A Study of Condition Numbers for First-Order Optimization

Charles Guille-Escuret*
Mila,
Université de Montréal

Baptiste Goujaud*
Mila

Manuela Girotti
Mila,
Université de Montréal,
Concordia University

Ioannis Mitliagkas
Mila,
Université de Montréal,
Canada CIFAR AI chair

Abstract

The study of first-order optimization algorithms (FOA) typically starts with assumptions on the objective functions, most commonly smoothness and strong convexity. These metrics are used to tune the hyperparameters of FOA. We introduce a class of perturbations quantified via a new norm, called *-norm. We show that adding a small perturbation to the objective function has an equivalently small impact on the behavior of any FOA, which suggests that it should have a minor impact on the tuning of the algorithm. However, we show that smoothness and strong convexity can be heavily impacted by arbitrarily small perturbations, leading to excessively conservative tunings and convergence issues. In view of these observations, we propose a notion of continuity of the metrics, which is essential for a robust tuning strategy. Since smoothness and strong convexity are not continuous, we propose a comprehensive study of existing alternative metrics which we prove to be continuous. We describe their mutual relations and provide their guaranteed convergence rates for the Gradient Descent algorithm accordingly tuned. Finally we discuss how our work impacts the theoretical understanding of FOA and their performances.

1 Introduction

Optimization of a high-dimensional cost function is at the core of fitting most machine learning models. In practice this is almost always performed by gradient-based first-order optimization algorithms (FOA). The analysis of their convergence properties typically assumes that their hyper-parameters are tuned based on some function properties; for example it is well-known that if f is μ-strongly convex and L-smooth, then gradient descent (GD) with step size $\alpha = \frac{2}{\mu+L}$ achieves a global linear convergence rate of $1-\frac{2}{\kappa+1}$ where $\kappa = \frac{L}{\mu}$ is called the condition number (see e.g. Nesterov (2004)). The condition number gives an indication of the tightness of the bounds on the curvature of f, and therefore of the difficulty to optimize it: the bigger the value of κ is, the slowest is the convergence of the algorithm.

In Lessard et al. (2016), the authors introduce a piecewise quadratic function $f_{\mathrm{LRP}} \in C^1(\mathbb{R})$, with a smaller second-order derivative for $x \in [1, 2]$ than elsewhere. They show that the Heavy-Ball (HB) algorithm (Polyak, 1964), when tuned using the L of smoothness and the μ of strong convexity of f_{LRP}, does not converge to the unique absolute minimizer $x = 0$ for some initialization $x_0 > 0$. On the other hand, standard GD with step size $\alpha = \frac{2}{\mu+L}$ does converge, but at a very low rate due to the high condition number of f_{LRP}. Although tuning the HB algorithm based on the L-smoothness and μ-strong convexity of f_{LRP} is arguably a heuristic strategy (since this optimal tuning rule is only provided for *quadratic* functions, see Polyak (1987)), the example in Lessard et al. (2016) highlights a striking pathological behaviour: a localized, bounded perturbation of the Hessian of the objective function yields a disastrous effect on its condition number and on the trajectory of the iterates of the FOA.

In this paper we analyze this phenomenon and we propose a unifying framework to study the convergence of FOA and design robust tunings of their hyperparameters. We first introduce a new topology, based on the definition of a star-norm $\|\cdot\|_*$ (Section 4.1). Such a norm will be the fundamental tool that will be used throughout the paper in order to assess the "closeness" between objective functions: Theorem 4.4 states that two functions whose difference is small in the $\|\cdot\|_*$-norm sense have comparable behaviour under continuous FOA. Therefore, the tuning strategy for the

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s). *equal contribution.

hyperparameters of the optimization algorithm should account for this similarity. However, the standard tuning based on smoothness and strong convexity fails to do so (Theorem 4.10) and it is easy to construct examples that illustrate this weakness.

Based on such a topology, we then define the notion of continuity of a condition number (Section 4.3), which in turn reflects the continuity of some properties of the objective function that we call *upper/lower conditions* (Section 5), smoothness and strong convexity being two examples of them. Having a continuous condition number is essential to the robustness of both tuning methods and convergence rates of the FOA. Our approach implies that even when the objective function verifies some of the strongest conditions (strong convexity and smoothness), relying on weaker ones to tune the FOA can lead to better and more consistent convergence behaviours.

2 Related Work

Because strong convexity and smoothness are strong requirements that are not verified by some classic machine learning models such as logistic regression (which verifies convexity but not strong convexity), several works have already explored substitute assumptions. Alternatives to strong convexity, which we will call *lower conditions* have been the most thoroughly studied, under some overlapping names. These include *local-quasi-convexity* (Hazan et al., 2015), *weak quasi-convexity* (Hardt et al., 2018), *restricted secant inequality* RSI (Zhang and Yin, 2013), *error bounds* EB (Luo and Tseng, 1993), *quadratic growth* QG (Anitescu, 1999), *Polyak-Łojasiewicz* PL (Polyak, 1963), further generalized as *Kurdyka-Łojasiewicz* KL (Kurdyka, 1998; Bolte et al., 2008). The scattering of these notions in the literature has led to some confusing names. For example, *optimal strong convexity* OSC (Liu and Wright, 2015) is also called *semi-strong convexity* (Gong and Ye, 2014) and *weak strong convexity* (Ma et al., 2016), despite being a different notion from the *weak strong convexity* of Karimi et al. (2016), which was formerly called *quasi-strong convexity* QSC in Necoara et al. (2019). Similarly, the *restricted strong-convexity* from Agarwal et al. (2012) is a different notion from the *restricted strong convexity* of Zhang and Yin (2013). To avoid further confusion, we will use the name *star-strong convexity* *SC for the notion of WSC/QSC of Karimi et al. (2016).

