MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

From Convexity to Strong Convexity and Beyond: Bridging The Gap In Convergence Rates

Romero, Orlando; Benosman, Mouhacine; Pappas, George

TR2024-131 October 02, 2024

Abstract

In this paper, we re-examine the role of convexity and smoothness on gradient-based unconstrained optimization. While existing literature establishes the fundamental limits for gradient-based optimization algorithms for the class FL of L-smooth convex functions and the subclass Fm,L of L-smooth and m-strongly convex functions, there is a notable gap in the stark transition from their respective sublinear and linear/exponential convergence rates that persists even as m - 0. This gap is notable since the classical rate of 0(1/k) for gradient descent in FL is often overly conservative compared to what is observed in practice for convex functions that are not strongly convex. In this work, we partially close the aforementioned gap by leveraging the notion of uniform smoothness and convexity, and their respective moduli, to quantify and more comprehensively characterize the smoothness and convexity of a given function. We show how, through a simple rescaling of gradient descent informed by the modulus of smoothness, we can recover the classic rates as edge cases and establish novel rates for a wide variety of functions. Further, we examine how uniform convexity can be replaced with the Kurdyka- Lojasiewicz inequality, with the so-called "desingularizing function" replacing the role of the modulus of convexity in the novel rates. This characterization yields novel geometric insights on the relationship between the optimization landscape and the attainable convergence rates.

IEEE Conference on Decision and Control (CDC) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

From Convexity to Strong Convexity and Beyond: Bridging The Gap In Convergence Rates

Orlando Romero[1] Mouhacine Benosman[2] George J. Pappas[1]

***Abstract*— In this paper, we re-examine the role of convexity and smoothness on gradient-based unconstrained optimization. While existing literature establishes the fundamental limits for gradient-based optimization algorithms for the class $\mathcal{F}_L$ of L-smooth convex functions and the subclass $\mathcal{F}_{\mu,L}$ of L-smooth and μ-strongly convex functions, there is a notable gap in the stark transition from their respective sublinear and linear/exponential convergence rates that persists even as $\mu \to 0$. This gap is notable since the classical rate of $\mathcal{O}(1/k)$ for gradient descent in $\mathcal{F}_L$ is often overly conservative compared to what is observed in practice for convex functions that are not strongly convex. In this work, we partially close the aforementioned gap by leveraging the notion of *uniform* smoothness and convexity, and their respective *moduli*, to quantify and more comprehensively characterize the smoothness and convexity of a given function. We show how, through a simple rescaling of gradient descent informed by the modulus of smoothness, we can recover the classic rates as edge cases and establish novel rates for a wide variety of functions. Further, we examine how uniform convexity can be replaced with the Kurdyka-Łojasiewicz inequality, with the so-called "desingularizing function" replacing the role of the modulus of convexity in the novel rates. This characterization yields novel geometric insights on the relationship between the optimization landscape and the attainable convergence rates.**

I. Introduction

Large-scale optimization and convex optimization are integral to applications in a vast range of areas such as deep learning, supply chain management, power systems, scientific computation, and many more [1], [2]. Theoretical understanding of the computational effort needed to solve optimization problems informs us about bottlenecks and scalability concerns, and ultimately provides intuition on how to best tackle these issues [3], [4]. Complexity analysis in optimization is, therefore, crucial for the effective resource management in a multitude of critical applications.

Convexity plays a key role in large-scale optimization, both at the practical and theoretical level, since it is one of the simplest properties, while still broadly applicable, that ensures we can find *global* minimizers numerically for generic functions [5], [6]. Further, convexity enjoys a variety of amenable properties that allows us to efficiently analyze the convergence rate of gradient-based optimization algorithms, particularly when paired with smoothness assumptions [7].

It is well established that, for the class $\mathcal{F}_L$ of L-smooth convex functions, the gradient descent (GD) algorithm $x_{k+1} = x_k - \eta \nabla f(x_k)$ with fixed learning rate $\eta > 0$ achieves a convergence rate $O(1/k)$, provided that $\eta \leq 1/L$ [7]. On the other hand, in the class $\mathcal{F}_{\mu,L}$ of L-smooth and μ-strongly convex functions, the same GD algorithm attains a much faster linear rate $\mathcal{O}\left((1-\eta\mu)^k\right)$ with $\rho = 1 - \eta\mu$ [7], [8]. While the optimization algorithm did not change in the two cases discussed above, the convergence rate is remarkably different. This abrupt transition motivates us to more generally examine the role that smoothness and convexity play in establishing convergence rates.

Related Work

In [8], the authors conduct a comprehensive overview of several alternatives to strong convexity found in the literature at the time, in order to establish linear convergence. It is established that the Polyak-Łojasiewicz (PŁ) inequality [14] is the weakest of all conditions considered that guarantees linear convergence to *global* local minima, with the quadratic growth condition being weaker but admitting non-global minima to exist. Further, the authors establish that all conditions considered are equivalent in the case of convexity. Interestingly, the PŁ inequality yields to an almost trivial proof of linear convergence, as the authors show. However, it is known that a rate $\mathcal{O}\left(\left(\frac{1-\kappa}{1+\kappa}\right)^{2k}\right)$ with $\kappa = \frac{L}{\mu}$ can be established in the class $\mathcal{F}_{\mu,L}$ when using GD with learning rate $\eta = \frac{2}{L+\mu}$, which is faster than the rate $\mathcal{O}\left(\left(1-\frac{1}{\kappa}\right)^k\right)$ derived using the μ-PŁ inequality for GD with learning rate $\eta = 1/L$.

