Imaginary Zeroth-Order Convex Optimization: linear convergence rates

Helia Atarod, Wouter Jongeneel and Daniel Kuhn

September 29, 2022

Abstract

In this article we exploit the properties of nonlinear convex real-analytic functions to sharpen a sublinear convergence rate to a linear convergence rate. Numerical experiments corroborate this theorem.

Keywords—zeroth-order optimization, derivative-free optimization, complex-step derivative, gradient estimation, Łojasiewicz inequality, numerical optimization.

AMS Subject Classification (2020)—65D25, 65G50, 65K05, 65Y04, 65Y20, 90C56.

1 Introduction

In this article we study optimization problems of the form

$$\underset{x\in\mathcal{X}}{\text{minimize}} \quad f(x), \tag{1.1}$$

where $f:\mathcal{D}\to\mathbb{R}$ is a smooth objective function defined on an open set $\mathcal{D}\subseteq\mathbb{R}^n$, and $\mathcal{X}\subseteq\mathcal{D}$ is a non-empty closed feasible set. The set of minimizers is denoted by $\mathcal{X}^\star$. In [JYK21; Jon21] the authors exploit real-analyticity of the objective to derive an imaginary *zeroth-order optimization* framework that is particularly well applicable to convex optimization problems. However, the fact that the objective is real-analytic can be further exploited in the convergence analysis. In this work we will show that convex optimization with real-analytic objective functions becomes effectively optimization with *strongly* convex objective functions.

Consider the metric space $(\mathbb{R}^n,\|\cdot\|_2)$ and let $\mathcal{U}\subseteq\mathbb{R}^n$ be open and non-empty. Consider some real-analytic function $f\in C^\omega(\mathcal{U})$ with $p^\star\in\mathcal{U}$ being a ***critical point*** of f, that is, $\nabla f(p^\star)=0$. Then the ***Łojasiewicz inequality*** says that there is a rational constant $\theta\in[\frac{1}{2},1)$, a constant $C\geq 0$ and a set $\mathcal{W}\subseteq U$ such that

$$|f(x)-f(p^\star)|^\theta \leq C\|\nabla f(x)\|_2 \quad \forall x\in\mathcal{W}. \tag{1.2}$$

Example 1.1 (The scalar case of (1.2))**.** ...

An instance of the Łojasiewicz inequality, independently due to Polyak [Pol63], is often exploited in optimization, that is, for τ-strongly convex functions one can show that the ***Polyak-Łojasiewicz inequality*** (PL inequality) $\|\nabla f(x)\|_2^2\geq 2\lambda(f(x)-f(p))$ holds with $\lambda=\tau$ [Nes03, Equation 2.1.19].

We will use this type of inequalities to improve upon the convergence rate for convex real-analytic functions as given in [JYK21, Theorem 4.1]. In particular, we use (1.2) and the proof of [JYK21, Theorem 5.1].

If $f : \mathcal{X} \to \mathbb{R}$ is convex and satisfies the PL inequality for some λ, then f satisfies the ***quadratic growth*** (QG) condition

$$f(x) - f(x^\star) \geq \frac{\lambda}{2}\|x - x^\star\|_2^2 \tag{1.3}$$

for all $x \in \mathcal{X}$, which is weaker than strong convexity [KNS16, Theorem 2]. The QG condition in combination with convexity goes by the name of "*optimal strong convexity*" [LW15]. In particular, if $\mathcal{X}^\star$ is not merely a singleton, (1.3) becomes

$$f(x) - f(P_{\mathcal{X}^\star}(x)) \geq \frac{\lambda}{2}\|x - P_{\mathcal{X}^\star}(x)\|_2^2, \tag{1.4}$$

for $P_{\mathcal{X}^\star}(\cdot)$ the projection operator onto $\mathcal{X}^\star$, which is sometimes written simply as x_p.

Exactly the condition (1.3) is used in the proof of [JYK21, Theorem 5.1]. As such, analyzing convex optimization with $f \in C^\omega$ should be akin to strongly convex optimization.

Example 1.2 (Convex real-analytic)**.** *The following functions are convex and real-analytic, but not strongly convex.*

(i) $f : \mathbb{R}^2 \to \mathbb{R}$ *defined by* $f(x) = (x_1 + x_2)^2$.