Alternatives to smoothness, which we will call *upper conditions*, have also been proposed, though more sporadically, such as *local smoothness* (Hazan et al., 2015), *restricted smoothness* (Agarwal et al., 2012), *relative smoothness* (Lu et al., 2018; Hanzely et al., 2018; Zhou et al., 2019), *restricted Lipschitz-continuous gradient* RLG (Zhang and Yin, 2013).

Most *lower conditions* can naturally be translated into an equivalent *upper condition*, by shifting the inequality from a lower bound to an upper bound. For example, smoothness is an upper condition equivalent to strong convexity, and *weak-smoothness* (Hardt et al., 2018) is an upper condition equivalent of the PL condition, which is further generalized to the stochastic case as *expected smoothness* in Gower et al. (2019). Similarly, RSI, WSC, EB, and QG all have natural equivalent *upper conditions*. In an attempt to reduce the number of similar names and their associated confusion, we will for instance refer to the PL condition as $\mathrm{PL}^-(\mu)$ and to its equivalent *upper condition* as $\mathrm{PL}^+(L)$.

In Karimi et al. (2016) the authors propose a study of the implications between some *lower conditions*, although under the assumption of global smoothness, and omitting the constant conversion induced by the implications. To the best of our knowledge, a study of the implications between *upper conditions* is missing from the literature. We collect all the relations between *upper* and *lower* conditions in two implication graphs (Figure 1), together with the constant conversions. We also study upper bounds on the convergence rates of gradient descent assuming that the objective function satisfies each pair of upper/lower conditions (Table 1).

While alternative conditions have been extensively researched, the main goal of the works mentioned above has always been to extend convergence results to a larger class of functions. On the other hand, our work aims at introducing a new approach for tackling the optimization task and at bringing a deeper understanding on the convergence of FOA and its connection with properties of the objective function itself.

3 Setup and notation

In this paper, we focus on minimizing an objective function $f : \mathbb{R}^d \to \mathbb{R}$ using first-order algorithms (FOA). The objective function is assumed to be continously differentiable $f \in C^1(\mathbb{R}^d)$, with a convex set of global minima $X^* \subseteq \mathbb{R}^d$; we denote $f^* = \min_{x \in \mathbb{R}^d} f(x)$. For $x \in \mathbb{R}^d$, we denote the distance between x and X^* as $d(x, X^*) = \inf_{x^* \in X^*} \|x - x^*\|_2$. We recall that since X^* is convex, for every $x \in \mathbb{R}^d$ there exists a unique element $x_p^* \in X^*$ (called the projection of x onto X^*) such that $\|x - x_p^*\|_2 = d(x, X^*)$.

For our analysis we will consider the following class of deterministic FOA:

Definition 3.1 (Continuous FOA). A first-order algorithm $\mathcal{A}_\theta$, possibly depending on a set of hyperparameters θ, is continuous if $\forall\, n \in \mathbb{N}$ the $(n+1)$-iterate

$$x_{n+1} = \mathcal{A}_\theta\Big(\{x_i\}_{i=0\ldots n}, \{f(x_i)\}_{i=0\ldots n}, \{\nabla f(x_i)\}_{i=0\ldots n}\Big), \tag{1}$$

is continuous with respect to all of its arguments.

Trivially, any algorithm that can be expressed as a finite composition of continuous operations is continuous. This class of FOA includes all the major algorithms like GD and Heavy Ball (HB) methods with step size and momentum hyperparameters not depending on the local values of f. However, some methods like Polyak step size (Polyak, 1987) are not guaranteed to be continuous without additional assumptions on the objective function.

We denote $\mathcal{B}(X^*, r) = \{y \in \mathbb{R}^d |\, d(y, X^*) < r\}$ to be the set of points in $\mathbb{R}^d$ whose distance from set X^* is smaller than r and, for a set of functions $\mathcal{F}$, we denote $f + \mathcal{F}$ the set of functions g such that $g - f \in \mathcal{F}$.

Unless stated otherwise, rates of convergence refer to the convergence of $f(x_n) - f^*$, not $d(x_n, X^*)$.

4 Continuity of first-order algorithms and condition numbers

In this section we introduce the theoretical framework to analyze the behaviour of FOA for objective functions that are "close". The first necessary component is a norm $\|\cdot\|_*$ that will induce the right kind of topology to evaluate the similarity between objective functions. Proofs of all key results are collected in Appendix A.

4.1 Star norm and stability of FOA behaviors

Consider an objective function $f \in C^1(\mathbb{R}^d)$. The purpose of the $\|\cdot\|_*$-norm will be to evaluate the impact of a perturbation of f on the convergence properties of FOA. In particular, if two functions f and g are such that $\|f - g\|_*$ is small, it is desirable for the FOA to behave similarly on them. Since we are focussing on optimization algorithms that depend on the first derivatives of the function, we require the $\|\cdot\|_*$-norm to give some control over the amplitude of the gradient of the perturbation of f. Additionally, notice that as the iterates approach the minima of the objective function, the updates typically become finer, so that even a small perturbation of the function gradient can greatly affect the convergence behaviour. This supports the intuition that the same perturbation of the gradient will have more impact close to the set of minima X^*, and less impact far away.

In view of the above discussion, we introduce the following definition of the $\|\cdot\|_*$-norm, which measures the maximal perturbation of the gradient weighted by the inverse of the distance to X^*.