In [9], the authors propose the notion of *strong smoothness* of order $p > 1$ and exploit the notion of *μ-uniform convexity of order p* and the closely related *μ-gradient dominance of order p* (with $p = 2$ corresponding to μ-strong convexity and μ-PŁ, respectively) to establish convergence rates on "descent algorithms of order p", which notably includes their proposed *rescaled gradient descent* algorithm

$$x_{k+1} = x_k - \eta^{\frac{1}{p-1}} \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_*^{\frac{p-2}{p-1}}}, \tag{1}$$

with $\eta > 0$ and $p > 1$. We will consider a unified notion of smoothness and convexity that generalizes the conditions considered in [9], while also being both less restrictive and simplifying the convergence analysis.

The notion of smoothness and convexity that we will consider is borrowed and adapted from the older works [10] and [11], but it should also be noted that [12] recently revisited these conditions (and more), modernized them, and provided new insights (notably, in connection to generalization bounds in statistical learning). A consequence of uniform convexity is the *Kurdyka-Łojasiewicz* (KŁ) inequality, which has been

[1]University of Pennsylvania, Philadelphia, PA.
[2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA. M. Benosman was supported solely by MERL.

extensively studied in [13]. The KŁ inequality has recently been used by [14], [15] to establish new iteration complexity bounds on the stochastic gradient descent algorithm (SGD). In this paper, we will also consider the KŁ inequality, with our results extending those of [14], [15] in the deterministic setting, but with the novelty of now allowing for cost functions that need not be L-smooth.

Contributions

In this work, we revisit the notions of *uniform smoothness* and *uniform convexity* and seek to establish a gradient-based optimization algorithm with provable convergence rates that subsumes the classical rates of $\mathcal{O}\left(\frac{L}{k}\right)$ and $\mathcal{O}\left(\left(1-\frac{\mu}{L}\right)^k\right)$ for GD with learning rate $\eta = \frac{1}{L}$ over the classes $\mathcal{F}_L$ and $\mathcal{F}_{\mu,L}$, respectively. More concretely, our contributions are:

- We leverage uniform smoothness to propose a new gradient-based optimization algorithm, in terms of the *modulus of smoothness* σ, which we call *σ-rescaled gradient descent* (σ-RGD).
- Next, we establish a descent lemma and a generalization of the PŁ inequality.
- Using these tools, we establish the linear convergence rate $f(x_k) - f^\star = \mathcal{O}\left(\left(1-\frac{1}{c}\right)^k\right)$ for σ-RGD under σ-smoothness and ϕ-convexity, where $c = \sup_{s>0} \frac{\phi^*(s)}{\sigma^*(s)} \geq 1$ is a generalized condition number.
- Next, we revisit GD under L-smoothness and reduce ϕ-convexity to φ-KŁ. With this, we establish the rate

$$f(x_k) - f^\star \leq E_\varphi^{-1}\left(E_\varphi(f(x_0) - f^\star) + \frac{\eta k}{2}\right),$$

  where $E_\varphi(s)$ is a function that we introduce as the *desingularizing energy*. This rate allows novel geometric interpretations on the relationship between the optimization landscape and the rates achieved by GD.
- Lastly, we state our most comprehensive result: under σ-smoothness and φ-KŁ, our σ-RGD algorithm satisfies the convergence rate

$$f(x_k) - f^\star \leq E_{\sigma,\varphi}^{-1}\Big(E_{\sigma,\varphi}\big(f(x_0) - f^\star\big) + k\Big),$$

  where $E_{\sigma,\varphi}(s) := \int_s^\infty \frac{1}{\sigma^*(1/\varphi'(r))} dr$.

II. Problem Formulation

We consider the unconstrained optimization problem

$$\min_{x\in\mathcal{X}} f(x) \tag{2}$$

of minimizing $f : \mathcal{X} \to \mathbb{R}$ using a gradient oracle. We will assume that $(\mathcal{X}, \|\cdot\|)$ is a real Banach space with dual space $(\mathcal{X}^*, \|\cdot\|_*)$ and duality pairing $\langle\cdot,\cdot\rangle : \mathcal{X}^* \times \mathcal{X} \to \mathbb{R}$ so that $\|g\|_* = \sup_{\|x\|=1}\langle g, x\rangle$ for $g \in \mathcal{X}^*$. We will assume continuous Gateaux differentiability, meaning that the Gateaux differential $df(x;v) := \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}$ is well defined, continuous (in both $x \in \mathcal{X}$ and $v \in \mathcal{X}$), and linear in v, and thus $df(x;\cdot)$ is an element of $\mathcal{X}^*$. We will re-brand it as the *gradient*, denoted as usual by $\nabla f(x)$, interpreted as the (assumed to exist and be unique) element of $\mathcal{X}^*$ for which $\langle\nabla f(x), v\rangle = \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}$ holds for all $v \in \mathcal{X}$. We will always assume that f is bounded from below, i.e. $f^\star := \inf f > -\infty$.

