ECE 285 Winter 2024:1–14, 2024

Performance Estimation for Weakly Convex Functions

Pranav Reddy P3REDDY@UCSD.EDU
UC San Diego

Abstract

We propose applying the techinques of performance estimation, first developed by Yoel Drori and Marc Teboulle, to the class of first-order methods of weakly-convex functions. The performance estimation problem aims to reduce the analysis of first-order methods to solving a nonconvex quadratically constrained quadratic program. With the relaxation and interpolation conditions proposed by Adrien Taylor, we can almost fully reduce the problem of finding an optimal first-order algorithm for convex functions to solving a semidefinite program. In this project we propose developing a new set of interpolation conditions for the class of ρ-weakly convex functions and apply the techinques of performance estimation to develop improved first-order algorithms.

Keywords: Performance estimation problem, first-order methods, nonsmooth optimization, nonconvex optimization, weakly convex, semidefinite programming.

1. Introduction

Previous conditions for weakly convex, convex, convex with bounded subgradients, smooth convex, and strongly convex functions have been proven in Taylor et al. (2017a). Existing work, however, does not, to our knowledge, include the class of ρ-weakly convex functions, which are of importance in machine learning and signal processing (Davis et al. (2018)).

1.1. Related Works

In 2014, Drori and Teboulle (2012) proposed the performance estimation problem (PEP) as a computer-assisted tool for analysis of gradient methods on smooth convex functions. By relaxing the infinite-dimensional problem to a finite-dimensional semidefinite program, they were able to improve upon existing methods in an automated manner. The essential insight that analysis of optimization methods is equivalent to finding the worst-case performance of a method on a class of functions was greatly extended by Taylor et al. (2016, 2017a) who introduced a set of interpolation conditions to make the relaxation into an exact reformulation. Although numerically challenging, Gupta et al. (2023) shows that there exist effective algorithms for solving such problems, and Taylor et al. (2017a) shows that analytic feasible solutions can be a powerful tool in proving convergence rates. Other works by Grimmer (2023); Shahriari-Mehr and Panahi (2024); Colla and Hendrickx (2023) have shown applications in smooth optimization and distributed problems. Hu et al. (2023) uses the PEP techniques on manifold optimization problems to prove a $O\left(\frac{1}{\sqrt{n}}\right)$ convergence rate of extragradient methods. The PEP has also seen applications for inexact methods, as shown by Kornilov et al. (2023).

1.2. Problem Formulation

The field of optimization revolves around designing algorithms and proving their convergence properties. For example, the classical proof of the $O\left(\frac{1}{n}\right)$ convergence rate of gradient descent for an L-smooth convex function assumes certain step-sizes of the algorithm acting on a set of functions. This is proven via a sequence of linear combinations of inequalities, eventually giving a bound on the decrease over n iterations. In the other direction, Nesterov (2014) shows that any method where future iterates are in the linear span of prior iterates and (sub)gradients cannot achieve better than a $O\left(\frac{1}{n^2}\right)$ convergence rate. In this manner, we can see that past works have bounded the worst-case performance and shown that these bounds are tight. Informally, this can be viewed as the primal and dual formulations of an optimization problem, where the decision variable is not an element of $\mathbb{R}^n$ but a function in a given function class, such as $\mathcal{F}_{\mu,L}$, the set of L-smooth μ-strongly convex functions. Formally, given a method $\mathcal{M}$, and performance measure $\mathcal{E}$, a class of functions $\mathcal{F}$, and some intial conditions $\mathcal{C}$, we are interested in the worst performance of $\mathcal{M}$:

$$\begin{aligned} \sup_{f} \quad & \mathcal{E}\left(\{x_i, f_i, g_i\}_{i=0,1,\ldots,N,\star}\right) \\ \text{subject to} \quad & f \in \mathcal{F}, \\ & x_\star \text{ is optimal}, \\ & \{(x_i, f_i, g_i)\}_{i=1,\ldots,N} \text{ are generated by } \mathcal{M}, \\ & (x_0, f_0, g_0) \text{ satisfy the initial conditions } \mathcal{C}. \end{aligned} \tag{f-PEP}$$

Taylor et al. (2016) showed that given a finite set of linear-qudratic interpolation conditions, the infinite-dimensional constraint $f \in \mathcal{F}$ can be reduced to $N+2$ variables in $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$ – a total of $(N+2)(2d+1)$ variables. This makes the primal-dual perspective formal: feasible points of the primal SDP establish lower bounds for convergence rates, while feasible points in the dual SDP establish upper bounds for convergence rates. However, interpolation conditions for ρ-weakly convex functions with M-bounded subgradients are presently unknown, although it is known that the naive discretization is only necessary and not sufficient.