Definition 4.1 (Star norm). Let $X^* \subseteq \mathbb{R}^d$ and

$$\begin{aligned}\mathcal{F}_{X^*} = \{h \in C^1(\mathbb{R}^d) \mid \forall\, x^* \in X^*, h(x^*) = 0 \text{ and}\\ \exists\, L \in \mathbb{R} : \|\nabla h(x)\|_2 \le L\ d(x, X^*), \forall\, x \in \mathbb{R}^d\}.\end{aligned}$$

We define the *star norm*, $\|\cdot\|_*$, on $\mathcal{F}_{X^*}$ as

$$\forall\, h \in \mathcal{F}_{X^*}, \quad \|h\|_* = \sup_{x \in \mathbb{R}^d \setminus X^*} \frac{\|\nabla h(x)\|_2}{d(x, X^*)}. \tag{2}$$

Example 4.2. Consider the function $h(x) = \sqrt{x^2+1} - 1$, which can be thought as a differentiable version of the absolute value function; then, $h \in \mathcal{F}_{\{0\}}$, with $\|h\|_* = 1$.

Remark 4.3. We emphasize that neither X^* nor $\mathcal{F}_{X^*}$ depend of the objective function f, which does not need to be in $\mathcal{F}_{X^*}$ itself. Requiring $h(x^*) = 0$ ensures that the $\|\cdot\|_*$-norm is indeed a norm on $\mathcal{F}_{X^*}$. Equivalently, we could have considered the quotient space ${\mathcal{F}_{X^*}}/{[c]}$, where $[c]$ is the set of constant functions, equipped with $\|\cdot\|_*$; however, this would have introduced too many technicalities along the paper, therefore we did not proceed in this direction.

Let $x_i(f, \mathcal{A}_\theta, x_0)$ denote the i-th iterate obtained by applying a prescribed algorithm $\mathcal{A}_\theta$ to f starting in x_0. We now argue that two functions that are close in the sense of the star norm will have similar behaviors for continuous FOA.

Theorem 4.4. *Let $f \in C^1(\mathbb{R}^d)$ with a set of global minimizers X^* and $\|\cdot\|_*$ the corresponding star norm. Let $\mathcal{A}_\theta$ be a continuous first-order algorithm and $\mathcal{K} \subset \mathbb{R}^d$ a compact set. Then, the following result holds:*

$$\begin{gathered}\forall\, \epsilon > 0,\ \forall\, i \in \mathbb{N}, \exists\, \eta = \eta(\epsilon, i, \mathcal{K}) > 0 \text{ such that}\\ \forall\, h \in \mathcal{F}_{X^*},\ \text{if } \|h\|_* < \eta,\ \text{then}\\ \forall\, x_0 \in \mathcal{K} : \|x_i(f, \mathcal{A}_\theta, x_0) - x_i(f+h, \mathcal{A}_\theta, x_0)\|_2 < \epsilon.\end{gathered}$$

The following corollary proves that for a target neighborhood of X^* and any $\delta > 0$, if h is sufficiently small in the sense of $\|\cdot\|_*$, then $\forall\, x_0 \in \mathcal{K}$, applying $\mathcal{A}_\theta$ to $f + h$ starting in x_0 will attain the target neighborhood in exactly the same number of steps as for f, up to a distance tolerance of δ.

Corollary 4.5. *Under the same hypotheses as Theorem 4.4, let $\varepsilon > 0$ and $\mathcal{B}(X^*, \varepsilon)$ a target neighborhood of X^*. Let us assume that $\mathcal{A}_\theta$ applied to f converges to X^* and $\forall\, x_0 \in \mathcal{K}$, let $N_{x_0} \in \mathbb{N}$ the smallest number of iterations such that $x_{N_{x_0}}(f, \mathcal{A}_\theta, x_0) \in \mathcal{B}(X^*, \varepsilon)$. Then, $\forall\, \delta > 0$, $\exists\, \eta > 0$ s.t. for any $h \in \mathcal{F}_{X^*}$, if $\|h\|_* < \eta$, then $\forall\, x_0 \in \mathcal{K}$,*

$$\begin{gathered}x_{N_{x_0}-1}(f+h, \mathcal{A}_\theta, x_0) \notin \mathcal{B}(X^*, \varepsilon - \delta) \quad \text{and}\\ x_{N_{x_0}}(f+h, \mathcal{A}_\theta, x_0) \in \mathcal{B}(X^*, \varepsilon + \delta).\end{gathered}$$

Theorem 4.4 and Corollary 4.5 show that if h is sufficiently small in the sense of the norm $\|.\|_*$, then the behaviour of a continuous FOA on f and $f+h$ will be similar, and thus it is natural to assume that the tuning of hyperparameters θ should also be similar. However, as the next section shows, this is not always the case.

4.2 Standard tuning fails continuity test

Consider the family of piecewise quadratic functions $\{f_\varepsilon\}_{\varepsilon\geq 0}\subset C^1(\mathbb{R})$:

$$f_\varepsilon(x)=\begin{cases}x^2 & x\leq 1\\ x^2+\frac{1}{\varepsilon}x^2-\frac{2x}{\varepsilon}+\frac{1}{\varepsilon} & 1\leq x\leq 1+\varepsilon^2\\ x^2+2\varepsilon x-2\varepsilon-\varepsilon^3 & x\geq 1+\varepsilon^2\end{cases} \quad (3)$$

We can view each function f_ε as a perturbation of the quadratic $f_0(x)=x^2$, which is 2-smooth and 2-strongly convex: $\forall\,\varepsilon\geq 0$, $f_\varepsilon(x)=f_0(x)+h_\varepsilon(x)$ with $h_\varepsilon\in\mathcal{F}_{X^*=\{0\}}$. It is also easy to see that $\|h_\varepsilon\|_*\to 0$ as $\varepsilon\to 0$.