Recall that a class $\mathcal{K}$ function is a continuous and strictly increasing function $\alpha : [0, r) \to \mathbb{R}$, for some $0 < r \leq \infty$, such that $\alpha(0) = 0$. With this, we introduce the main properties that will be used to study the optimization problem discussed above. The following definitions are based on [12], [11], [10]

Definition 1. We say that f is σ-smooth if

$$f(x) \leq f(y) + \langle\nabla f(y), x - y\rangle + \sigma(\|x - y\|), \tag{3}$$

holds for every $x, y \in \mathcal{X}$, for some differentiable and convex class $\mathcal{K}$ function σ such that $\sigma'(0) = 0$.

Definition 2. We say that f is ϕ-convex if

$$f(x) \geq f(y) + \langle\nabla f(y), x - y\rangle + \phi(\|x - y\|), \tag{4}$$

holds for every $x, y \in \mathcal{X}$, for some differentiable and convex class $\mathcal{K}$ function ϕ such that $\phi'(0) = 0$.

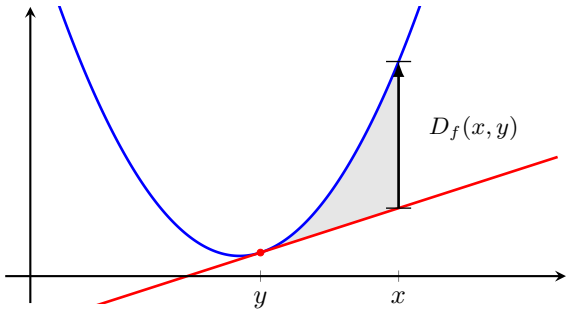

Fig. 1: The Bregman divergence $D_f(x,y) := f(x) - f(y) - \langle\nabla f(y), x-y\rangle$ can be used to characterize ϕ-convexity and σ-smoothnesss. More precisely, f is σ-smooth and ϕ-convex if and only if $\phi(\|x-y\|) \leq D_f(x,y) \leq \sigma(\|x-y\|)$. They characterize the optimization landscape by bounding the curvature of the function.

Let us denote the family of σ-smooth convex functions as $\mathcal{F}_\sigma$, and the subclass of σ-smooth and ϕ-convex functions as $\mathcal{F}_{\sigma,\phi}$. With these definitions, we are ready to state the problem we seek to solve:

Problem 1. Design a 1-step gradient-based optimization algorithm $x_{k+1} = F_\sigma(x_k, \nabla f(x_k))$ for the class $\mathcal{F}_\sigma$, with a provable convergence rate $f(x_k) - f^\star \leq R_{\phi,\sigma}(k, f(x_0) - f^\star)$ over the class $\mathcal{F}_{\phi,\sigma}$. In particular, the algorithm and convergence rate should subsume the rates $f(x_k) - f^\star = \mathcal{O}\left(\frac{L}{k}\right)$ and $f(x_k) - f^\star = \mathcal{O}\left(\left(1-\frac{\mu}{L}\right)^k\right)$ for GD with leargning rate $\eta = \frac{1}{L}$ over the classes $\mathcal{F}_L$ and $\mathcal{F}_{\mu,L}$, respectively.

III. Preliminaries

Before we can study convergence for the gradient-based algorithm that we will propose, let us establish two useful lemmas that will ensure well-behavedness.

Recall that, for a function $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty, +\infty\}$ with co-domain on the extended real line, its *domain* is given

by $\operatorname{dom} f = \{x \in \mathcal{X} : f(x) \in \mathbb{R}\}$. When discussing a class $\mathcal{K}$ function $\alpha : [0, r) \to \infty$, we will implicitly extend it to $\alpha : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, by setting $\alpha(s) = +\infty$ for $s \notin [0, r)$. Additionally, recall that the *convex conjugate* of $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty, +\infty\}$ is given by

$$f^*(y) = \sup_{x \in \mathcal{X}} \{\langle y, x\rangle - f(x)\}$$

for $y \in \mathcal{X}^*$.

Lemma 1. *If α is a differentiable class $\mathcal{K}$ function such that $\alpha'(0) = 0$, then α^* is also a differentiable class $\mathcal{K}$ function.*

Proof. Since α is differentiable in its domain, then so is α^* in its domain. Note that $\alpha^*(0) = \max_t\{0 \cdot t - \alpha(t)\} = 0$ because $\alpha(t) \geq 0$ with $\alpha(0) = 0$. Let $[0, r)$ be the domain of α, with $0 < r \leq \infty$. Let $s \in [0, r)$. By the maximizing argument property, we have

$$(\alpha^*)'(s) \in \operatorname*{arg\,max}_{0 \leq t < r} \{st - \alpha(t)\}$$

and thus, noting that $st - \alpha(t) = 0$ for $t = 0$, it follows that $(\alpha^*)'(s) \geq 0$. Now, note that if $(\alpha^*)'(s) = 0$, then $0 = s \cdot 0 - \alpha(0) \geq st - \alpha(t)$ for all t. Therefore, $s \leq \alpha(t)/t$ for $t > 0$. Taking $t \to 0$, we conclude that $s \leq \alpha'(0) = 0$, and thus $s = 0$. Thus, $(\alpha^*)'(s) \geq 0$ for all $s \in [0, r)$, with equality only at $s = 0$. Therefore, α^* is strictly increasing over its domain. ■

