mu^2} + \frac{\sigma_*^2 L}{\mu^4}, \quad \forall t \geq 0, \tag{3.25}$$

which is another upper bound.

(iv) Inequality (3.23) still holds in this case because $f \in Ł_b^{1/2}(\mathbb{R}^d)$ it is not used yet, using $f \in Ł_b^q(\mathbb{R}^d), q \in (\frac{1}{2}, 1)$ (and μ as in iii)) we obtain

$$\dot{\tilde{G}}(t) \leq -\mu^2 \mathbb{E}(\tilde{g}(t)^{2q}) + \frac{\sigma_*^2 L}{2}, \ \ \forall t \geq 0.$$

Using $q > \frac{1}{2}$, by Jensen's inequality we have $\tilde{G}(t)^{2q} \leq \mathbb{E}(\tilde{g}(t)^{2q})$, then

$$\dot{\tilde{G}}(t) \leq -\mu^2 \tilde{G}(t)^{2q} + \frac{\sigma_*^2 L}{2}, \ \ \forall t \geq 0. \tag{3.26}$$

By Proposition 3.4 with $a = \mu^2, b = 2q, c = \frac{\sigma_*^2 L}{2}, y(0) = f(X_0) - min(f)$ we have that $\tilde{G}$ satisfies

$$\tilde{G}(t) \leq \left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1/2q} + (f(X_0) - min(f))e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-1/2q} t}, \ \ \forall t \geq 0.$$

This is

$$\mathbb{E}[f(X(t)) - min(f)] \leq (f(X_0) - min(f))e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-1/2q} t} + \left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1/2q}, \ \ \forall t \geq 0. \tag{3.27}$$

Moreover, since $f \in \Gamma_0(\mathbb{R}^d) \cap Ł_b^q(\mathbb{R}^d)$ we can use the inequality of Proposition 1.14 in (3.27) and obtain that there exists $\tilde{\mu} > 0$ such that

$$\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \frac{1}{2\tilde{\mu}^{2/p}}\left((f(X_0) - min(f))e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-1/2q} t} + \left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1/2q}\right)^{2/p}, \ \ \forall t \geq 0, \tag{3.28}$$

and we get another upper bound.

□

Finally, we are ready to deduce the complexities associated with the upper bounds shown in the previous Propositions.

Corollary 3.8 (Complexities of Proposition 3.6) *Let $\varepsilon_0 > 0$ and $\varepsilon \in (0, \varepsilon_0)$ arbitrary.*

(i) If $\sigma_*^2 = \mathcal{O}(\varepsilon), t = \Omega_0\left(\frac{1}{\varepsilon}\right)$. Since $f \in \Gamma_0(H)$ then

$$\mathbb{E}\left[\bar{f}(t)\right] - min(f) \leq \varepsilon.$$

(ii) If $f \in \Gamma_\mu(\mathbb{R}^{\mathrm{d}})$, $\sigma_*^2 = \mathcal{O}(\varepsilon), t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$. Then

$$\mathbb{E}\left(\frac{\|X(t) - x^*\|^2}{2}\right) \leq \varepsilon.$$

Proof. (i) In order to guarantee $\mathbb{E}\left[\bar{f}(t)\right] - min(f) \leq \varepsilon$, we will require that:

$$\frac{\sigma_*^2}{2} \leq \frac{\varepsilon}{2} \quad \text{and} \quad \frac{\|X_0 - x^*\|^2}{2t} \leq \frac{\varepsilon}{2}.$$

Then $\sigma_*^2 \leq \varepsilon$ and

$$\begin{aligned} \frac{\|X_0 - x^*\|^2}{2t} &\leq \frac{\varepsilon}{2} \\ \iff \frac{1}{t} &\leq \frac{\varepsilon}{\|X_0 - x^*\|^2} \\ \iff t &\geq \frac{\|X_0 - x^*\|^2}{\varepsilon}. \end{aligned}$$

So by taking a particular $t = \Omega_0\left(\frac{1}{\varepsilon}\right)$, we can satisfy the last inequality and we conclude.