The following properties hold:

Proposition 4.6. *For any $\varepsilon>0$, the function f_ε is μ_ε-strong convex and L_ε-smooth, with $\mu_\varepsilon=2$ and $L_\varepsilon=2+\frac{2}{\varepsilon}$; moreover, these constants are optimal: i.e. f_ε is not μ-strongly convex for $\mu>2$ and not L-smooth for $L<2+\frac{2}{\varepsilon}$.*

Furthermore, GD tuned with step size $\alpha=\frac{2}{\mu_0+L_0}=\frac{1}{2}$ applied to f_ε ($\forall\,\epsilon>0$) converges with linear rate ε; however, if GD is tuned with $\alpha=\frac{2}{\mu_\varepsilon+L_\varepsilon}=\frac{\varepsilon}{2\varepsilon+1}$, it does not converge with linear rate q for any $q<\frac{1-\varepsilon^2}{(2\varepsilon+1)(1+\varepsilon^2)}$.

If we tune GD according to the values of smoothness and strong convexity of f_0 and optimize f_ε, the linear rate tends to 0 as $\varepsilon\to 0$ (in fact, we obtain convergence in at most two steps). On the other hand, if we tune GD based on the tightest strong convexity and smoothness constants μ_ε and L_ε of f_ε, the linear rate tends to 1 as $\varepsilon\to 0$. Notice that the condition number $\frac{L_\varepsilon}{\mu_\varepsilon}$ of f_ε diverges as $\varepsilon\to 0$, thus leading to a very conservative tuning and increasingly slow convergence rate, while the tuning of f_0 leads to superlinear convergence.

The above example suggests that a sane tuning strategy for the hyperparameters of a FOA should be robust (continuous) with respect to $\|\cdot\|_*$-small perturbations of a given function. It also shows that the standard tuning based on L-smoothness and μ-strong convexity lacks this property.

4.3 Continuity of condition numbers

We now formally introduce the notions of upper and lower conditions which represent generalizations of smoothness and strong convexity, and the notion of continuity of a condition.

Definition 4.7 (Upper conditions). We use the term *upper condition* to describe a generalization of smoothness and we formalize it as a family of sets of functions, $\mathcal{C}^+(L)\subseteq C^1(\mathbb{R}^d)$, which satisfies $\mathcal{C}^+(L_1)\subseteq\mathcal{C}^+(L_2)$ for all $L_1\leq L_2$.

Definition 4.8 (Lower conditions). We use the term *lower condition* to describe a generalization of strong convexity. We formalize it as a family of sets of functions, $\mathcal{C}^-(\mu)\subseteq C^1(\mathbb{R}^d)$, which satisfies $\mathcal{C}^-(\mu_1)\supseteq\mathcal{C}^-(\mu_2)$ for all $\mu_1\leq\mu_2$.

In Definition 5.1 and Definition 5.3, we list some known upper and lower conditions extensively studied in the literature .

Definition 4.9 (Continuity of a condition). We say that $\mathcal{C}^+$ is continuous in $f\in\bigcup_{L>0}\mathcal{C}^+(L)$ with convex set of global minima X^* if for any $L>0$ s.t. $f\in\mathcal{C}^+(L)$, $\forall\,\varepsilon>0,\exists\,\eta>0$ s.t. $\forall\,h\in\mathcal{F}_{X^*}$, if $\|h\|_*\leq\eta$, then $f+h\in\mathcal{C}^+(L+\varepsilon)$.

Similarly, $\mathcal{C}^-$ is continuous in $f\in\bigcup_{\mu>0}\mathcal{C}^-(\mu)$ with set of global minima X^* if for any $\mu>0$ s.t. $f\in\mathcal{C}^-(\mu)$, $\forall\,\varepsilon>0,\exists\,\eta>0$ s.t. $\forall\,h\in\mathcal{F}_{X^*}$, if $\|h\|_*\leq\eta$, then $f+h\in\mathcal{C}^-(\mu-\varepsilon)$.

We say that $\mathcal{C}^+$ is continuous if it is continuous in all $f\in\bigcup_{L>0}\mathcal{C}^+(L)$ that admits a convex set of global minima, and $\mathcal{C}^-$ is continuous if it is continuous in all $f\in\bigcup_{\mu>0}\mathcal{C}^-(\mu)$ that admits a convex set of global minima.

Note that this definition is independent from the standard notion of continuity, as we only allow f to be approximated by functions in $f+\mathcal{F}_{X^*}$.