Lemma 2. *If $\alpha : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ has domain within $[0, \infty)$, then $(\alpha \circ \|\cdot\|)^* = \alpha^* \circ \|\cdot\|_*$.*

Proof. Let $g \in \mathcal{X}^*$. Then,

$$\begin{aligned}
(\alpha \circ \|\cdot\|)^*(g) &= \sup_{x \in \mathcal{X}} \{\langle g, x\rangle - \alpha(\|x\|)\} \\
&= \sup_{t \geq 0} \sup_{\|x\|=1} \{\langle g, tx\rangle - \alpha(t)\} \\
&= \sup_{t \geq 0} \left\{ t \sup_{\|x\|=1} \langle g, x\rangle - \alpha(t) \right\} \\
&= \sup_{t \in \operatorname{dom}(\alpha)} \{t\|g\|_* - \alpha(t)\} \\
&= \alpha^*(\|g\|_*) \\
&= (\alpha^* \circ \|\cdot\|_*)(g).
\end{aligned}$$

Assuming σ-smoothness with known σ, we propose the following algorithm, similar in spirit to Nemirovski's mirror descent, which we call σ-rescaled gradient descent (σ-RGD):

$$\begin{aligned}
x_{k+1} &= x_k - z_k \\
z_k &\in \operatorname*{arg\,max}_{z \in \mathcal{X}} \{\langle \nabla f(x_k), z\rangle - \sigma(\|z\|)\}
\end{aligned} \tag{5}$$

Therefore, by the maximizing argument property of convex conjugates, we have

$$x_{k+1} \in x_k - \partial(\sigma \circ \|\cdot\|)^*(\nabla f(x_k)). \tag{6}$$

For the important case of $\mathcal{X} = \mathbb{R}^d$ with the Euclidean norm $\|\cdot\| = \|\cdot\|_2$, we thus have

$$x_{k+1} = x_k - (\sigma^*)'(\|\nabla f(x_k)\|_2) \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}, \tag{7}$$

with the convention that $(\sigma^*)'(\|\nabla f(x)\|_2) \frac{\nabla f(x)}{\|\nabla f(x)\|_2} = 0$ when x is a stationary point of f.

IV. Main Results

Similar to the typical analysis of gradient descent, let us first establish a useful descent lemma.

Proposition 1 (σ-descent lemma)**.** *If f is σ-smooth, then the σ-RGD algorithm* (6) *satisfies the descent condition*

$$f(x_{k+1}) \leq f(x_k) - \sigma^*(\|\nabla f(x_k)\|_*)$$

for $k \in \{0, 1, \ldots\}$.

Proof. By definition of σ-smoothness, we have

$$\begin{aligned}
f(x_{k+1}) &= f(x_k) - \Big(\langle \nabla f(x_k), z_k\rangle - \sigma(\|z_k\|)\Big) \\
&\overset{(5)}{=} f(x_k) - (\sigma \circ \|\cdot\|)^*(\nabla f(x_k)),
\end{aligned}$$

and the result follows by invoking Lemma 2. ■

A. Linear Rate for σ-RGD

In the generalized framework we have adopted, the analysis that leads to a linear convergence rate is less technical and more intuitive, so we will start there. Before we proceed, let us establish a generalization of the Polyak-Łojasiewicz (PŁ) inequality.

Proposition 2 (ϕ-Polyak-Łojasiewicz)**.** *If f is ϕ-convex, then*

$$\|\nabla f(x)\|_* \geq (\phi^*)^{-1}(f(x) - f^\star)$$

for every $x \in \mathcal{X}$.

Proof. For every $y \in \mathcal{X}$, we have

$$\begin{aligned}
f^\star &= \inf_{x \in \mathcal{X}} f(x) \\
&\geq \inf_{x \in \mathcal{X}} \big\{f(y) + \langle \nabla f(y), x - y\rangle + \phi(\|x - y\|)\big\} \\
&= f(y) - \sup_{x \in \mathcal{X}} \big\{\langle \nabla f(y), y - x\rangle + \phi(\|y - x\|)\big\} \\
&= f(y) - (\phi \circ \|\cdot\|)^*(\nabla f(y)) \\
&= f(y) - \phi^*\left(\|\nabla f(y)\|_*\right).
\end{aligned}$$

Relabeling y and rearranging terms, we find that

$$\phi^*\left(\|\nabla f(x)\|_*\right) \geq f(x) - f^\star, \qquad \forall x \in \mathcal{X}$$

and the result follows by invoking Lemma 1. ■

Equipped with this generalization of the PŁ inequality, we can readily combine it with the descent lemma to establish a linear rate of convergence, provided that σ and ϕ satisfy a relationship that is akin to $L \geq \mu$ in the class $\mathcal{F}_{\mu,L}$.