(ii) In order to guarantee $\mathbb{E}\left(\frac{\|X(t)-x^*\|^2}{2}\right) \leq \varepsilon$, we will require that:

$$\frac{\sigma_*^2}{2\mu} \leq \frac{\varepsilon}{2} \quad \text{and} \quad \frac{\|X_0 - x^*\|^2}{2} e^{-\mu t} \leq \frac{\varepsilon}{2}.$$

Then $\sigma_*^2 \leq \mu\varepsilon$ and

$$\begin{aligned} \frac{\|X_0 - x^*\|^2}{2} e^{-\mu t} &\leq \frac{\varepsilon}{2} \\ \iff e^{-\mu t} &\leq \frac{\varepsilon}{\|X_0 - x^*\|^2} \\ \iff -\mu t &\leq ln\left(\frac{\varepsilon}{\|X_0 - x^*\|^2}\right) \\ \iff t &\geq \frac{1}{\mu} ln\left(\frac{\|X_0 - x^*\|^2}{\varepsilon}\right). \end{aligned}$$

So by taking a particular $t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$, we can satisfy the last inequality and we conclude.

□

Corollary 3.9 (Complexities of Proposition 3.7) *Let $\varepsilon_0 > 0$ and $\varepsilon \in (0, \varepsilon_0)$ arbitrary.*

(i) If $f \in \text{Ł}_b^{1/2}(\mathbb{R}^{\mathrm{d}}), \sigma_*^2 = \mathcal{O}(\varepsilon)$ and $t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$. Then

$$\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \varepsilon.$$

(ii) If $f \in \text{Ł}_b^q(\mathbb{R}^{\mathrm{d}})$, $q \in (\frac{1}{2}, 1)$, $\sigma_*^2 = \mathcal{O}(\varepsilon^{\frac{1}{2(1-q)}})$, $t = \tilde{\Omega}_0\left(\frac{1}{\varepsilon^{\frac{2q-1}{2(1-q)}}}\right)$. Then

$$\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \varepsilon.$$

(iii) If $f \in C_L^{1.1}(\mathbb{R}^{\mathrm{d}}) \cap \text{Ł}_b^{1/2}(\mathbb{R}^{\mathrm{d}})$, $\sigma_*^2 = \mathcal{O}(\varepsilon)$, $t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$. Then

$$\mathbb{E}\left[f(X(t)) - min(f)\right] \leq \varepsilon.$$

(iv) If $f \in C_L^{1.1}(\mathbb{R}^{\mathrm{d}}) \cap \text{Ł}_b^q(\mathbb{R}^{\mathrm{d}})$, $q \in (\frac{1}{2}, 1)$, $\sigma_*^2 = \mathcal{O}(\varepsilon^{2q})$, $t = \tilde{\Omega}_0\left(\frac{1}{\varepsilon^{2q-1}}\right)$. Then

$$\mathbb{E}\left[f(X(t)) - min(f)\right] \leq \varepsilon.$$

Proof. (i) In order to guarantee $\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \varepsilon$, we will require that:

$$\frac{2\sigma_*^2}{\mu^2} \leq \frac{\varepsilon}{2} \quad \text{and} \quad \frac{d(X_0,S)^2}{2} e^{-\mu^2 t} \leq \frac{\varepsilon}{2}.$$

Then $\sigma_*^2 \leq \frac{\mu^2 \varepsilon}{4}$ and

$$\begin{aligned}
\frac{d(X_0,S)^2}{2} e^{-\mu^2 t} &\leq \frac{\varepsilon}{2} \\
\iff e^{-\mu^2 t} &\leq \frac{\varepsilon}{d(X_0,S)^2} \\
\iff -\mu^2 t &\leq ln\left(\frac{\varepsilon}{d(X_0,S)^2}\right) \\
\iff t &\geq \frac{1}{\mu^2} ln\left(\frac{d(X_0,S)^2}{\varepsilon}\right).
\end{aligned}$$

So by taking a particular $t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$, we can satisfy the last inequality and we conclude.