(ii) $f : \mathbb{R}^n \to \mathbb{R}$ *defined by* $f(x) = 0\ \forall x \in \mathbb{R}^n$.

(iii) $f : \mathbb{R}^n \to \mathbb{R}$ *defined by* $f(x) = \|Ax - b\|_2^2$ *with* $\ker(A) \neq \{0\}$.

(iv) ...

(v) ...

[We can have more example functions]

1.1 Related work Zeroth-order optimization is particularly suitable for simulation-based and data-driven optimal control problem *cf.* [Faz+18]. ...

1.2 Contributions By analyzing the set of nonlinear convex real-analytic functions we are able to sharpen the sublinear rate of the form $O(K^{-1})$, as proven in [JYK21, Theorem 4.1], to a linear rate of the form $O(\alpha^K)$ for some $\alpha \in (0, 1)$.

2 Notions of regularity

A function is said to be C^k-smooth when it is k times continuously differentiable. We highlight a stronger regularity notion of great importance in this article.

Definition 2.1 (Real analytic function)**.** *The function* $f : \mathcal{D} \to \mathbb{R}$ *is real analytic on* $\mathcal{D} \subseteq \mathbb{R}^n$ *if for every* $x' \in \mathcal{D}$ *there exist* $f_\alpha \in \mathbb{R}$, $\alpha \in \mathbb{Z}_+^n$, *and an open set* $U \subseteq \mathcal{D}$ *containing* x' *such that*

$$f(x) = \textstyle\sum_{\alpha \in \mathbb{Z}_+^n} f_\alpha \cdot (x - x')^\alpha \quad \forall x \in U. \tag{2.1}$$

We use $C^\omega(\mathcal{D})$ *to denote the family of all real analytic functions on* $\mathcal{D}$.

Indeed, the power series representation (2.1) corresponds to the Taylor series of f around x'.

Using the notation from [Nes03] a function f is said to be $C_L^{k,r}(\mathcal{D})$-***smooth*** when f is k times continuously differentiable with additionally having its r^{th}-derivative being L-Lipschitz over some open set $\mathcal{D} \subseteq \mathbb{R}^n$. Here, k is an element of $\mathbb{N}_{\geq 0} \cup \{\infty\} \cup \{\omega\}$. That is, if $f \in C^{1,1}_{L_1(f)}(\mathcal{D})$, then, f has a ***Lipschitz gradient***, *i.e.*,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_1(f)\|x - y\|_2, \quad \forall x, y \in \mathcal{D}. \tag{2.2}$$

which is equivalent [NS17, Equation (6)] to the inequality

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \tfrac{1}{2} L_1 \|x - y\|_2^2 \quad \forall x, y \in \mathcal{D}. \tag{2.3}$$

Similarly, if $f \in C^{2,2}_{L_2(f)}(\mathcal{D})$, then, f has a ***Lipschitz Hessian***, *i.e.*,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L_2(f)\|x - y\|_2 \quad \forall x, y \in \mathcal{D}. \tag{2.4}$$

Then, Consider the setting of $f \in C^{\omega}(\mathcal{D})$ being $\tau(f)$-***strongly convex*** over $\mathcal{D}$, *i.e.*, there is some $\tau(f) > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\tau(f)\|y - x\|_2^2, \quad \forall x, y \in \mathcal{D}. \tag{2.5}$$

In particular (2.5) implies that for $\mathcal{D}$ such that $x^\star \in \mathrm{int}(\mathcal{D})$ one has

$$f(x) - f(x^\star) \geq \tfrac{1}{2}\tau(f)\|x - x^\star\|_2^2, \quad \forall x \in \mathcal{D}. \tag{2.6}$$

If additionally $f \in C^{\omega,1}_{L_1(f)}$, then by the PL-condition $\|\nabla f(x)\|_2^2 \geq 2\tau(f)(f(x) - f(x^\star))$ [Nes03, Equation 2.1.19] one has

$$\tau(f)\|x - x^\star\|_2 \leq \|\nabla f(x)\|_2 \leq L_1(f)\|x - x^\star\|_2. \tag{2.7}$$

3 Zeroth-order algorithm

First we recall the imaginary zeroth-order optimization framework from [JYK21; Jon21].