1.3. Notation

The ambient vector space is $\mathbb{R}^n$, unless otherwise specified. We use N to denote the number of iterations of a given algorithm, and $\star$ as an index to denote the optimal value or point (e.g. $f^\star$ or $x^\star$). We denote the Euclidean 2-norm by $\|\cdot\|$. We will not consider other norms. We also denote the set of L-smooth, μ-strongly convex functions by $\mathcal{F}_{\mu,L}$, where $0 \leq \mu < L \leq \infty$. Note that if $\mu = 0$ then we recover the set of convex (not necessarily strongly convex) functions, $F_{0,L}$. If $\mu < 0$, then we have the set of weakly convex functions. Similarly if $L = \infty$ we can assume that f is not necessarily L-smooth, giving $\mathcal{F}_{\mu,\infty}$. Thus, $F_{0,\infty}$ will denote the set of proper closed convex functions. Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$, we denote its convex conjugate by $f^\star$. This overlaps with the optimality notation, but it will be clear from context whether we are referring to conjugation or optimality. We also denote the limiting subdifferential of f at the point x by $\partial f(x)$. When f is convex, this coincides with the convex subdifferential.

2. Formulating the Performance Estimation Problem

2.1. Interpolation for Smooth Strongly Convex Functions

We reconsider (f-PEP). The central issue is that the decision variable, f ranges over an infinite dimensional space $\mathcal{F}$. For example, the set of all proper closed convex functions, denoted $\mathcal{F}_{0,\infty}$, forms a cone (not proper) in the space of all proper lower semicontinuous functions (a proof of this can be found in Rockafellar (1970)), which is infinite-dimensional. To reduce this to a finite dimensional constraint, we require a set of inequalities of finitely many variables which can produce an equivalent problem.

In his PhD thesis, Taylor et al. (2016) uses convex conjugation (also called Legendre-Fenchel conjugation) to describe a finite set of linear-quadratic inequalities which give sufficient and necessary conditions for a set of points, function values, and subgradients to be interpolated by a convex function. We describe the process in more detail in Appendix A.

Definition 1 *Let $\{(x_i, f_i, g_i)\}_{i=0,1,\dots,N,\star} \subseteq \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ be a set of points, function values, and subgradients. Let $\mathcal{F}$ be a collection of subdifferentiable functions from $\mathbb{R}^n \to \mathbb{R}$. We say that $\{(x_i, f_i, g_i)\}_{i=1,\dots,N,\star}$ is $\mathcal{F}$-**interpolable** if there exists $f \in \mathcal{F}$ such that*

$$\begin{aligned} f(x_i) &= f_i, \\ g_i &\in \partial f(x_i) \quad \forall i = 0, 1, \dots, N, \star. \end{aligned}$$

Thus, we have the following theorems:

Theorem 2 (Taylor et al. (2017a), Theorem 3.2) *Consider a function $f \in \mathcal{F}_{0,\infty}$. Then, $f \in \mathcal{F}_{\mu,L}$ if and only if $f^\star \in \mathcal{F}_{\frac{1}{\mu},\frac{1}{L}}$.*

Proof *This follows from the fact that $g \in \partial f(x) \iff x \in \partial f^\star(g)$. A proof of this equivalence can be found in Rockafellar (1970).* ■

Theorem 3 (Taylor et al. (2016), Theorem 4) *Let $\{(x_i, f_i, g_i)\}_{i=0,1,\dots,N,\star} \subseteq \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$. Then, $\{(x_i, f_i, g_i)\}_{i=0,1,\dots,N,\star}$ is $\mathcal{F}_{\mu,L}$-interpolable if and only if*

$$f_i - f_j - \langle g_i, x_i - x_j \rangle \geq \frac{1}{2(1 - \frac{\mu}{L})} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right)$$

is satisfied for all pairs of indices $i, j = 0, 1 \dots, N, \star$.

Using these theorems, we can reformulate (f-PEP) into a finite dimensional problem. We consider a concrete example:

Example 1 *Consider gradient descent with step size $\frac{1}{k}$. That is, $\mathcal{M}$ is defined by*

$$x^{(k+1)} = x^{(k)} - \frac{1}{k} \nabla f(x^{(k)}).$$

We choose $\mathcal{E}$ as the objective gap $f(x^{(N)}) - f(x^\star)$. Other possible choices might be distance from optimality, $\|x^{(N)} - x^\star\|^2$, or gradient norm, $\|\nabla f(x^{(N)})\|^2$. So that we can use Theorem 3, we choose $\mathcal{F} = \mathcal{F}_{\mu,L}$, the set of μ-strongly convex L-smooth functions. Lastly, for our intial conditions $\mathcal{C}$, we *assume that* $\|x^{(0)} - x^\star\| \leq R$*, for some nonnegative radius* R*. Note that* μ, L*, and* R *are variables decided beforehand (otherwise the cost can be arbitrarily large), and these match assumptions made in standard texts (see Nesterov (2014)). Thus, (f-PEP) becomes*

$$
\begin{aligned}
\max_{\{(x_i, f_i, g_i)\}_{i=0,1,\dots,N,\star}} \quad & f_N - f_\star \\
\textit{subject to} \quad & f_i - f_j - \langle g_i, x_i - x_j \rangle \geq \frac{1}{2(1-\frac{\mu}{L})} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 \right. \\
& \qquad \left. - 2\frac{\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right), \\
& g_\star = 0, \\
& x_{k+1} = x_k - \frac{1}{k+1} g_k, \quad k = 0, 1, \dots, N-1, \\
& \|x_0 - x_\star\|^2 \leq R^2.
\end{aligned}
$$