Theorem 1. *If f is σ-smooth and ϕ-convex, with σ, ϕ such that, there exists some $c \geq 1$ for which $\phi(cs) \geq c\sigma(s)$ holds for all $s \geq 0$, then the σ-RGD algorithm* (6) *satisfies*

$$f(x_k) - f^\star \leq \left(1 - \frac{1}{c}\right)^k (f(x_0) - f^\star) \tag{8}$$

for all $k \in \{0, 1, \ldots\}$.

The constant $c \geq 1$ can be understood as a generalized *condition number*. In particular, we can show that, setting

$$c = \sup_{s>0} \frac{\phi^*(s)}{\sigma^*(s)}, \tag{9}$$

the relationship $\phi(cs) \geq c\sigma(s)$ holds. However, such c need not be finite, which would render the rate (8) vacuous. In the the culmination of this section, we will obtain a non-vacuous rate that subsumes the rates $\mathcal{O}(L/k)$ and $\mathcal{O}((1-\mu/L)^k)$ in $\mathcal{F}_L$ and $\mathcal{F}_{\mu,L}$ for vanilla GD with learning rate $\eta = 1/L$.

To illustrate the aforementioned vacuousness, consider functions in the class $\mathcal{F}_{\mu,L}$. In there, we have $\sigma(s) = \frac{L}{2}s^2$ and $\varphi(s) = \frac{\mu}{2}s^2$, and thus

$$c = \sup_{s>0} \frac{(2\mu)^{-1}s^2}{(2L)^{-1}s^2} = \frac{L}{\mu}.$$

Clearly then, as $\mu \to 0$, we find that $c \to \infty$.

Proof. The proof closely resembles the argument used in [8] for the linear convergence of vanilla gradient descent under L-smoothness and the μ-Polyak-Łojasiewicz inequality.

Since $\phi(cs) \geq c\sigma(s)$, then, by the order reversing and scaling properties of the convex conjugate [**?**], it follows that $\phi^*(s/c) \leq c\sigma^*(s/c)$. By a straightforward change of variables, we can see that $\phi^*(s) \leq c\sigma^*(s)$. With this, we have that $\delta_k := f(x_k) - f^\star$ satisfies

$$\begin{aligned}
\delta_{k+1} &\leq \delta_k - \sigma^*(\|\nabla f(x_k)\|_*) && (\sigma\text{-descent lemma}) \\
&\leq \delta_k - \frac{\phi^*(\|\nabla f(x_k)\|_*)}{c} && (\phi^* \leq c\sigma^*) \\
&\leq \delta_k - \frac{\delta_k}{c} && (\phi\text{-Polyak-Łojasiewicz}) \\
&= \left(1 - \frac{1}{c}\right)\delta_k,
\end{aligned}$$

and the result follows by direct recursion. ∎

Example 1. Suppose that f is and (L,p)-smooth, i.e.

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

and (μ,p)-convex, i.e.

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

for some $L \geq \mu > 0$ and $p \geq 2$. Then, the RGD algorithm

$$x_{k+1} = x_k - \eta \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|^{\frac{p-2}{p-1}}}$$

converges at a linear rate

$$\delta_k \leq \left[1 - \eta\left(1 - \frac{L\eta^{p-1}}{p}\right)\frac{p}{p-1}\mu^{\frac{1}{p-1}}\right]^k \delta_0,$$

where $\delta_k = f(x_k) - f^\star$, provided that $0 < \eta \leq (p/L)^{\frac{1}{p-1}}$. For the optimal choice of $\eta > 0$, which is $\eta = 1/L^{\frac{1}{p-1}}$, the rate becomes

$$f(x_k) - f^\star \leq \left(1 - \frac{1}{\kappa^{\frac{1}{p-1}}}\right)^k (f(x_0) - f^\star),$$

where $\kappa := L/\mu$.

B. Novel Rates for Gradient Descent

Let us now consider the case when $\mathcal{X} = \mathbb{R}^d$ is equipped with the usual Euclidean norm, and $\sigma(s) = \frac{L}{2}s^2$. This way, our proposed RGD algorithm reduces to vanilla gradient descent with the optimal learning rate $\eta = \frac{1}{L}$. The results presented in this paper do not heavily depend on the convexity of f, and instead largely follow from the descent inequality in combination with the generalization of the PŁ inequality discussed earlier. This generalization can be seen as a particular case of the more general standalone condition known as the Kurdyka-Łojasiewicz (KŁ) inequality, which notably does not require convexity (much like PŁ). However, in both cases, the condition does require *invexity*, meaning that every stationary point is a global minimizer (assuming that the conditions are to hold globally).

Definition 3 ([13])**.** We say that f satisfies the Kurdyka-Łojasiewicz (φ-KŁ) condition if

$$\|\nabla(\varphi \circ (f - f^\star))\| \geq 1 \tag{10}$$

holds pointwise for some class $\mathcal{K}$ function φ such that $\varphi(s)$ is continuously differentiable except at $s = 0$. The function φ will be referred to as a *desingularizing function* for f at $x^\star$.