(ii) Recall that $p = \frac{1}{1-q} \geq 1$, in order to guarantee $\mathbb{E}\left(\frac{d(X(t),S)^2}{2}\right) \leq \varepsilon$, we will require that:

$$\frac{1}{2}\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{2/p} = \frac{\varepsilon}{2} \quad \text{and} \quad \frac{d(X_0,S)^2}{2} e^{-\tilde{\mu} p \left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{1-2/p} t} \leq \frac{\varepsilon}{2}.$$

Then $\sigma_*^2 = \tilde{\mu}\varepsilon^{\frac{p}{2}}$ and

$$\begin{aligned}
\frac{d(X_0,S)^2}{2} e^{-\tilde{\mu}p\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{1-2/p}t} &\leq \frac{\varepsilon}{2} \\
\iff e^{-\tilde{\mu}p\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{1-2/p}t} &\leq \frac{\varepsilon}{d(X_0,S)^2} \\
\iff -\tilde{\mu}p\left(\frac{\sigma_*^2}{\tilde{\mu}}\right)^{1-2/p}t &\leq ln\left(\frac{\varepsilon}{d(X_0,S)^2}\right) \\
\iff -\tilde{\mu}p\varepsilon^{\frac{p}{2}-1}t &\leq ln\left(\frac{\varepsilon}{d(X_0,S)^2}\right) \\
\iff t &\geq \frac{1}{\tilde{\mu}p}\frac{ln\left(\frac{d(X_0,S)^2}{\varepsilon}\right)}{\varepsilon^{\frac{p}{2}-1}}.
\end{aligned}$$

So by taking a particular $t = \Omega_0\left(\frac{ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon^{\frac{p}{2}-1}}\right)$, we can satisfy the last inequality and we conclude.

(iii) In order to guarantee $\mathbb{E}\left[f(X(t)) - min(f)\right] \leq \varepsilon$, we will require that:

$$\frac{\sigma_*^2 L}{\mu^2} \leq \frac{\varepsilon}{2} \quad \text{and} \quad \frac{\|X_0 - x^*\|^2}{2} e^{-\mu t} \leq \frac{\varepsilon}{2}.$$

Then $\sigma_*^2 \leq \frac{\mu^2\varepsilon}{2L}$ and

$$\begin{aligned}
(f(X_0) - min(f))e^{-\mu^2 t} &\leq \frac{\varepsilon}{2} \\
\iff e^{-\mu^2 t} &\leq \frac{\varepsilon}{2(f(X_0) - min(f))} \\
\iff -\mu^2 t &\leq ln\left(\frac{\varepsilon}{2(f(X_0) - min(f))}\right) \\
\iff t &\geq \frac{1}{\mu^2} ln\left(\frac{2(f(X_0) - min(f))}{\varepsilon}\right).
\end{aligned}$$

So by taking a particular $t = \Omega_0\left(ln\left(\frac{1}{\varepsilon}\right)\right)$, we can satisfy the last inequality and we conclude.

(iv) In order to guarantee $\mathbb{E}\left[f(X(t)) - min(f)\right] \leq \varepsilon$, we will require that:

$$\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{\frac{1}{2q}} = \frac{\varepsilon}{2} \quad \text{and} \quad (f(X_0) - min(f))e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-\frac{1}{2q}}t} \leq \frac{\varepsilon}{2}.$$