The equivalence of the above problem and (f-PEP) can be seen as follows. Any function $f \in \mathcal{F}$ *which is feasible to (f-PEP) must produce iterates which obey the given iterate update rule. We can simply extract the iterates of gradient descent applied to* f *to generate a feasible point of the above problem with the same cost. Now, suppose we are given a set of indices* $\{(x_i, f_i, g_i)\}_{i=0,1,\dots,N,\star}$*. If this set is feasible for the above problem, then by Theorem 3, there exists a convex function* $\hat{f}$ *such that* $\hat{f}$ *interpolates the points* $\{(x_i, f_i, g_i)\}_{i=0,1,\dots,N,\star}$*. Therefore, gradient descent produces the same iterates and therefore the cost is the same. Thus, the two problems are equivalent.*

Remark 4 *Although the above formulation looks complicated, notice that the constraints are all quadratic or linear in the decision variables. We have* $(2n+1) \times (N+2)$ *variables, so although large, this problem is in fact a QCQP. We will see later how to lift this to a more tractable SDP relaxation.*

2.2. SDP Formulation of the PEP

We describe the process of relaxing (f-PEP) to an SDP relaxation. First, we define two matrices which collect information from all the past oracle calls:

1. $P_N = \begin{bmatrix} x_0 & x_1 & \dots x_N & x_\star & g_0 & g_1 & \dots g_N & g_\star \end{bmatrix} \in \mathbb{R}^{n \times 2(N+2)}$
2. $F_N = \begin{bmatrix} f_0 & f_1 & \dots f_N & f_\star \end{bmatrix}^\top \in \mathbb{R}^{N+2}$

In order to extend this problem to a QCQP, we define the Gram matrix

$$
\begin{aligned}
G_N &= P_N^\top P_N \\
&= \begin{bmatrix}
\langle x_0, x_0 \rangle & \dots & \langle x_0, x_N \rangle & \langle x_0, x_\star \rangle & \langle x_0, g_0 \rangle & \dots & \langle x_0, g_N \rangle & \langle x_0, g_\star \rangle \\
\langle x_1, x_0 \rangle & \dots & \langle x_1, x_N \rangle & \langle x_1, x_\star \rangle & \langle x_1, g_0 \rangle & \dots & \langle x_1, g_N \rangle & \langle x_1, g_\star \rangle \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\langle g_0, x_0 \rangle & \dots & \langle g_0, x_N \rangle & \langle g_0, x_\star \rangle & \langle g_0, g_0 \rangle & \dots & \langle g_0, g_N \rangle & \langle g_0, g_\star \rangle \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\langle g_N, x_0 \rangle & \dots & \langle g_N, x_N \rangle & \langle g_N, x_\star \rangle & \langle g_N, g_0 \rangle & \dots & \langle g_N, g_N \rangle & \langle g_N, g_\star \rangle \\
\langle g_\star, x_0 \rangle & \dots & \langle g_\star, x_N \rangle & \langle g_\star, x_\star \rangle & \langle g_\star, g_0 \rangle & \dots & \langle g_\star, g_N \rangle & \langle g_\star, g_\star \rangle
\end{bmatrix}.
\end{aligned}
$$

Note that $G_N \in \mathbb{S}_+^{2(N+2)}$, so its dimension does not depend on the dimension of the ambient vector space, $\mathbb{R}^n$. Additionally, for certain cases we can reduce the dimension further by applying first-order optimality conditions or an affine translation, which we will specify in more detail later.

If the interpolation conditions for $\mathcal{F}$ correspond to a set of linear inequalities in the entries of G_N and F_N, then we can redefine the constraints as linear functions of G_N and F_N.

Example 2 *We continue Example 1 from the previous section. Let e_i denote the ith standard basis vector of $\mathbb{R}^{N+2}$. Notice that the cost $f_N - f_\star$ can be rewritten as $\langle e_{N+2}, F_N\rangle - \langle e_1, F_N\rangle$. Similarly, the constraint $g_\star = 0$ is equivalent to $\langle g_\star, g_\star\rangle = 0$. In terms of G_N this becomes $e_{N+2}^\top G_N e_{N+2} = \langle G_N, e_{N+2}e_{N+2}^\top\rangle = 0$. The constraint $\|x_0 - x_\star\|^2 \le R^2$ can be rewritten similarly.*

Let us now consider the constraint $x_{k+1} = x_k - \frac{1}{k+1}g_k$ for $k = 0, 1, \ldots, N-1$. This is equivalent to $x_{k+1} - x_k + \frac{1}{k}g_k = 0$, which is equivalent to $\|x_{k+1} - x_k + \frac{1}{k}g_k\|^2 = 0$. We can rewrite this in terms of linear combinations of the matrix P_N, to get

$$0 = \|x_{k+1} - x_k + \frac{1}{k}g_k\|^2 = \|P_N(e_{k+1} - e_k + \frac{1}{k}e_{N+2+k+1})\|^2 = \langle m_{k+1}m_{k+1}^\top, G_N\rangle,$$

where we define $m_{k+1} = e_{k+1} - e_k + \frac{1}{k}e_{N+2+k+1}$, the vector corresponding to the linear equality constraint of the method $\mathcal{M}$. Note that we could relax this equality to an inequality to handle an inexact update for a fixed tolerance as well.