3.1 Imaginary gradient estimation To make sure that the gradient estimator is well-defined we assume the following.

Assumption 3.1 (Analytic extension)**.** *The objective function $f : \mathcal{D} \to \mathbb{R}$ of problem* (1.1) *admits an analytic extension to the strip $\mathcal{D} \times i \cdot (-\bar{\delta}, \bar{\delta})^n$ for some $\bar{\delta} \in (0, 1)$.*

Now the gradient estimator is constructed via a surrogate function f_δ of f, which is defined as

$$f_\delta(x) = V_n^{-1} \textstyle\int_{\mathbb{B}^n} \Re\big(f(x + i\delta y)\big)\mathrm{d}y. \tag{3.1}$$

Here, $\delta \in (0, \bar{\delta})$ is the radius of the ball we smooth over. It turns out that the gradient of f_δ has a representation particularly suitable for a zeroth-order optimization framework.

Proposition 3.2 (Gradient of the smoothed complex-step function [JYK21, Proposition 3.3])**.** *If $f \in C^{\omega}(\mathcal{D})$ satisfies Assumption 3.1, then f_δ defined as in* (3.1) *is differentiable, and we have*

$$\nabla f_\delta(x) = \frac{n}{\delta}\mathbb{E}_{y\sim\sigma}\left[\Im\left(f(x + i\delta y)\right) y\right] \quad \forall x \in \mathcal{D},\ \delta \in (0, \bar{\delta}), \tag{3.2}$$

where σ denotes the uniform distribution on $\mathbb{S}^{n-1}$.

Differently put, by Proposition 3.2 we find that the gradient of f_δ admits the unbiased single-point estimator

$$g_\delta(x) = \frac{n}{\delta}\Im\left(f(x + i\delta y)\right) y \quad \text{with} \quad y \sim \sigma. \tag{3.3}$$

This estimator has been analyzed in [JYK21; Jon21]. We will analyze (3.3) in an algorithm akin to gradient descent.

Algorithm 1 Imaginary zeroth-order optimization

1: **Input:** initial iterate $x_1 \in \mathcal{X}$, stepsizes $\{\mu_k\}_{k\in\mathbb{N}}$, smoothing parameters $\{\delta_k\}_{k\in\mathbb{N}}$
2: **for** $k = 1, 2, \dots, K-1$ **do**
3: sample $y_k \sim \sigma$
4: set $g_{\delta_k}(x_k) = \frac{n}{\delta_k}\Im\left(f(x_k + i\delta_k y_k)\right) y_k$
5: set $x_{k+1} = \Pi_{\mathcal{X}}\left(x_k - \mu_k\, g_{\delta_k}(x_k)\right)$
6: **end for**
7: **Output:** last iterate x_K

3.2 Algorithm and convergence proof In the remainder we will assume that the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 as well as all directional samples $\{y_k\}_{k\in\mathbb{N}}$ and the corresponding gradient estimators $\{g_{\delta_k}(x_k)\}_{k\in\mathbb{N}}$ represent random objects on an abstract filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_{k\in\mathbb{N}}, \mathbb{P})$, where $\mathcal{F}_k$ denotes the σ-algebra generated by the independent and identically distributed (i.i.d.) samples $y_1, \dots, y_{k-1}$. Hence, x_k is $\mathcal{F}_k$-measurable. We let $\mathbb{E}[\cdot]$ denote the expectation operator with respect to $\mathbb{P}$.

Now we continue with formalizing what was alluded to the introduction. To that end, we assume the following.

Assumption 3.3 (Nonlinearity)**.** *The objective function $f \in C^{\omega}(\mathcal{D})$ of problem (1.1) is such that for all $x^\star \in \mathcal{X}$ and any $y \in \mathbb{S}^{n-1}$ the function $\partial_t^2 f(x^\star + ty)$ is not identically 0.*

Note that Assumption 3.3 admits a variety of formulations. This assumption effectively rules out affine functions.

In what follows, let μ_n denote the Lebesgue measure on $\mathbb{R}^n$.