Intuitively, the idea is that, when φ is applied to $f - f^\star$, it "bends" the cost function f without changing the location of its minimizer but making the function "sharp" (no longer differentiable) near the minimizer. We can quantify how much "energy" is needed to achieve this by introducing the *desingularizing energy* function

$$E_\varphi(s) = \int_s^\infty \big(\varphi'(r)\big)^2 \mathrm{d}r \tag{11}$$

with domain $s > 0$. Clearly, $E_\varphi(s)$ is strictly decreasing, with $s > 0$ dictating the size of the neighborhood of $x^\star$ to discard in the desingularizing energy. For many functions of interest, we will have $E_\varphi(s) \to \infty$ as $s \to 0$. If the improper integral diverges even for $s > 0$, we can make some minor adjustments to the definition above and the way we use it later on. However, for the sake of simplicity, let us assume that $E_\varphi(s)$ is well defined.

Example 2. Consider the function $f(x) = c\|x - x^\star\|^p$ with $c > 0$, and $p \geq 2$. We can easily verify that $\varphi(s) = c^{-1/p}s^{1/p}$ and $E_\varphi(s) = \frac{c^{-2/p}}{p(p-2)}s^{-\frac{p-2}{p}}$, with $E_\varphi(s) = \infty$ for $p = 2$.

Note that, as $c \to 0$, the function becomes flatter, thus requiring more energy to bend it sharp, as reflected in $\lim_{c\to 0} E_\varphi(s) = \infty$ for every $s > 0$. See Fig. 2a. Likewise, $E_\varphi(s) \sim \frac{1}{p^2}\frac{1}{s}$ as $p \to \infty$, which shows us that the function becomes sharper *away* from $x^\star$ (as reflected by $E_\varphi(s) \to 0$ as $s \to \infty$) and flatter *near* $x^\star$ (as reflected by $E_\varphi(s) \to \infty$ as $s \to 0$). See Fig 2b.

We are ready to state our next main result:

Theorem 2. *Suppose that f is L-smooth and φ-KŁ, with φ strictly concave. Then, the gradient descent (GD) algorithm*

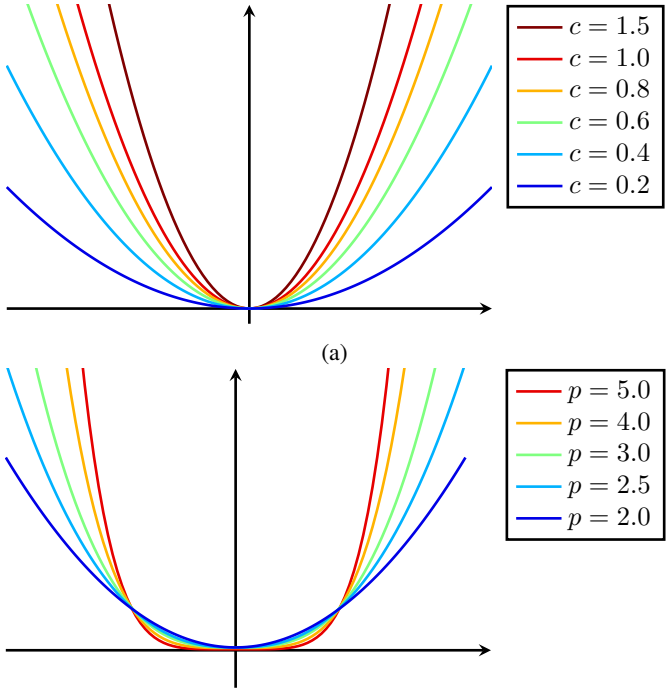

Fig. 2: $f(x) = c\|x - x^\star\|^p$ with varying $c > 0$ and $p \geq 2$.

$x_{k+1} = x_k - \eta \nabla f(x_k)$ *converges at a rate*

$$f(x_k) - f^\star \leq E_\varphi^{-1}\left(E_\varphi(f(x_0) - f^\star) + \frac{\eta k}{2}\right) \tag{12}$$

for all $k \in \{0, 1, \ldots\}$, *provided that* $0 < \eta_k \leq \frac{1}{L}$.

Proof. The condition (10) can be rewritten as $|\varphi'(f(x) - f^\star)\|\nabla f(x)\| \geq 1$. Plugging this inequality in the descent inequality obtained when $\sigma(s) = (L/2)s^2$ yields

$$\delta_{k+1} \leq \delta_k - \frac{\eta_k}{2}\frac{1}{(\varphi'(\delta_k)^2},$$

where $\delta_k := f(x_k) - f^\star$. This difference inequality can be seen as the forward-Euler discretization of the differential inequality $\dot{\delta} \leq -\alpha(\delta)$, with time steps $t_k = kh$ and step sizes $h = \eta/2$, for some class $\mathcal{K}$ function $\alpha(\cdot)$ (in this case, $\alpha(s) = 1/(\varphi'(s))^2$). The proof will consist of carefully comparing $\{\delta_k\}$ with the solution of the worst-case ODE

$$\dot{\delta} = -\alpha(\delta) \tag{13}$$

from one step to the next, and keeping track of the accumulated error. Intuitively, if $\delta(t_0) \approx \delta_0$, then $\delta(t_k) \approx \delta_k$. Let us formalize this by considering the linear interpolation of $\{(t_k, \delta_k) : k = 0, 1, \ldots\}$, given by