The interpolation constraint can be handled similarly, although it is much more involved. For the sake of space, and since it does not add any additional insight into the process, we will leave it out. This reformulation gives

$$\begin{aligned}
\max_{\substack{F_N \in \mathbb{R}^{N+2} \\ G_N \in \mathbb{S}^{2(N+2)}}} \quad & \langle e_{N+2} - e_1, F_N\rangle \\
\textit{subject to} \quad & c_{ij}^\top F_N + \langle D_{ij}, G_N\rangle \le 0, \quad i,j = 0, 1 \ldots, N, \star \\
& \langle e_{N+2}e_{N+2}^\top, G_N\rangle = 0, \\
& \langle m_{k+1}m_{k+1}^\top, G_N\rangle, \quad k = 0, \ldots, N-1, \\
& \langle (e_0 - e_{N+2})(e_0 - e_{N+2})^\top, G_N\rangle \le R^2 \\
& G_N \succeq 0 \\
& \operatorname{rank}(G_N) \le n.
\end{aligned}$$

The rank constraint $\operatorname{rank}(G_N) \le n$ ensures that by a Cholesky factorization, we can always recover the matrix P_N and obtain a set of $\mathcal{F}_{\mu,L}$-interpolable points $\{(x_i, f_i, g_i)\}_{i=0,1,\ldots,N,\star}$, so the SDP and QCQP are equivalent. By dropping the rank constraint, we find an SDP relaxation of (f-PEP).

Remark 5 *An interesting fact is that by an affine translation, we can assume that $x_\star = 0$ and $f_\star = 0$. By the first-order conditions of convex optimization, we also get that $g_\star = 0$. This could allow us to eliminate 2 rows of G_N and an entry of F_N, resulting in eliminating $2(N+2)+1$ variables.*

We will now generalize this example to arbitrary classes of functions, methods, and initial conditions.

Definition 6 *Let $\mathcal{F}$ be a class of functions, $\mathcal{M}$ an iterative first-order method, $\mathcal{C}$ a set of initial conditions, and $\mathcal{E}$ a performance measure depending only on the first-order oracle calls. We say that any of these are **Gram representable** if it has a set of interpolation conditions which can be represented as linear function of the Gram matrix G_N.*

Remark 7 *This definition is actually more restrictive than necessary, see Taylor et al. (2017a).*

We now define the generalized form of Example 2.

Definition 8 *Let $\mathcal{F}$ be a class of functions, $\mathcal{M}$ an iterative first-order method, $\mathcal{C}$ a set of initial conditions, and $\mathcal{E}$ a performance measure depending only on the first-order oracle calls. Suppose that all of these are Gram representable. Then we can reformulate (f-PEP) as*

$$\begin{aligned} \max_{\substack{F_N \in \mathbb{R}^{N+2} \\ G_N \in \mathbb{S}^{2(N+2)}}} \quad & \langle e, F_N \rangle + \langle E, G_N \rangle \\ \textit{subject to} \quad & a_{ij} + \langle C_{ij}, F_N \rangle + \langle D_{ij}, G_N \rangle \leq 0, \quad i,j = 0,1\dots,N,\star \\ & b_k + \langle m_k, F_N \rangle + \langle M_k, G_N \rangle \leq 0, \quad k = 1,\dots,N, \\ & d + \langle c, F_N \rangle + \langle C, G_N \rangle \leq 0 \\ & G_N \succeq 0. \end{aligned} \tag{SDP-PEP}$$

The first constraint corresponds to the interpolation conditions for $\mathcal{F}$. The second is the reformulation of the method $\mathcal{M}$. We allow for inequalties to generalize to inexact updates, the kth such constraint might represent the tolerance at step k. The third constraint represents the intial conditions, $\mathcal{C}$. Finally, we drop the rank constraint seen in Example 2.

Remark 9 *Dropping the rank constraint does not change much in practice. Indeed, simply assuming that the dimension, n, of the ambient space, $\mathbb{R}^n$, is larger than $2(N+2)$ guarantees that the rank constraint is trivial. In the setting of large-scale optimization, the number of iterations may be far smaller than the dimension of the space, so the relaxation is exact in this context.*

2.3. Weakly Convex Functions

Definition 10 (Davis and Drusvyatskiy (2018), Lemma 2.1) *A function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is called ρ-**weakly convex** if $\varphi(x) + \frac{\rho}{2}\|x\|^2$ is convex. Equivalently, for all $x, y \in \mathbb{R}^n$ with $v \in \partial\varphi(x)$ and $w \in \partial\varphi(y)$,*

1. *The approximate secant inequality*

$$\varphi(\lambda x + (1-\lambda)x) \leq \lambda\varphi(x) + (1-\lambda)\varphi(x) + \frac{\rho\lambda(1-\lambda)}{2}\|x-y\|^2$$

for all $\lambda \in [0,1]$.

2. *The subgradient inequality holds*

$$\varphi(y) \geq \varphi(x) + \langle v, y-x \rangle - \frac{\rho}{2}\|x-y\|^2.$$

3. The subdifferential map is hypomontone:

$$\langle v - w, x - y\rangle \geq -\rho\|x - y\|^2.$$

Another proof of these results can be found in (Rockafellar, 1998, Theorem 12.17).

Weakly convex functions have various nice properties, which can be found in Davis and Drusvyatskiy (2018). A set of interpolation conditions for weakly convex functions are known, and the following theorem characterizes them.