Then $\sigma_*^2 = \frac{2^{1-2q}\mu^2}{L}\varepsilon^{2q}$ and

$$\begin{aligned}
(f(X_0) - min(f))\, e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-\frac{1}{2q}}t} &\leq \frac{\varepsilon}{2} \\
\iff e^{-2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-\frac{1}{2q}}t} &\leq \frac{\varepsilon}{2(f(X_0) - min(f))} \\
\iff -2\mu^2 q\left(\frac{\sigma_*^2 L}{2\mu^2}\right)^{1-\frac{1}{2q}} t &\leq ln\left(\frac{\varepsilon}{2(f(X_0) - min(f)}\right) \\
\iff -2\mu^2 q\left(\frac{\varepsilon}{2}\right)^{2q-1} t &\leq ln\left(\frac{\varepsilon}{2(f(X_0) - min(f)}\right) \\
\iff t &\geq \frac{2^{2(q-1)}}{\mu^2 q}\frac{ln\left(\frac{2(f(X_0)-min(f)}{\varepsilon}\right)}{\varepsilon^{2q-1}}.
\end{aligned}$$

So by taking a particular $t = \Omega_0\left(\frac{ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon^{2q-1}}\right)$, we can satisfy the last inequality and we conclude.

□

In this Chapter, we have shown and proved remarkable results about upper bounds and complexities of (CSGD). As some final observations we can mention the following: Firstly, the forms of the upper bounds are similar to the ones shown for (SGD) (see Propositions 2.9,2.13,2.17). Moreover, Corollary 3.8 ensures us that the same complexity rates as in (GD) and (CGD) will be obtained for the convex and strongly convex case. Furthermore, Corollary 3.9 ensures us the same holds for the Łojasiewicz case with $q \in [\frac{1}{2}, 1)$.

With these results we conclude the proposed framework of this Thesis, after having studied convergences rates, upper bounds and complexities of algorithms such as: Gradient Descent (GD), Continuous Gradient Descent (CGD), Stochastic Gradient Descent (SGD) and a Continuous-time model of *SGD* (CSGD). The contributions of this Thesis were the upper bounds and complexities of (SGD) under Łojasiewicz assumptions, and all results on upper bounds and complexities related to (CSGD).

Conclusions and Future Work

3.3 Conclusions

Convergence rates and complexities of Gradient Descent were recalled under different properties of the objetive function such as: convexity, strong convexity and Bounded Łojasiewicz Inequality. It can be seen that the strong convexity and the Bounded Łojasiewicz Inequality with $q = \frac{1}{2}$ case, share the same convergence rates. On the other hand, if f satisfies the Bounded Łojasiewicz Inequality with $q \in (\frac{1}{2}, 1)$, the closer q is to $\frac{1}{2}$, the better is the complexity, and as q is closer to 1, the complexity worsens until the limit ($q = 1$), where is obtained the result of the convex case (ε^{-1}), which is the worse case for our deterministic algorithms.

The connection of Gradient Descent with its continuous version and its properties was made clear, displaying that they share the same convergence rates and complexities under convexity, strong convexity and Bounded Łojasiewicz Inequality. Stochastic algorithms such as (SGD) were studied, showing classic results in the convex and strongly convex case, and providing new results in the Łojasiewicz case. In the table below are shown the complexities of the algorithms already mentioned.

Property	Complexity GD	Complexity CGD	Complexity SGD
Γ_0	ε^{-1}	ε^{-1}	ε^{-2}
Γ_μ	$ln(\varepsilon^{-1})$	$ln(\varepsilon^{-1})$	ε^{-1}
$Ł_b^{1/2}$	$ln(\varepsilon^{-1})$	$ln(\varepsilon^{-1})$	ε^{-1}
$Ł_b^q$, $q \in (1/2, 1)$	$\varepsilon^{-(2q-1)}$	$\varepsilon^{-(2q-1)}$	$\varepsilon^{-(4q-1)}$