Lemma 3.4. *Suppose that $f \in C^{\omega}(\mathcal{D})$ is convex and satisfies Assumption 3.3. Then, there is an open neighbourhood $\mathcal{W} \subseteq \mathcal{D}$ of $\mathcal{X}^\star$ such that for μ_n-a.e. $x \in \mathcal{W}$ (1.2) holds with $\theta = \frac{1}{2}$.*

Proof (sketch). As $f \in C^{\omega}$, f satisfies (1.2) for *some* θ. Moreover, as $f \in C^{\omega}$ we know that for all $x \in \mathcal{X}^\star$ the function $\partial_t^2 f(x^\star + ty)$, for any $y \in \mathbb{S}^{n-1}$, is either identically zero, or μ_n-a.s. non-zero, the former being impossible by assumption. This however means that, at least locally, (1.2) must hold with $\theta = \frac{1}{2}$ for almost every x in some neighbourhood of $x^\star$. □

An important ramification of Lemma 3.4 is that under those conditions the quadratic growth condition (1.3) holds for μ_n-a.e. $x \in \mathcal{W}$. This follows directly from the proofs in [KNS16]. Note that even for f being convex, one cannot always extend the domain of (1.2) from $\mathcal{W}$ to $\mathcal{D}$ *cf.* (4.1).

Now we have the machinery to generalize the rate from [JYK21, Theorem 5.1] to merely (non-linear) convex functions.

Theorem 3.5 (Convergence rate of Algorithm 1 for convex optimization)**.** *Suppose that $f \in C^{\omega}(\mathcal{D})$ is a convex function satisfying Assumption 3.1 and Assumption 3.3 as well as the Lipschitz conditions (2.2) and (2.4) with $L_1 > 0$ and $L_2 \geq 0$. Also assume that $\mathcal{X}$ has non-empty interior, is closed and convex, that $\nabla f(x^\star) = 0\ \forall x^\star \in \mathcal{X}^\star$ and that $\mathcal{X} \subseteq \mathcal{W}$, for $\mathcal{W}$ as in Lemma 3.4. Denote by $\{x_k\}_{k\in\mathbb{N}}$ the iterates generated by Algorithm 1 with constant stepsize $\mu_k = \mu = 1/(2nL_1)$ and adaptive smoothing parameter $\delta_k \in (0, \kappa\bar{\delta}]$ for all $k \in \mathbb{N}$, where $\kappa \in (0,1)$, and define $R = \|x_1 - x^\star\|_2$. If $\delta_k = \delta/k$ for all $k \in \mathbb{N}$, then, there is a constant $C \geq 0$ and a $\lambda \in (0, L_1]$ such that the following inequality holds for all $K \in \mathbb{N}$*

$$\mathbb{E}[f(x_K) - f(x^\star)] \leq \tfrac{1}{2}L_1\left(\delta^2 C + \left(1 - \tfrac{\lambda}{4nL_1}\right)^{K-1}\left(R^2 - \delta^2 C\right)\right). \tag{3.4}$$

A semi-explicit formula for C in terms of n, L_1, L_2 and τ is derived in the proof of Theorem 3.5.

Proof. As illustrated in the introduction, we can proceed as in the proof of [JYK21, Theorem 5.1]. To start, as in the proof of [JYK21, Theorem 4.1], we set $C_1 = 3(\frac{1}{6}L_2 + C_\kappa)$ and $r_k = \|x_k - x^\star\|_2$ for all $k \in \mathbb{N}$, and we initially assume that $\mathcal{X} = \mathcal{D}$. Now, as $x_k \in \mathcal{W}$, then, combining [JYK21, Equation (4.1)] from the proof of [JYK21, Theorem 4.1], that is,

$$\mathbb{E}\left[r_{k+1}^2 \,\middle|\, \mathcal{F}_k\right] \leq r_k^2 - \mu\left(f(x_k) - f(x^\star)\right) + n\mu\delta_k^2 C_1 r_k + \mu^2 n^2 C_1^2 \delta_k^4. \tag{3.5}$$

with the QG condition (1.3), which we can do by Lemma 3.4 for μ-a.e. $x_k \in \mathcal{W}$, yields