Theorem 11 (Taylor et al. (2017a), Theorem 3.10) *The set* $\{(x_i, f_i, g_i)\}_{i=0,1,\ldots,N,\star} \subseteq \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ *is* $\mathcal{F}_{-\rho,L}$*-interpolable if and only if*

$$f_i + \frac{\rho}{2}\|x_i\|^2 \geq f_j + \frac{\rho}{2}\|x_j\|^2 + \langle g_j + \rho x_j, x_i - x_j\rangle + \frac{1}{2L}\|g_i + \rho x_i - g_j - \rho x_j\|^2$$

is satisfied for all pairs of indices $i, j = 0, 1 \ldots, N, \star$.

Proof *Note that* $\{(x_i, f_i, g_i)\}_{i=0,1,\ldots,N,\star}$ *being* ρ*-weakly convex interpolable is equivalent to the set* $\left\{(x_i, f_i + \frac{\rho}{2}\|x_i\|^2, g_i + \rho x_i)\right\}_{i=0,1,\ldots,N,\star}$ *being* $\mathcal{F}_{0,L+\rho}$*-interpolable. An application of Theorem 3 completes the proof.* ■

However, in practice we do not usually consider the L-smooth case. Instead we consider an alternate class of functions:

Definition 12 *Let* $f\colon \mathbb{R}^n \to \mathbb{R}$ *be a subdifferentiable function. We say that* f *has* M***-bounded subgradients*** *if for all* x*, if* $g \in \partial f(x)$ *then* $\|g\| \leq M$.

The class of ρ-weakly convex functions with M-bounded subgradients appear in Davis and Drusvyatskiy (2018) and Nemirovski et al. (2009). Similar conditions appear in Grimmer (2018), and Zhu et al. (2023) provides a survey of such subgradient bounds and their associated convergence rates. However, as of writing, interpolation conditions for the class of ρ-weakly convex functions are unknown, although Rubbens et al. (2023) states they have (as of writing) an unpublished result on this topic.

2.4. Interpolating Weakly Convex Functions with Bounded Subgradients

Definition 10 provides some candidates to examine for a set of interpolation conditions. However, Rubbens et al. (2023) state that the subgradient inequality in Definition 10, together with the constraint $\|g_i\|^2 \leq M^2$, are necessary but not sufficient. Therefore, if we use those two constraints in (SDP-PEP), we will obtain a relaxation which includes all ρ-weakly-convex functions with M-bounded subgradients, but could potentially be a strictly larger class of functions. An overview of numerical results using this method can be found in Section 3.

To interpolate convex functions, Taylor et al. (2016) used convex conjugation. However, the convex conjugate of a ρ-weakly convex function may not be proper. For example, consider $\varphi(x) = -\frac{\rho}{2}\|x\|^2$. The convex conjugate can be computed as $f^\star(y) = \sup_x \left\{\langle y, x\rangle - \left(-\frac{\rho}{2}\|x\|^2\right)\right\} = \infty$ for all y. Thus, the conjugate function does not provide anything useful in this context.

The proof of Theorem 11 bypasses this restriction by considering the convex function $f(x) := \varphi(x) + \frac{\rho}{2}\|x\|^2$. However, notice that if $g \in \partial\varphi(x)$, then $g + \rho x \in \partial f(x)$, and the norm $\|g + \rho x\|^2$ is unbounded as a function of x, so the property of bounded subgradients is lost.

To repair this, we can consider the generalized conjugate from Rockafellar (1998).

Definition 13 (Rockafellar (1998), Chapter 11, Section L) *Let $\Phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be any function. For any function $f\colon \mathbb{R}^n \to \mathbb{R}$, the Φ-**conjugate** of f is defined as*

$$f^{\Phi}(y) = \sup_{x\in\mathbb{R}^n} \left\{\Phi(x,y) - f(x)\right\}.$$

*Similarly, the Φ-**biconjugate** of f is*

$$f^{\Phi\Phi}(x) = \sup_{y\in\mathbb{R}^n} \left\{\Phi(x,y) - f^{\Phi}(x)\right\}.$$

The goal is to use the generalized conjugate function $\Phi(x,y) = \langle x, y\rangle - \frac{\rho}{2}\|x\|^2$ to act as the convex conjugate for the class of ρ-weakly convex functions. However, the critical step is to analyze how the subgradient is transformed under conjugation, but as of writing I have not yet fully analyzed this.

3. Numerical Experiments

As previously stated, exact interpolation conditions for the class of ρ-weakly convex functions with M-bounded subgradients is unknown. However, a set of necessary conditions is given in Definition 10, together with the constraint $\|g_i\|^2 \leq M^2$. We consider the subgradient update methods given in Davis et al. (2018).

Algorithm 1: Subgradient Method

Data: $x^{(0)} \in \mathbb{R}^n$, $\varepsilon > 0$
Result: $x^{(N)} \in \mathbb{R}^n$

```
while k ≥ 0 do
    Choose g^(k) ∈ ∂φ(x)
    if ||g^(k)|| ≤ ε then
        Stop, exit algorithm
    else
        x^(k+1) ← x^(k) − (c^(k) / ||g^(k)||) g^(k)
    end
end
```

The step sizes $c^{(k)}$ are predetermined, and we consider a few different choices:

1. Polyak stepsize: $c^{(k)} = \frac{\varphi(x^{(k)}) - \varphi(x^{\star})}{\|g^{(k)}\|}$,
2. Constant stepsize: $c^{(k)} = \alpha$, where α is a constant,
3. Geometric stepsize: $c^{(k)} = \alpha \cdot q^k$, where α is a constant and $q \in (0,1)$.