The following comments and observations can be made about this table: in (SGD) the $\sigma^2 < \infty$ hypothesis (finite variance of ∇f_s at $x^* \in S$) was assumed for the convex and strongly convex case. Nevertheless, for the Łojasiewicz case the $\bar{\sigma}^2 < \infty$ hypothesis (finite second moment of ∇f_s in the entire sequence) was assumed. The complexity of (SGD) for

the Łojasiewicz case when $q \in (\frac{1}{2}, 1)$, gets progressively worse as q grows and ranges from ε^{-1} to ε^{-3}. The column of complexities of (SGD) looks different than the others columns because there is a constant that depends on a fixed variance (σ^2 or $\bar{\sigma}^2$) and on a stepsize γ, which appears as a sum in the upper bounds. So, in order to make the constant term small enough to obtain a complexity, we have to impose a small stepsize, this makes the algorithm go slower and this effect is reflected in their complexities, which are worse in this case rather than the others ((GD) and (CGD)) and they cannot be improved (see Propositions 2.7 and 2.8).

Inspired by [19], [23], [25] and [26] we showed a Continuous-Time Model of SGD (CSGD) and we calculated interesting upper bounds, noticing that we had to use similar propositions to those in the (SGD) case (see Lemma 2.16 and Proposition 3.4). Although at first glance these results are not easy to interpret, if we assume that we can make the term σ_*^2 arbitrarily small, then we can obtain the following complexities:

Property	Complexity CSGD (with $\sigma_*^2 = \mathcal{O}(\varepsilon)$)
Γ_0	ε^{-1}
Γ_μ	$ln(\varepsilon^{-1})$
$Ł_b^{1/2}$	$ln(\varepsilon^{-1})$
$Ł_b^q$, $q \in (1/2, 1)$	$\varepsilon^{-(2q-1)}$

The first thing we can notice is that this complexities match when comparing the results of Gradient Descent (GD) or Continuous Gradient Descent (CGD) presented in the first table. However, they do not match the results of Stochastic Gradient Descent (SGD) because, for the Continuous SGD (CSGD) we assumed that we could make σ_*^2 arbitrarily small and that in (SGD), $\bar{\sigma}^2$ was fixed. The assumption about σ_*^2 was made in order to get a complexity (analogous to asking for the stepsize to be small enough in the (SGD) case).

We can also notice that the upper bounds of (CSGD) in the Łojasiewicz case seem like a linear convergence rate plus a constant depending on σ_*^2 which needs to be small enough, since the involved exponential also depends on σ_*^2, it makes the dynamic move much slower than linear. This is clearly seen when viewing the complexity in the table. So, on one hand, the mechanics used to obtain complexities of (CSGD) are similar to the ones used in (SGD) in terms of upper bounds and getting worse complexities, but on the other, these complexities coincide with those obtained in the classic cases ((GD) and (CGD)).

3.4 Future Work

There are a lot of ways to improve or extend the results of this Thesis, or to ask other interesting questions that are not solved yet.

A first step would be to find upper bounds and complexities for Proposition 2.17 without the hypothesis $\bar{\sigma}^2 < \infty$ and only with the hypothesis $\sigma^2 < \infty$ and $\mathcal{L}-$smooth.

In a different direction, if we consider X a solution of (CSGD), then the assumption that X is bounded is strong. For instance, if we consider that f is constant, then the solution of (CSGD) is "proportional" to the Brownian motion, which is not bounded. So, we should be able to find sufficient conditions on f which ensure that the solutions of (CSGD) are bounded a.s. or such that they are bounded with high probability. This is hard, and so far, most people just make this assumption (see [27],[28]). On the other hand, up until now we only have results in expectation, so one path to explore would be to look for results in probability and in a.s. context.

A concept that does not appear in this Thesis is the KŁ property, which is a generalization of the Łojasiewicz Inequality. Convergence rates under KŁ property can be seen in [1],[2]. It might be interesting to find upper bounds or complexities of (CSGD) under KŁ property (and convexity), since this property is usually sufficient to ensure better convergence rates results in the deterministic case. Also, in this case, KŁ property implies that the sequence (trajectory) has finite length, this is useful to not just have convergence on the objectives but also on the iterates (trajectories). So in (CSGD) under this property, we should look for convergence rates/upper bounds results of the trajectories and not just on the objectives.