$$\mathbb{E}\left[r_{k+1}^2 | \mathcal{F}_k\right] \leq \left(1 - \tfrac{\mu\lambda}{2}\right) r_k^2 + \mu C_1 n \delta_k^2 r_k + \mu^2 C_1^2 n^2 \delta_k^4,$$

for some $\lambda > 0$. By taking unconditional expectations, and applying Jensen's inequality, we then find

$$\mathbb{E}[r_{k+1}^2] \leq \left(1 - \tfrac{\mu\lambda}{2}\right) \mathbb{E}[r_k^2] + \mu C_1 n \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \delta_k^4 \tag{3.6a}$$

$$\leq \mathbb{E}[r_k^2] + \mu C_1 n \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \delta_k^4. \tag{3.6b}$$

Note, the latter inequality holds regardless of $x_k \in \mathcal{W}$. Next, choose any $k' \in \mathbb{N}$ and sum the above inequalities over all $k \leq k' - 1$ to obtain

$$\begin{aligned}\mathbb{E}[r_{k'}^2] &\leq r_1^2 + \mu C_1 n \textstyle\sum_{k=1}^{k'-1} \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'-1} \delta_k^4 \\ &\leq r_1^2 + \mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4.\end{aligned}$$

By using the same reasoning as in the proof of [JYK21, Theorem 4.1], that is, by exploiting [SRB11, Lemma 1], the last bound implies

$$\sqrt{\mathbb{E}\left[r_{k'}^2\right]} \leq \mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 + r_1 + \left(\mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4\right)^{\frac{1}{2}}.$$

Substituting this inequality into (3.6a) for $k = k'$ and noting that $r_1 = R$ yields

$$\mathbb{E}[r_{k'+1}^2] \leq \left(1 - \tfrac{\mu\lambda}{2}\right) \mathbb{E}[r_{k'}^2] + \mu^2 C_1^2 n^2 \delta_{k'}^4 + \mu C_1 n \delta_{k'}^2 \left(\mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 + R + \left(\mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4\right)^{\frac{1}{2}}\right).$$

Indeed, if $x_k \notin \mathcal{W}$, we can replace $(1 - \mu\lambda/2)$ by 1. Then, as $\delta_k = \delta/k$ for all $k \in \mathbb{N}$ and as the constant stepsize satisfies $\mu = 1/(2nL_1)$, we may then use the standard zeta function inequalities, that is,

$$\textstyle\sum_{j=1}^J j^{-2} \leq \zeta(2) = \frac{1}{6}\pi^2 \quad \text{and} \quad \sum_{j=1}^J j^{-4} \leq \zeta(4) = \frac{1}{90}\pi^4 \quad \forall J \in \mathbb{N}, \tag{3.7}$$

to obtain

$$\mathbb{E}[r_{k'+1}^2] \leq \left(1 - \tfrac{\lambda}{4nL_1}\right) \mathbb{E}[r_{k'}^2] + C_1^2 \tfrac{\delta^4}{4L_1^2(k')^4} + C_1^2 \tfrac{\pi^2\delta^4}{24L_1^2(k')^2} + C_1 R \tfrac{\delta^2}{2L_1(k')^2} + C_1^2 \tfrac{\pi^2\delta^4}{4\sqrt{90}L_1^2(k')^2} \tag{3.8}$$

$$\leq \left(1 - \tfrac{\lambda}{4nL_1}\right) \mathbb{E}[r_{k'}^2] + C_1 R \tfrac{\delta^2}{L_1} + 3C_1^2 \tfrac{\delta^4}{L_1^2}, \tag{3.9}$$

where the last inequality follows from the elementary bounds $\frac{1}{2(k')^2} < 1$, $\frac{1}{4(k')^4} < 1$, $\frac{\pi^2}{24(k')^2} < 1$ and $\pi^2/(4\sqrt{90}(k')^2) < 1$. As $|\delta| < 1$, we may set $C = \frac{4n}{\lambda}(C_1 R + 3C_1^2/L_1)$ to obtain

$$\mathbb{E}[r_{k'+1}^2] \leq \left(1 - \tfrac{\lambda}{4nL_1}\right) \mathbb{E}[r_{k'}^2] + \tfrac{\lambda}{4nL_1}\delta^2 C.$$