Note that this is computable only if we know the optimal value in advance. Thus, this stepsize scheme is only of theoretical interest, although it can be used for estimating better stepsizes.

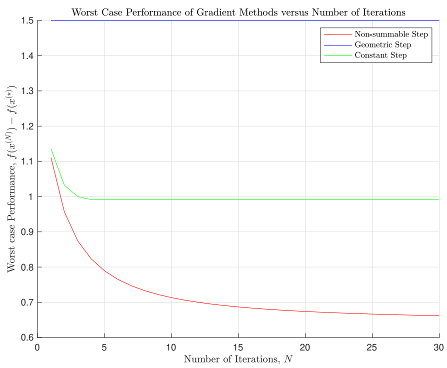

Figure 1: Comparison of Different Step Sizes on Weakly Convex Functions

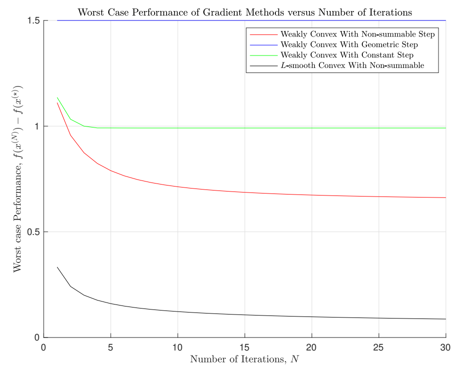

Figure 2: Comparison of Weakly Convex Functions and L-smooth Convex Functions

3.1.

The problem (SDP-PEP) works in a general form, but as subtely noted in Example 2, the constraint $x_{k+1} = x_k - \frac{1}{k} g_k$ defines an *implicit* rank constraint on our problem. Since the some columns of the matrix are in the span of the others, this is essentially requiring the rank of the matrix G_N to be below a certain threshold. To alleviate this, we redefine the variables $x_1, \ldots, x_N$ in terms of the gradient updates to get $x_k = x_0 - \sum_{i=1}^{k} \frac{1}{k} g_{k-1}$. This allows us to change our variables into constraints. Note that this applies to any method where the current iterate is in the span of the gradients. I ran the numerical experiments on a 2020 MacBook Pro using YALMIP with MOSEK, and set $\rho = M = 1$ and $L = 2$ for the L-smooth convex comparison setup. I ran the SDP with $N = 1, 2, \ldots, 30$, since anything larger incurred significant time cost. The code can be found in Appendix B.

4. Conclusions

We can see that although the weakly convex functions do not perform as well as the L-smooth convex functions, the non-summable decaying step sizes perform the best on the class of functions which satisfy the necessary conditions for ρ-weakly convex functions with M-bounded subgradients. Future work must focus on rigorously justifying the use of interpolation conditions, and currently the generalized conjugate of Rockafellar appears to be the best candidate. On the numerical side, it is worth investigating the ways in which the solve time can be sped up, since any $N > 50$ causes significant issues for the solver. The main issue appears to be the number of constraints, which grow at a rate of $O(N^2)$, due to needing an interpolation constraint for each pair of indices. Grimmer (2023) notes that many primal optimal solutions have rank one, so methods which can take advantage of this observed rank deficiency, such as the spectral bundle method of Liao et al. (2023) may be worth investigating as a potential alternative to interior-point methods.

References

Sebastien Colla and Julien M. Hendrickx. Automatic performance estimation for decentralized optimization. *IEEE Transactions on Automatic Control*, 68(12):7136–7150, December 2023. ISSN 2334-3303. doi: 10.1109/tac.2023.3251902. URL http://dx.doi.org/10.1109/TAC.2023.3251902.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions, 2018.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions, 2018.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach, 2012.

Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity, 2018.

Benjamin Grimmer. Provably faster gradient descent via long steps, 2023.

Shuvomoy Das Gupta, Bart P. G. Van Parys, and Ernest K. Ryu. Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods, 2023.

Zihao Hu, Guanghui Wang, Xi Wang, Andre Wibisono, Jacob Abernethy, and Molei Tao. Extragradient type methods for riemannian variational inequality problems, 2023.

Nikita Kornilov, Eduard Gorbunov, Mohammad Alkousa, Fedor Stonyakin, Pavel Dvurechensky, and Alexander Gasnikov. Intermediate gradient methods with relative inexactness, 2023.

Feng-Yi Liao, Lijun Ding, and Yang Zheng. An overview and comparison of spectral bundle methods for primal and dual semidefinite programs, 2023.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970. ISBN 9781400873173. doi: doi:10.1515/9781400873173. URL https://doi.org/10.1515/9781400873173.

Ralph Tyrell Rockafellar. *Monotone Mappings*, pages 533–578. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-642-02431-3. doi: 10.1007/978-3-642-02431-3_12. URL https://doi.org/10.1007/978-3-642-02431-3_12.