Based on the observations of Theorem 4.4 and Corollary 4.5, if we tune a continuous FOA based on a condition $\mathcal{C}^\pm$, it is desirable for $\mathcal{C}^\pm$ to be continuous in the sense we just introduced. However, the standard properties of smoothness and strong convexity fail to be continuous:

Theorem 4.10. *For any f $\bar{\mu}$-strongly convex and $\bar{L}$-smooth with a set of global minima $X^*\subseteq\mathbb{R}^d$, there exists a family $\{h_\varepsilon\}_{\varepsilon>0}$ in $\mathcal{F}_{X^*}$ such that $\lim_{\varepsilon\to 0}\|h_\varepsilon\|_*=0$ and $\forall\,L,\mu>0$, there is $\varepsilon_{L,\mu}$ such that $\forall\,\varepsilon\leq\varepsilon_{L,\mu}$, $f_\varepsilon=f+h_\varepsilon$ is not L-smooth and not μ-strongly convex.*

Not only smoothness and strong convexity are continuous nowhere, but also the discontinuity is not bounded: given any objective function f, it is possible to approximate it by a family of perturbed functions $\{f_\varepsilon\}_{\varepsilon>0}$ with arbitrarily bad conditioning. In particular, the explicitly construction of $\{f_\varepsilon\}_{\varepsilon>0}$ is given in the proof. Therefore, the main consequence of Theorem 4.10 is that tunings that rely on smoothness and strong convexity lack robustness.

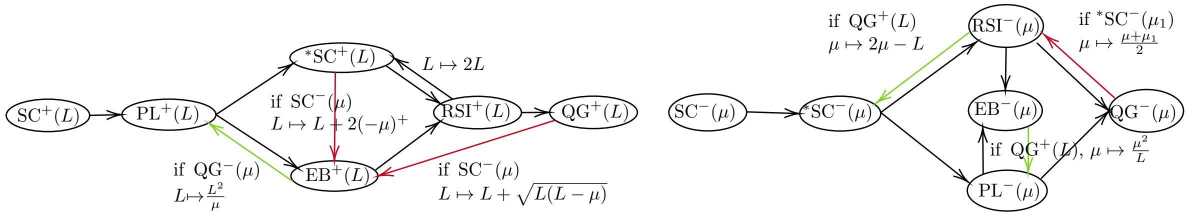

Figure 1: Graph of implications between upper and lower conditions. Red arrows only hold under $^*\mathrm{SC}^-(\mu)$ or $\mathrm{SC}^-(\mu)$ where μ can be negative. Green arrows only hold under QG^+ or QG^-.

5 Alternative conditioning

Motivated by the weakness of strong convexity and smoothness detailed in Subsection 4.2 and in Theorem 4.10, we propose here known alternative conditions that could be used to tune FOA.

Let $f \in C^1\left(\mathbb{R}^d\right)$ with convex set of minimizer X^*. We recall that any $x \in \mathbb{R}^d$ has an unique projection $x_p^* \in X^*$ on X^*: $\|x - x_p^*\|_2 = d(x, X^*)$.

Definition 5.1 (Lower conditions)**.** Let $\mu > 0$. We define:

- *(Strong convexity)* $f \in \mathrm{SC}^-(\mu)$ iff $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2$, $\forall x, y \in \mathbb{R}^d$.
- *(Star strong convexity)* $f \in {}^*\mathrm{SC}^-(\mu)$ iff $f^* \geq f(x) + \langle \nabla f(x), x_p^* - x\rangle + \frac{\mu}{2}\left\|x_p^* - x\right\|_2^2$, $\forall x \in \mathbb{R}^d$.
- *(Lower restricted secant inequality)* $f \in \mathrm{RSI}^-(\mu)$ iff $\langle \nabla f(x), x - x_p^*\rangle \geq \mu\left\|x - x_p^*\right\|_2^2$, $\forall x \in \mathbb{R}^d$.
- *(Lower error bound)* $f \in \mathrm{EB}^-(\mu)$ iff $\|\nabla f(x)\|_2 \geq \mu\left\|x - x_p^*\right\|_2$, $\forall x \in \mathbb{R}^d$.
- *(Lower Polyak-Łojasiewicz)* $f \in \mathrm{PL}^-(\mu)$ iff $\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu\left(f(x) - f^*\right)$, $\forall x \in \mathbb{R}^d$.
- *(Lower quadratic growth)* $f \in \mathrm{QG}^-(\mu)$ iff $f(x) - f^* \geq \frac{\mu}{2}\left\|x - x_p^*\right\|_2^2$, $\forall x \in \mathbb{R}^d$.

Remark 5.2. A function in $\mathrm{SC}^-(0)$ is called *convex* and a function in $^*\mathrm{SC}^-(0)$ is called *star-convex*. Additionally, if the inequality in the definition of $\mathrm{SC}^-(0)$ is strict, then f is *strictly convex*.

Definition 5.3 (Upper conditions)**.** Let $L > 0$. We define:

- *(Smoothness)* $f \in \mathrm{SC}^+(L)$ iff $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2$, $\forall x, y \in \mathbb{R}^d$.
- *(Star smoothness)* $f \in {}^*\mathrm{SC}^+(L)$ iff $f^* \leq f(x) + \langle \nabla f(x), x_p^* - x\rangle + \frac{L}{2}\left\|x_p^* - x\right\|_2^2$, $\forall x \in \mathbb{R}^d$.
- *(Upper restricted secant inequality)* $f \in \mathrm{RSI}^+(L)$ iff $\langle \nabla f(x), x - x_p^*\rangle \leq L\left\|x - x_p^*\right\|_2^2$, $\forall x \in \mathbb{R}^d$.
- *(Upper error bound)* $f \in \mathrm{EB}^+(L)$ iff $\|\nabla f(x)\|_2 \leq L\left\|x - x_p^*\right\|_2$, $\forall x \in \mathbb{R}^d$.
- *(Upper Polyak-Łojasiewicz)* $f \in \mathrm{PL}^+(L)$ iff $\frac{1}{2}\|\nabla f(x)\|_2^2 \leq L\left(f(x) - f^*\right)$, $\forall x \in \mathbb{R}^d$.
- *(Upper quadratic growth)* $f \in \mathrm{QG}^+(L)$ iff $f(x) - f^* \leq \frac{L}{2}\left\|x - x_p^*\right\|_2^2$, $\forall x \in \mathbb{R}^d$.