We must keep in mind that (CSGD) is just a continuous-time model of (MB-SGD) under a Gaussian assumption. It could be interesting to get more results of the obtained SDE without the Gaussian assumption. In this scenario, (CSGD) is driven not by a Brownian motion but by an $\alpha-$stable Lévy motion, a result about this can be seen in [19, Theorem 3].

Finally, a topic that is briefly mentioned in this Thesis is the variance reduction technique, applied in algorithms such as SVRG and SAGA. They can reach the same convergence rates as the classic deterministic algorithms in the convex and the strongly convex case. An open problem would be to describe their convergence rates under assumptions as Łojasiewicz Inequality, expecting a convergence rate of the order of $\mathcal{O}(k^{-\frac{1}{2q-1}})$ for an exponent $q > \frac{1}{2}$. The next step after that would be to describe their convergence rates under the KŁ property in general. An interesting approach to solve this problem would be to try to adapt the Lyapunov functions that are used in [6].

Bibliography

[1] From error bounds to the complexity of first-order descent methods for convex functions. In Jon Lee and Sven Leyffer, editors, *Mathematical Programming*, volume 165, page 471–507. Springer.

[2] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2014.

[3] Juan Peypouquet. De la teoría de semigrupos al análisis variacional numérico. 2021.

[4] Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *MIT Press*, 2011.

[5] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Training neural networks for and by interpolation. *International Conference on Machine Learning 2020*, 2020.

[6] Aaron Defazio. New optimisation methods for machine learning. *arXiv: Machine Learning*, 2016.

[7] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 2016.

[8] Jean Bernard Baillon and Georges Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel J. Math.*, pages 137–150, 1977.

[9] Guillaume Garrigos, Lorenzo Rosasco, and Silvia Villa. Convergence of the forward-backward algorithm: Beyond the worst-case with the help of geometry. *arXiv: Optimization and Control*, 2020.

[10] Haim Brézis. Opérateurs maximaux monotones et semi-groupes de contractions dans les espace hilbert. *North-Holland mathematics studies; 5.*, 1973.

[11] Juan Peypouquet and Sylvain Sorin. Evolution equations for maximal monotone operators: Asymptotic analysis in continuous and discrete time. *Journal of Convex Analysis*, (17):1113–1163, 2009.

[12] Herbert Robins and Sutton Monro. A stochastic approximation method. *Ann. Math.*

Statist. 22, pages 400–407, 1951.

[13] Robert Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. *Proceedings of the 36th International Conference on Machine Learning, PMLR 97*, pages 5200–5209, 2019.

[14] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv: Machine Learning*, 2012.

[15] Arkadi Nemirovsky, Anatoli Juditsky, and Guanghui Lan. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[16] Alekh Agarwal, Peter Bartlett, Pradeep Ravikumar, and Martin Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *arXiv: Machine Learning*, 2011.

[17] Arkadi Nemirovsky and David Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience series in discrete mathematics*, 1983.

[18] Rick Durrett. Probability: theory and examples. *Cambridge university press*, 2010.

[19] Umut Simsekli, Mert Gurbuzbalaban, Thanh Huy Nguyen, Gael Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv: Machine Learning*, 2019.

[20] Bernt Øksendal. Stochastic differential equations. *Springer*, 2003.

[21] Xuerong Mao. Stochastic differential equations and applications. *Elsevier*, 2007.

[22] Herbert Federer. Geometric measure theory, die grundlehren der mathematischen wissenschaften. *Springer*, 1969.

[23] Antonio Orvieto and Lucchi Aurelien. Continuous-time models for stochastic optimization algorithms. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[24] Sergei Ovchinnikov. *Measure, integral, derivative: a course on Lebesgues theory.*

[25] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

[26] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. *arXiv: Machine Learning*, 2017.

[27] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Springer*, 2018.

[28] Steffen Dereich and Sebastian Kassing. Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes. *arXiv: Machine Learning*, 2021.