Anne Rubbens, Nizar Bousselmi, Sébastien Colla, and Julien M. Hendrickx. Interpolation constraints for computing worst-case bounds in performance estimation problems. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, December 2023. doi: 10.1109/cdc49753.2023.10384170. URL http://dx.doi.org/10.1109/CDC49753.2023.10384170.

Firooz Shahriari-Mehr and Ashkan Panahi. Asynchronous decentralized optimization with constraints: Achievable speeds of convergence for directed graphs, 2024.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods, 2016.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3): 1283–1313, Jan 2017a. doi: 10.1137/16m108104x. URL https://doi.org/10.1137%2F16m108104x.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Performance estimation toolbox (pesto): Automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283, 2017b. doi: 10.1109/CDC.2017.8263832.

Daoli Zhu, Lei Zhao, and Shuzhong Zhang. A unified analysis for the subgradient methods minimizing composite nonconvex, nonsmooth and non-lipschitz functions, 2023.

Appendix A. Convex Interpolation

A.1. Convex Conjugation

The convex conjugate of $f\colon \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f^\star(y) = \sup_x \{\langle y, x\rangle - f(x)\}.$$

It has a number of useful properties.

Theorem 14 *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be any function, and let $f^\star(y)$ be its convex conjugate. Then,*

1. *$f^\star$ is convex and closed, even when f is not.*
2. *Fenchel's Inequality: $f(x) + f^\star(x^\star) \geq \langle x, x^\star\rangle$, and equality holds if and only if $x^\star \in \partial f(x)$.*
3. *$f^{\star\star} \leq f$, and if f is a proper closed convex function, then $f^{\star\star} = f$.*
4. *For a proper closed convex function f, $x^\star \in \partial f(x)$ if and only if $x \in \partial f^\star(x^\star)$.*

Proof *We will use an equivalent definition of convexity (Rockafellar (1970)). Define the epigraph of f as epi $f := \{(x, y) : f(x) \leq y\}$. Then, f is a convex function if and only if epi f is a convex set. The proof of 1 follows from realizing that the mapping $x^\star \mapsto \langle x, x^\star\rangle - f(x)$ is affine, and therefore convex, so its epigraph is convex, and additionally it is also closed. Taking the intersection of all* *epigraphs over all values of x means we take the intersection of convex and closed sets, which is also convex and closed, and this is precisely the epigraph of $f^\star$.*

The proof of 2 follows from the definition:

$$\begin{aligned} f^\star(x^\star) &= \sup_x\{\langle x^\star, x\rangle - f(x)\} \\ &\geq \langle x^\star, x\rangle - f(x) \\ &\iff \\ f(x) + f^\star(x^\star) &\geq \langle x^\star, x\rangle \end{aligned}$$

For the case of equality, we have that

$$\begin{aligned} x^\star \in \partial f(x) &\iff f(z) \geq f(x) + \langle x^\star, z - x\rangle \qquad \forall z \\ &\iff f(z) - \langle x^\star, z\rangle \geq f(x) - \langle x^\star, x\rangle \\ &\iff \langle x^\star, z\rangle - f(z) \leq \langle x^\star, x\rangle - f(x) \\ &\iff \sup_{z\in dom(f)}\{\langle x^\star, z\rangle - f(z)\} \leq \langle x^\star, x\rangle - f(x) \\ &\iff \langle x^\star, x\rangle - f(x) = f^\star(x^\star) \\ &\iff \langle x^\star, x\rangle = f(x) + f^\star(x^\star). \end{aligned}$$

For the proof of 3, we will first show that $f^{\star\star}(x) \leq f(x)$.

$$\begin{aligned} f^{**}(x) &= \sup_z\{x^\top z - f^*(z)\} \\ &= \sup_z\{x^\top z - \sup_y\{z^\top y - f(y)\}\} \\ &= \sup_z\{x^\top z + \inf_y\{f(y) - z^\top y\}\} \\ &= \sup_z \inf_y\{x^\top z + f(y) - z^\top y\} \\ &\leq \inf_y \sup_z\{z^\top(x-y) + f(y)\} \\ &\leq f(x) \end{aligned}$$

Now suppose f is proper, closed, and convex. Then, its epigraph is nonempty, closed, and convex. We know from the previous results that $f^{\star\star}$ is also closed and convex, and since $f^{\star\star} \leq f$, it must also be proper. Therefore, its epigraph is also nonempty, closed, and convex. We want to show that $f^{\star\star} \geq f$ as well. Suppose that $f^{\star\star}(x) < f(x)$ at some point x. Then, by the hyperplane separation theorem, we can strictly separate epi f from $(x, f^{\star\star}(x))$, and this hyperplane cannot be vertical. This gives, for some $\varepsilon > 0$ and vector y,

$$\begin{aligned} f(z) - \varepsilon &\geq \langle y, z - x\rangle + f^{\star\star}(x) \quad \forall\ z \\ \langle y, x\rangle - \varepsilon &\geq \langle y, z\rangle - f(z) + f^{\star\star}(x) \\ \langle y, x\rangle - \varepsilon &\geq f^\star(y) + f^{\star\star}(x) \\ \langle y, x\rangle &> f^\star(y) + f^{\star\star}(x) \end{aligned}$$

However, this contradicts Fenchel's inequality, so therefore $f^{\star\star} = f$. ■

A.2. Interpolation Conditions

We can use this to begin the proof of the interpolation results.