Remark 5.4. The proposed upper and lower conditions all coincide on quadratics, with optimal L and μ equal to the highest and lowest eigenvalues of the Hessian, respectively.

The upper and lower conditions above are related according to the graphs in Figure 1 (see proofs in Appendices C and D). If an implication changes the value of the constant, it is specified on the corresponding arrow. Some of the implications only hold under extended notions of $^*\mathrm{SC}^-(\mu)$ and $\mathrm{SC}^-(\mu)$, where μ is allowed to be negative (red arrows in Figure 1). These notions are weaker than star convexity and convexity, respectively. Finally some implications are made under an additional QG^+ or QG^- assumption (green arrows in Figure 1).

In Karimi et al. (2016), the authors already presented connections between the lower conditions, but under the assumption of global smoothness ($\mathrm{SC}^+(L)$) and without giving the conversion of constants. To the best of our knowledge, there is no study of the implications between upper conditions in the literature.

Theorem 4.10 showed smoothness and strong convexity are not continuous in the sense of Definition 4.9. On the other hand, the above alternatives are continuous conditions, therefore they are robust to the type of perturbations introduced in Section 4.1:

Theorem 5.5. *The lower conditions $^*\mathrm{SC}^-$, RSI^-, EB^-, QG^-, PL^- are continuous. The upper conditions $^*\mathrm{SC}^+$, RSI^+, EB^+, QG^+, PL^+ are continuous in all functions $f \in \mathrm{QG}^-(\mu)$, for some $\mu > 0$.*

Proof. See Appendix B.

Note that since $\mathrm{SC}^-(\mu)$, $^*\mathrm{SC}^-(\mu)$, $\mathrm{PL}^-(\mu)$, $\mathrm{RSI}^-(\mu)$, $\mathrm{EB}^-(\mu) \subset \mathrm{QG}^-(\frac{\mu^2}{L})$ (see Figure 1), $^*\mathrm{SC}^+$, RSI^+, EB^+,

Table 1: Linear rates for the GD algorithm for each pair of conditions, as function of $\kappa = \frac{L}{\mu}$. Rates marked with * hold under the additional assumption of star-convexity, while rates marked with † hold under the additional assumption of convexity. Rates are colored in green if corresponding to a continuous pair of conditions and red otherwise.

Rates of cv	$\mathrm{SC}^-(\mu)$	$^*\mathrm{SC}^-(\mu)$	$\mathrm{PL}^-(\mu)$	$\mathrm{RSI}^-(\mu)$	$\mathrm{EB}^-(\mu)$	$\mathrm{QG}^-(\mu)$
$\mathrm{SC}^+(L)$	$\left(\frac{\kappa-1}{\kappa+1}\right)^2$	$1-\frac{1}{\kappa}$	$1-\frac{1}{\kappa}$	$1-\frac{1}{\kappa^2}$ / $1-\frac{1}{2\kappa}$ $*$	$1-\frac{1}{\kappa^2}$	$1-\frac{1}{4\kappa}$ $*$
$\mathrm{PL}^+(L)$	$\left(1-\frac{1}{\kappa}\right)^2$	$1-\frac{1}{\kappa}$	$1-\frac{1}{4\kappa}$ $*$	$1-\frac{1}{\kappa^2}$ / $1-\frac{1}{2\kappa}$ $*$	$1-\frac{1}{4\kappa^2}$ $*$	$1-\frac{1}{4\kappa}$ $*$
$\mathrm{EB}^+(L)$	$\left(1-\frac{1}{\kappa}\right)^2$	$1-\frac{1}{\kappa^2}$	$1-\frac{1}{4\kappa^2}$ $*$	$1-\frac{1}{\kappa^2}$	$1-\frac{1}{4\kappa^4}$ $*$	$1-\frac{1}{4\kappa^2}$ $*$
$^*\mathrm{SC}^+(L)$	$\left(1-\frac{1}{\kappa}\right)^2$	$1-\frac{1}{\kappa^2}$ $\dagger$	$1-\frac{1}{4\kappa^2}$ $\dagger$	$1-\frac{1}{\kappa^2}$ $\dagger$	$1-\frac{1}{4\kappa^4}$ $\dagger$	$1-\frac{1}{4\kappa^2}$ $\dagger$
$\mathrm{RSI}^+(L)$	$\left(1-\frac{1}{\kappa}\right)^2$	$1-\frac{1}{4\kappa^2}$ $\dagger$	$1-\frac{1}{16\kappa^2}$ $\dagger$	$1-\frac{1}{4\kappa^2}$ $\dagger$	$1-\frac{1}{16\kappa^4}$ $\dagger$	$1-\frac{1}{16\kappa^2}$ $\dagger$
$\mathrm{QG}^+(L)$	$\left(1-\frac{1}{\kappa}\right)^2$	$1-\frac{1}{4\kappa^2}$ $\dagger$	$1-\frac{1}{16\kappa^2}$ $\dagger$	$1-\frac{1}{4\kappa^2}$ $\dagger$	$1-\frac{1}{16\kappa^4}$ $\dagger$	$1-\frac{1}{16\kappa^2}$ $\dagger$

QG^+, PL^+ are continuous in any function that verifies one of the proposed lower conditions.