Taken together, the Lipschitz inequality (2.2) and the quadratic growth condition (1.3) imply that $\lambda \leq L_1$, that is, one recovers (2.7) with λ taking the role of τ, which in turn ensures that $\lambda/(4nL_1) < 1$. Hence, the above inequality implies

$$\left(\mathbb{E}[r_{k'+1}^2] - \delta^2 C\right) \leq \left(1 - \tfrac{\lambda}{4nL_1}\right) \left(\mathbb{E}([r_{k'}^2] - \delta^2 C\right).$$

Then it follows that
$$\left(\mathbb{E}[r_K^2]-\delta^2 C\right)\leq\left(1-\tfrac{\lambda}{4nL_1}\right)^{K-1}\left(R-\delta^2 C\right).$$
The final claim follows by combining this inequality with the estimate $\mathbb{E}[f(x_K)-f(x^\star)]\leq\frac{1}{2}L_1\mathbb{E}[r_K^2]$, which follows from the Lipschitz condition (2.3). This completes the proof for $\mathcal{X}=\mathcal{D}$. To show that the claim remains valid when $\mathcal{X}$ is a non-empty closed convex subset of $\mathcal{D}$, we may proceed as in the proof of [JYK21, Theorem 4.1]. Details are again omitted for brevity. □

The convergence rate as proven in [JYK21] for τ-*strongly* convex functions is as follows
$$\mathbb{E}[f(x_K)-f(x^\star)]\leq\tfrac{1}{2}L_1\left(\delta^2 C+\left(1-\tfrac{\tau}{4nL_1}\right)^{K-1}\left(R^2-\delta^2 C\right)\right),\tag{3.10}$$
which is qualitatively the rate we found above *cf.* (3.4), yet λ took the role of τ. Recall that we know by [JYK21, Theorem 4.1] that under the conditions of Theorem 3.5, there is a constant $C_2\geq 0$ such that
$$\mathbb{E}\left[f(\bar{x}_K)-f(x^\star)\right]\leq\tfrac{n}{K}\left(\sqrt{2L_1}R+C_2\delta^2\right)^2,$$
for the averaged iterate $\bar{x}_K=\frac{1}{K}\sum_{k=1}^K x_k$. As such we sharpened their rate from sublinear to linear.

Let us clarify when a function *fails* to meet the conditions of Theorem 3.5. For instance, consider $f(x)=x^4$, this convex real-analytic function fails to satisfy the PL condition around $x^\star=0$. Indeed, $\partial_x^2 f(x)|_{x=x^\star}=0$.

Note that by our assumption $\mathcal{X}\subseteq\mathcal{W}$, Theorem 3.5 can be understood as a *local* or *asymptotic* result. We come back to this remark in the numerical section.

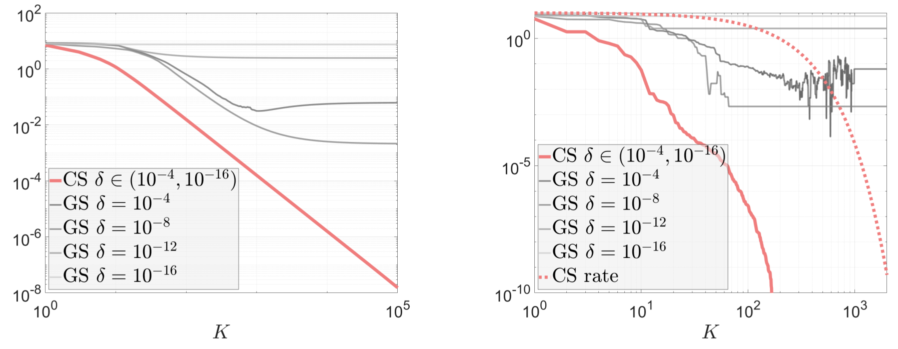
4 Numerical experiments

In this section we showcase our convergence rate.