$$\hat{\delta}(t) = \sum_{k=0}^{\infty} 1_{t\in[t_k,t_{k+1})}\left[\delta_k + \frac{t - t_k}{h}(\delta_{k+1} - \delta_k)\right] \tag{14}$$

as well as the piecewise solution with jumps, given by

$$\tilde{\delta}(t) = \Big(\text{sol. of (13) with } \delta(t_k) = \delta_k\Big)\Big|_t \tag{15}$$

for each $t \in [t_k, t_{k+1})$. Until otherwise specified, we will now always assume $t \in [t_k, t_{k+1})$ with fixed (but arbitrary) k. In order to proceed, let us first note that

$$\tilde{\delta}(t) = \delta_k + \int_{t_k}^{t} \frac{d}{ds}\tilde{\delta}(s)ds$$

by the Picard-Lindeöf theorem. Therefore, we can bound the local truncation error as follows:

$$\begin{aligned}
\hat{\delta}(t) - \tilde{\delta}(t) &= \frac{t - t_k}{h}(\delta_{k+1} - \delta_k) - \int_{t_k}^{t} \frac{d}{ds}\tilde{\delta}(s)ds \\
&= -(t - t_k)\alpha(\delta_k) + \int_{t_k}^{t} \alpha(\tilde{\delta}(s))ds \\
&\leq -(t - t_k)\alpha(\delta_k) + (t - t_k)\alpha(\tilde{\delta}(t_k)) \\
&= -(t - t_k)\alpha(\delta_k) + (t - t_k)\alpha(\delta_k) \\
&= 0,
\end{aligned}$$

where the inequality originates by noting that $\alpha \circ \tilde{\delta}$ is non-increasing. Subsequently, since k was arbitrary, we have $\hat{\delta}(t) \leq \tilde{\delta}(t)$ for all $t \geq 0$.

To proceed, we first adapt the argument in the proof of Lemma 3.4 in [16] to note that

$$\tilde{\delta}(t) = E_\varphi^{-1}(E_\varphi(\delta_k) + (t - t_k))$$

for $t \in [t_k, t_{k+1})$. Indeed, such $\delta(t)$ must satisfy $E_\varphi(\delta(t)) = E_\varphi(\delta_k) + t - t_k$. Differentiating, we find $-E_\varphi'(\delta(t))\dot{\delta}(t) = 1$. Noting that $E_\varphi'(s) = -(\varphi'(s))^2$, it follows that $\dot{\delta}(t) = -\frac{1}{(\varphi'(\delta(t)))^2} = -\alpha(\delta(t))$. Equipped with this solution, we find that

$$\hat{\delta}(t) \leq E_\varphi^{-1}(E_\varphi(\delta_k) + t - t_k)$$

for $t \in [t_k, t_{k+1})$. Taking the limit $t \to t_{k+1}$ from below and recalling that $t_{k+1} - t_k = h = \frac{\eta}{2}$, we find

$$\delta_{k+1} \leq E_\varphi^{-1}\left(E_\varphi(\delta_k) + \frac{\eta}{2}\right)$$

by noting that $\hat{\delta}(t) \to \delta_{k+1}$ as $t \to t_{k+1}$. We can rewrite the above inequality as $E_\varphi(\delta_{k+1}) \geq E_\varphi(\delta_k) + \frac{\eta}{2}$ (note thet E_φ is non-increasing) and subsequently, perform a telescoping sum (since k was arbitrary), leading to $E_\varphi(\delta_k) \geq E_\varphi(\delta_0) + \frac{\eta k}{2}$. The result follows by rearranging the terms. ■

In essence, our result summarizes how the curvature and overall shape of the function impacts the convergence rate of gradient descent. Specifically, functions with larger desingularizing energy, as induced by unfavorable geometric landscapes (e.g. flatter functions), are likely force GD to converge slower than functions with smaller desingularizing energy (e.g. functions closer to a quadratic one).

(needs touch-up)

Corollary 1. *If* f *is locally* L*-smooth and* (μ, p)*-convex with* $p > 2$*, then, in some neighborhood of the global minimizer* $x^\star$*, the GD algorithm* $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ *converges at a sublinear rate* $f(x_k) - f^\star = \mathcal{O}(1/k^{\frac{p}{p-2}})$.

Proof. The assumed convexity implies that the Lojasiewicz gradient inequality

$$\frac{p-1}{p}\|\nabla f(x)\|^{\frac{p}{p-1}} \geq \mu^{\frac{1}{p-1}}(f(x)-f^\star) \tag{16}$$

holds in some neighborhood of the global minimizer $x^\star$. Therefore, f is φ-KŁ with $\varphi(s) \propto s^{1/p}$. Thus, $E_\varphi(s) \propto 1/s^{\frac{p-2}{p}}$ and therefore $E_\varphi^{-1}(s) \propto 1/s^{\frac{p}{p-2}}$. Therefore, $f(x_k) - f^\star = \mathcal{O}(1/k^{\frac{p}{p-2}})$. ■

C. General Case

With the intuition from the previous two subsections, we can now readily state the general rate for RGD under σ-smoothness and φ-KŁ.