Theorem 15 *Let f be a proper closed convex function. Then, f is L-smooth ($f \in \mathcal{F}_{0,L}$) if and only if $f^\star$ is $\frac{1}{L}$-strongly convex ($f \in \mathcal{F}_{\frac{1}{L},\infty}$).*

Proof *We use the following characterizations of L-smooth and μ-strongly convex functions.*

$$f \in \mathcal{F}_{0,L} \iff$$

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

$$f \in \mathcal{F}_{\mu,\infty} \iff$$

$$\mu\|x - y\|^2 \leq \langle \nabla f(y) - \nabla f(x), y - x\rangle \leq \frac{1}{\mu}\|\nabla f(y) - \nabla f(x)\|^2$$

The result then follows from realizing that $\nabla f^*(\nabla f(x)) = x$. ■

We need one last lemma to transform μ-strongly convex functions to general convex functions.

Lemma 16 *Consider $f \in \mathcal{F}_{\mu,L}$ with $0 \leq \mu < L \leq \infty$. Define $\phi(x) := f(x) - \frac{\mu}{2}\|x - x_\star\|^1$. Then, $\phi \in \mathcal{F}_{0,L-\mu}$ if and only if $f \in \mathcal{F}_{\mu,L}$. The mapping of $f \mapsto f - \frac{\mu}{2}\|\cdot\|^2$ is known as **minimal curvature subtraction**.*

Proof *Suppose $f \in \mathcal{F}_{\mu,L}$. Then,*

$$\begin{aligned}
\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle &= \langle \nabla f(x) - \nabla f(y) - \mu(x - y), x - y\rangle \\
&= \langle \nabla f(x) - \nabla f(y), x - y\rangle - \mu\|x - y\|^2 \\
&\leq L\|x - y\|^2 - \mu\|x - y\|^2 \\
&\leq (L - \mu)\|x - y\|^2.
\end{aligned}$$

We use the inequality $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|^2$ if and only if $f \in \mathcal{F}_{0,L}$ (assuming f is C^1). A proof of this can be found in Nesterov (2014).

Suppose $\phi \in \mathcal{F}_{0,L-\mu}$. Then, $f(x) = \phi(x) + \frac{\mu}{2}\|x - x_\star\|^2$. By convexity and $(L-\mu)$-smoothness:

$$\begin{aligned}
\langle \nabla f(x) - \nabla f(y), x - y\rangle &= \langle \nabla\phi(x) - \nabla\phi(y) + \mu(x - y), x - y\rangle \\
&= \langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle + \mu\|x - y\|^2 \\
&\leq (L - \mu)\|x - y\|^2 + \mu\|x - y\|^2 \\
&= L\|x - y\|^2.
\end{aligned}$$

This proves L-smoothness. For strong convexity, we have that

$$\begin{aligned}
0 &\leq \langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \\
&= \langle \nabla f(x) - \nabla f(y), x - y\rangle - \mu\|x - y\|^2 \\
\mu\|x - y\|^2 &\leq \langle \nabla f(x) - \nabla f(y), x - y\rangle.
\end{aligned}$$

This inequality is an equivalent definition of strong convexity for C^1 functions (Nesterov (2014)). ■

Now, we can state the main result of Taylor et al. (2016):

Theorem 17 *A set* $\{(x_i, f_i, g_i)\}_{i\in I}$ *is* $\mathcal{F}_{\mu,L}$*-interpolable if and only if*

$$f_i - f_j - \langle g_j, x_i - x_j\rangle \geq \frac{1}{2\left(1-\frac{\mu}{L}\right)}\left(\frac{1}{L}\|g_i - g_j\|^2 + \mu\|x_i - x_j\|^2 - 2\frac{\mu}{L}\langle g_j - g_i, x_j - x_i\rangle\right)$$

Proof *The full proof is very long and doesn't add anything new. We focus on the main ideas.*

1. *Show that* $\{(x_i, f_i, g_i)\}_{i\in I}$ *is* $\mathcal{F}_{\mu,L}$*-interpolable if and only if* $\{(g_i, f_i, x_i)\}_{i\in I}$ *is* $\mathcal{F}_{\frac{1}{L},\frac{1}{\mu}}$*-interpolable.*
2. *Then, starting with a* μ*-strongly convex,* L*-smooth function* f*, we use minimal curvature subtraction to obtain an* $(L-\mu)$*-smooth function* $\tilde{f}$*.*
3. *Then, we take* $\tilde{f}$ *and map it to its conjugate* $\tilde{f}^\star$ *to obtain a* $\frac{1}{L-\mu}$*-strongly convex function.*
4. *We again use minimal curvature subtraction to obtain a convex* $h^\star$*, which we can use our nonsmooth convex interpolation result for.*

∎

Appendix B. Numerical Experiments

The code can be found at

https://github.com/pranavnreddy/ECE285Project.

We chose to use a constant stepsize $c_k = 1$, a geometric step size $c_k = 100 \cdot 0.8^k$, and a non-summable decaying stepsize $c_k = \frac{1}{k}$. For the convex L-smooth function we chose to use the best performing stepsize, the non-summable decaying $c_k = \frac{1}{k}$.