Example 4.1 (Smooth approximate ℓ_1-regularization)**.** *Following [FG16], we are interested in solving a smoothly approximated version of a ℓ_1-regularized convex program. Specifically, we consider the pseudo-Huber loss given by*
$$\psi_\theta(x)=\theta\sum_{i=1}^m\left(\sqrt{1+x_i^2/\theta^2}-1\right)\tag{4.1}$$
and we are interested in minimizing the objective $f(x)=\frac{1}{2}\|Ax-b\|_2^2+\lambda\psi_\theta(x)$ over $x\in\mathbb{R}^n$ for some $\lambda>0$ and data $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$. It follows from [FG16, Lemma 2] that $L_1=\lambda/\theta+\|A^\top A\|_2$, whereas it follows from [FG16, Lemma 6] that $L_2=\lambda/\theta^2$. Now, we compare Algorithm 1 to the method proposed in [NS17]. Here we let A and b be random with unit covariance matrices for $m=4$, $n=2$. Moreover, $\lambda=\theta=10^{-4}$ and $x_1=(0,0)$. We show the costs $f(\bar{x}_K)$ and $f(x_k)$ in Figure 4.1i and Figure 4.1ii, respectively, for a decreasing smoothing parameter δ. Again, as in [JYK21] a difference in numerical stability can be observed. More importantly, for x_K we observe a convergence rate that qualitatively matches Theorem 3.5 indeed.

At last, we consider a degenerate quadratic function, that is, a convex function that is not strongly convex.

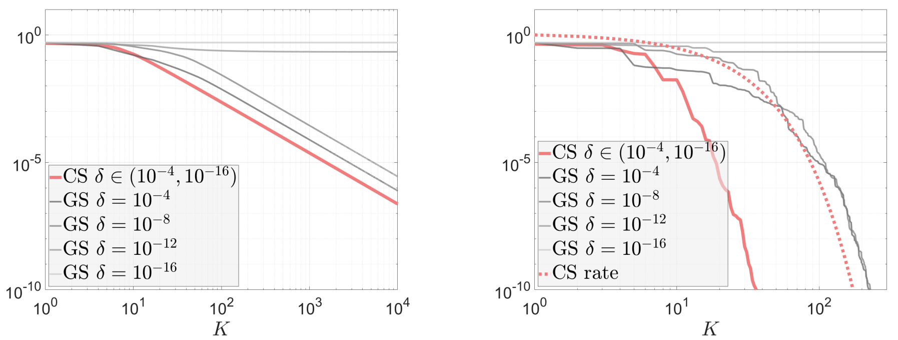
Example 4.2 (Degenerate quadratic function)**.** *We redo Example 4.1, but for a different objective and with $x_1=(1,1)$. Let $f\in C^\omega(\mathbb{R})$ be defined by $f:(x_1,x_2)\mapsto\frac{1}{2}x_1^2$. This function has $L_1=1$. The results are shown in Figure 4.2, again, we observe the numerical stability of the complex-step method and additionally, the convergence rate from Theorem 3.5.*



(i) Suboptimality gap $f(\bar{x}_K) - f(x^\star)$ for Example 4.1.

(ii) Suboptimality gap $f(x_K) - f(x^\star)$ for Example 4.1.

Figure 4.1: The single-point complex smoothing (CS) method (Algorithm 1.(a)) compared to the multi-point Gaussian smoothing (GS) method from [NS17, Equation (54)] on a variety of objective functions for a time-varying smoothing parameter $\delta_k = \delta/k$. For the "*CS rate*" we plot the sequence $z_K = \frac{1}{2}L_1(1 - \frac{1}{4nL_1})^{K-1}R^2$.



(i) Suboptimality gap $f(\bar{x}_K) - f(x^\star)$ for Example 4.2.

(ii) Suboptimality gap $f(x_K) - f(x^\star)$ for Example 4.2.

Figure 4.2: The single-point complex smoothing (CS) method (Algorithm 1.(a)) compared to the multi-point Gaussian smoothing (GS) method from [NS17, Equation (54)] on a variety of objective functions for a time-varying smoothing parameter $\delta_k = \delta/k$. For the "*CS rate*" we plot the sequence $z_K = \frac{1}{2}L_1(1 - \frac{1}{4nL_1})^{K-1}R^2$.