6 Gradient descent convergence

To give some insights on the strengths of the listed conditions, we collected in Table 1 the guaranteed linear convergence rates of $f(x_n) - f^*$ of the GD algorithm with constant step size and proper tuning, obtained for each pair of upper/lower conditions $f \in \mathcal{C}^+(L) \cap \mathcal{C}^-(\mu)$, as function of the condition number $\kappa = \frac{L}{\mu}$. The conditions are ordered from the strongest to the weakest, when applicable. The rates that are marked with an asterisk or a $\dagger$ symbol are only guaranteed under an additional assumption of convexity or star convexity, respectively. Many of these rates do not exist in the literature to the best of our knowledge. In particular, the rates under $\mathrm{QG}^+(L) \cap \mathrm{SC}^-(\mu)$, and all the rates inherited from the known ones, as in Figure 1, are novel. The rate under $\mathrm{PL}^+ \cap {}^*\mathrm{SC}^-(\mu)$ is a particular case of Theorem 3.1 in Gower et al. (2019) applied to the deterministic case. For the sake of completeness, we reported rates under additional convexity assumption, although convexity suffers from the same continuity issue as strong convexity and smoothness.

We refer to Appendix E for the proofs; the exact value of the step size for the convergence of GD under each pair of upper/lower conditions is also given.

Some care needs to be taken when comparing the κ's from different entries of the table, as the quantities involved (L and μ) differ according to the upper/lower conditions considered. Notice that the condition $\mathrm{PL}^+(L)$ paired with any lower condition shows a convergence rate with the same dependence in κ as $\mathrm{SC}^+(L)$, with the added bonus that $\mathrm{PL}^+(L)$ is continuous. Additionally, the pair $\mathrm{EB}^+(L) \cap \mathrm{RSI}^-(\mu)$ has a linear rate that depends quadratically in κ, however, this pair of conditions is weaker than other pairs ($\mathrm{PL}^+(L)$, $\mathrm{PL}^-(\mu)$, $^*\mathrm{SC}^-(\mu)$), therefore the condition number κ for this case might be drastically smaller and it may yield a better convergence rate. Thus, these two pairs $\mathrm{PL}^+(L) \cap \mathcal{C}^-(\mu)$ and $\mathrm{EB}^+(L) \cap \mathrm{RSI}^-(\mu)$ look particularly promising for effectively tuning the step size of the GD algorithm.

We recall that quadratics give a lower bound $\left(\frac{\kappa-1}{\kappa+1}\right)^2$ for the convergence rate for GD with fixed step size. Since all the conditions listed in this paper coincide on quadratics, such a lower bound applies to any pair of upper/lower conditions. However, it may be not tight for some pairs of conditions.

Finally, we complete Table 1 by mentioning the sublinear convergence speed we have under any upper condition and convexity or star-convexity ($\mathcal{C}^+(L) \cap \mathrm{SC}^-(0)$ or $\mathcal{C}^+(L) \cap {}^*\mathrm{SC}^-(0)$). While it is known that GD has a rate of convergence of order $\mathcal{O}\left(\frac{1}{n}\right)$ if $f \in \mathrm{SC}^+(L) \cap \mathrm{SC}^-(0)$ (see e.g. Bansal and Gupta (2017)), the same rate can be achieved under $\mathrm{PL}^+(L) \cap {}^*\mathrm{SC}^-(0)$ for the best iterate (or the average under convexity). For a complete proof, we refer to Appendix E.

In Ghadimi et al. (2015), the authors prove the same rate of convergence under convexity and $\mathrm{QG}^+(L)$ for the average iterate. They also obtain the same rate under those very weak conditions for the last iterate using an extra momentum term (following the heavy ball procedure).

7 Discussion

In this section we discuss how the use of alternative conditions impacts our understanding of the behavior of Polyak's Heavy-Ball method (Polyak, 1964):

$$x_{n+1} = x_n - \alpha \nabla f(x_n) + \beta(x_n - x_{n-1}) \tag{4}$$

where the step size α and the momentum β are the hyperparameters. It is well known that the optimal hyperparameters of HB on a strongly convex *quadratic* function with minimum eigenvalue μ and maximum eigenvalue L of the Hessian are:

$$\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}, \qquad \beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2 \tag{5}$$

with $\kappa = \frac{L}{\mu}$. This tuning yields a linear convergence rate for $f(x_n) - f^*$ equal to $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$, which approaches the lower bound for L-smooth and μ-strongly convex functions (Bubeck, 2015).

In Lessard et al. (2016) the authors introduced the following piecewise quadratic function: $f_{\mathrm{LRP}} \in C^1(\mathbb{R})$ with derivative

$$f'_{\mathrm{LRP}}(x) = \begin{cases} 25x & x < 1 \\ x + 24 & 1 \leq x \leq 2 \\ 25x - 24 & x > 2 \end{cases} \tag{6}$$

The authors showed that for the initial value $x_0 = 3.3$ and using the same tuning rule (5) but with the L of smoothness and the μ of strong convexity, HB does not converge.

While applying the tuning rule (5) to non-quadratic functions itself is arbitrary, the choice of smoothness and strong convexity values as generalizations of the maximum and minimum eigenvalues is particularly problematic: all conditions introduced in Section 5 coincide on quadratic functions and there is no strong evidence to prefer $\mathrm{SC}^-(\mu)$ and $\mathrm{SC}^+(L)$ as tuning conditions.

Since this function has been used as an example of inconsistent behavior from HB, it is natural to question how using continuous conditions as generalizations of the biggest and smallest eigenvalues of a quadratic function may affect the convergence.