Theorem 3. *If f is σ-smooth and φ-KŁ with strictly concave φ, then the σ-RGD algorithm* (6) *satisfies the convergence rate*

$$f(x_k) - f^\star \leq E_{\sigma,\varphi}^{-1}\Big(E_{\sigma,\varphi}\big(f(x_0)-f^\star\big)+k\Big), \tag{17}$$

where $E_{\sigma,\varphi}(s) := \int_s^\infty \frac{1}{\sigma^*(1/\varphi'(r))}dr$.

Proof. The proof follows the same general steps as that of Theorem 2, with only minor adjustments needed on the differential inequality to account for the more general smoothness function σ. ■

V. Conclusion and Future Work

We analyzed the convergence of our proposed σ-RGD algorithm under σ-smoothness and ϕ-convexity or φ-KŁ. We saw that, under a suitable relationship between σ and ϕ, our algorithm converges at a linear rate. Further, under L-smoothness and the Euclidean norm, our algorithm reduces to vanilla gradient descent (GD), which allowed us to show that the transition from sublinear to linear rate of GD under convexity and strong convexity is not actually abrupt but instead smoothly depends on ϕ. Further, our analysis provides insights into how to optimization landscape affects the complexity analysis for gradient-based optimization. Lastly, we provide a general rate under σ-smoothness and the Kurdyka-Lojasiewicz inequality.

For future work, this general result will be further refined and studied in the context of deep learning, as motivated by works such as [17], [18]. Indeed, we believe that our framework can be used to further explain why gradient clipping can accelerate gradient descent. To some extent, our σ-RGD can be seen as a smooth form of gradient clipping, particularly for functions that locally behave quadratic, but grow fast outside the vicinity of local minima (e.g. $\sigma(s) = \Theta(s^2)$ for $s \to 0$ and $\sigma(s) = \Theta(s^{2p})$ with $p \gg 1$, as is the case for sum of squared polynomials of order 2 through p). For future work, we will also investigate the role of smoothness and convexity on accelerated methods, hopefully extending some of the results and ideas explored in [19]. We will also revisit backtracking from the lens of uniform smoothness and convexity. Lastly, we will extend some of our results to the online and stochastic optimization setting, and we would like to establish lower bounds.

References

[1] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[2] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

[3] D. Bertsimas and J. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific, 1997.

[4] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali, *Linear Programming and Network Flows*. Wiley-Interscience, 2004.

[5] D. Bertsekas, *Convex Optimization Theory*, ser. Athena Scientific optimization and computation series. Athena Scientific, 2009. [Online]. Available: https://books.google.com/books?id=lC1EEAAAQBAJ

[6] C. Zalinescu, *Convex Analysis In General Vector Spaces*. World Scientific Publishing Company, 2002. [Online]. Available: https://books.google.com/books?id=jcbUCgAAQBAJ

[7] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer Publishing Company, Incorporated, 2018.

[8] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Machine Learning and Knowledge Discovery in Databases*, P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, Eds. Cham: Springer International Publishing, 2016, pp. 795–811.

[9] A. C. Wilson, L. Mackey, and A. Wibisono, *Accelerating rescaled gradient descent: fast optimization of smooth functions*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[10] C. Zălinescu, "On uniformly convex functions," *Journal of Mathematical Analysis and Applications*, vol. 95, no. 2, pp. 344–374, 1983.

[11] D. Azé and J.-P. Penot, "Uniformly convex and uniformly smooth convex functions," in *Annales de la Faculté des sciences de Toulouse : Mathématiques*, vol. 6(4), 1995.

[12] T. Kerdreux, A. d'Aspremont, and S. Pokutta, "Local and global uniform convexity conditions," *arXiv preprint arXiv:2102.05134*, 02 2021.

[13] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, "Characterizations of Lojasiewicz inequalities: Subgradient flows, talweg, convexity," *Transactions of the American Mathematical Society*, vol. 362, no. 6, pp. 3319–3363, Jun. 2010. [Online]. Available: https://hal.science/hal-00243094

[14] K. Scaman, C. Malherbe, and L. D. Santos, "Convergence rates of non-convex stochastic gradient descent under a generic lojasiewicz condition and local smoothness," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 19 310–19 327. [Online]. Available: https://proceedings.mlr.press/v162/scaman22a.html

[15] I. Fatkhullin, J. Etesami, N. He, and N. Kiyavash, "Sharp analysis of stochastic optimization under global kurdyka-lojasiewicz inequality," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 15 836–15 848. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/65ae674df2fb642518ae8d2b5435e1b8-Paper-Conference.pdf

[16] H. K. Khalil, *Nonlinear systems; 3rd ed.* Upper Saddle River, NJ: Prentice-Hall, 2002, the book can be consulted by contacting: PH-AID: Wallet, Lionel. [Online]. Available: https://cds.cern.ch/record/1173048

[17] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BJgnXpVYwS

[18] H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie, "Convex and non-convex optimization under generalized smoothness," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=8aunGrXdkl

[19] A. C. Wilson, L. Mackey, and A. Wibisono, "Accelerating rescaled gradient descent: Fast optimization of smooth functions," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.