Bibliography

[Faz+18] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. "Global convergence of policy gradient methods for the Linear Quadratic regulator". *International Conference on Machine Learning*. 2018, pp. 1467–1476.

[FG16] K. Fountoulakis and J. Gondzio. "A second-order method for strongly convex ℓ_1-regularization problems". *Mathematical Programming* 156.1 (2016), pp. 189–219.

[Jon21] W. Jongeneel. "Imaginary Zeroth-Order Optimization" (2021). arXiv: `2112.07488`.

[JYK21] W. Jongeneel, M.-C. Yue, and D. Kuhn. "Small errors in random zeroth-order optimization are imaginary" (2021). arXiv: `2103.05478`.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition". *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, pp. 795–811.

[LW15] J. Liu and S. J. Wright. "Asynchronous stochastic coordinate descent: Parallelism and convergence properties". *SIAM Journal on Optimization* 25.1 (2015), pp. 351–376.

[Nes03] Y. Nesterov. *Introductory Lectures on Convex Optimization: a Basic Course*. Springer Science & Business Media, 2003.

[NS17] Y. Nesterov and V. Spokoiny. "Random gradient-free minimization of convex functions". *Foundations of Computational Mathematics* 17.2 (2017), pp. 527–566.

[Pol63] B. Polyak. "Gradient methods for the minimisation of functionals". *Ussr Computational Mathematics and Mathematical Physics* 3 (1963), pp. 864–878.

[SRB11] M. Schmidt, N. Roux, and F. Bach. "Convergence rates of inexact proximal-gradient methods for convex optimization". *Neural Information Processing Systems* 24 (2011), pp. 1458–1466.

A To do

A.0.0.1 Primary goal Use (1.2), to "*reconstruct*" the proof of [JYK21, Theorem 5.1], but *without* assuming strong convexity, we can assume that $f \in C^\omega$ is convex and has a Lipschitz gradient. If needed, we can also assume that f has a Lipschitz Hessian. [I made a start, but technical details require more care.]

[Code in folder.]

[At last, we need to understand how a trajectory $\{x_k\}_{k\in\mathbb{N}}$ under Algorithm 1 behaves. In particular, how does $|\{x_k \in \mathcal{W}\}|$ grow with $k \to +\infty$? As $y_k \sim \sigma$ and $f \in C^\omega$, x_k will visit every open set of $\mathcal{D}$ with strictly positive probability. The only fixed point of $x_{k+1} = x_k - \mu_k g_{\delta_k}(x_k)$ for $\delta_k \to 0$ is $x_k = x^\star \in \mathcal{W}$].

A.0.0.2 Secondary goal If there is time left, can we say anything about a non-convex case?

A.0.0.3 Questions

(i) Warming up: find a convex real-analytic function that is not strongly convex? Can you do it for $n = 1$? [Add to Example 1.2]

(ii) Use a power series (Taylor series) argument to show (1.2) for $n = 1$. This should reveal why $\theta \geq \frac{1}{2}$. [Add to Example 1.1]

(iii) Develop an understanding, by means of an example of the "*how local*" (1.2) is.

(iv) Can we say anything about θ, C and/or W? If not, this means we only capture the convergence rate *regime*, not the actual rate (as C is most likely unknown). [This seems to be true only regarding C, θ can be quantified, what about W?]

(v) Show that (4.1) is not strongly convex.

(vi) Prove Lemma 3.4 (if true!). [Can we get it to be global?]

(vii) Can we do a more interesting example?

Further resources For a short explanation of how the PL condition can be exploited, see[1]. The work by Łojasiewicz[2], in French. Link to the English version of [Pol63][3]. The arXiv version of [KNS16][4].

$\mathscr{P} \equiv (\Sigma, \mathscr{O}, \mathscr{T})$

[1] https://labs.utdallas.edu/conlab/linear-convergence-of-gradient-and-proximal-gradient-methods-under-the-polyak-lojasiewicz-condition/

[2] https://perso.univ-rennes1.fr/michel.coste/Lojasiewicz.pdf

[3] https://www.sciencedirect.com/science/article/pii/0041555363903823?

[4] https://arxiv.org/pdf/1608.04636v3.pdf