All the upper conditions on f_{LRP} give the same parameter $L = 25$, while the lower conditions give $\mu_{\mathrm{SC}^-} = 1$, $\mu_{^*\mathrm{SC}^-} = 7$, $\mu_{\mathrm{RSI}^-} = \mu_{EB^-} = 13$, $\mu_{\mathrm{PL}^-} = \frac{169}{19}$ and $\mu_{\mathrm{QG}^-} = 19$.

In Figure 2 we present the linear convergence rates experimentally obtained for 200,000 values of (α, β). We also indicate the hyperparameters corresponding to tuning rule (5) using different lower conditions. We immediately observe that while the generalization based on strong convexity falls into the black region (no linear convergence), other conditions all offer excellent convergence properties. This might suggest that the divergent behavior is caused by relying on strong convexity for tuning rather than by the algorithm itself, and highlights how weaker conditions are essential to understand FOA behaviors, even on functions that verify smoothness and strong convexity.

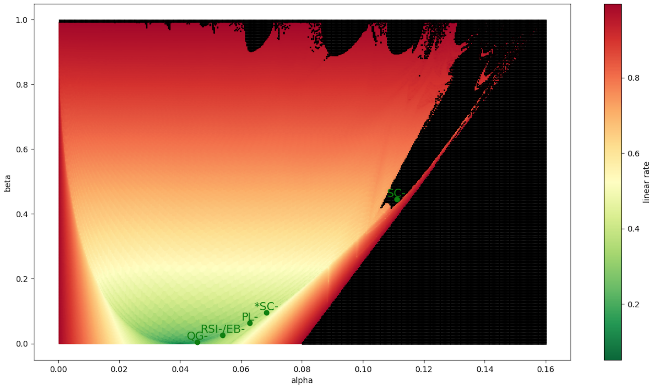

Figure 2: Convergence rate of HB on f_{LRP} with starting point $x_0 = 3.3$ for different tunings of α, β. Black areas mean no linear convergence

8 Conclusion

In this paper we presented an argument on the necessity to adopt different conditions from the ones classically used (smoothness and strong convexity), in order to tune the hyperparameters of FOA in a meaningful way. Via a new notion of continuity of a condition number, we have established that the properties of strong convexity and smoothness have an important weakness resulting in a lack of robustness for first-order algorithms tuned on them. We have presented promising alternatives that do not share this weakness and given examples of the benefits of a theoretical framework based on these conditions. We have proposed an extensive study of the relationships between these conditions and provided their guaranteed convergence rates for GD.

The study of the convergence properties of optimization algorithms largely depends on their tuning, hence understanding its underlying conditions leads to a better comparison between them and improves the algorithm performances, as illustrated in Section 7.

While it is well known that some optimization algorithms (e.g. Nesterov Accelerated Gradient, Nesterov (1983)) can approach the lower bound of convergence rates achievable for μ-strongly convex and L-smooth functions, as function of $\kappa = \frac{L}{\mu}$, lower bounds based on alternative condition numbers will result in different optimality results, which we leave to future work.

Acknowledgements

We thank Nicolas Loizou for his valuable advice. We also thank Leonard Boussioux and Aymeric Dieuleveut for their proofreading and useful feedback. We are grateful to the anonymous reviewers for their valuable comments in improving the clarity of our manuscript.

This work was partially supported by the FRQNT new researcher program (2019-NC-257943), the NSERC Discovery grant (RGPIN-2019-06512), a startup grant by IVADO, a Samsung collaborative grant and a Canada CIFAR AI chair.

References

Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 2012.

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. on Optimization*, 10(4):1116–1135, 1999. ISSN 1052-6234.

Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(\frac{1}{n})$. *Advances in Neural Information Processing Systems (NIPS)*, pages 773–791, 2013.

Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods. *arXiv preprint arXiv:1712.04581*, 2017.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of łojasiewicz inequalities and applications. *arXiv:0802.0826*, 2008.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity.* Foundations and Trends (R) in Machine Learning. Now Publishers, 2015.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 07 2011.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

Pinghua Gong and Jieping Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *arXiv:1406.1102*, 2014.

Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. *arXiv:2006.1031*, 2020.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. SGD: General analysis and improved rates. *arXiv:1901.09401*, 2019.

Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. Technical Report MSR-TR-2018-22, Microsoft, 2018.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19, 2018.

Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1594–1602. Curran Associates, Inc., 2015.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, volume abs/1608.04636, pages 795–811, Cham, 2016. Springer International Publishing.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48:769–783, 1998.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26:57–95, 2016.

Ji Liu and Stephen J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM J. Optim.*, 25(1):351—376, 2015.

Stanislaw Łojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Seminari di geometria*, 1983:115–117, 1982.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1): 333–354, 2018.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46 (1):157–178, 1993.

Chenxin Ma, Rachael Tappenden, and Martin Takàč. Linear convergence of the randomized feasible descent method under the weak strong convexity assumption. *Journal of Machine Learning Research*, 17(228):1–24, 2016.

Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming volume*, 175:69–107, 2019.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer US, 2004.

Boris T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864 – 878, 1963.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.

Boris T. Polyak. *Introduction to optimization.* Optimization Software, 1987.

Hui Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. *Optimization Letters*, 11(4):817–833, 2017.

Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. Cam report, UCLA, 2013.

Jian Zhang and Ioannis Mitliagkas. YellowFin and the art of momentum tuning. *Systems and ML*, 2019.

Yi Zhou, Yingbin Liang, and Lixin Shen. A simple convergence analysis of bregman proximal gradient algorithm. *Computational Optimization and Applications*, 73(3):903–912, 2019.