Proceedings of Machine Learning Research vol 242:1–13, 2024

Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods

Feng-Yi Liao FLIAO@UCSD.EDU
Department of Electrical and Computer Engineering, University of California San Diego

Lijun Ding LDING47@WISC.EDU
Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison

Yang Zheng ZHENGY@UCSD.EDU
Department of Electrical and Computer Engineering, University of California San Diego

Abstract

Many machine learning problems lack strong convexity properties. Fortunately, recent studies have revealed that first-order algorithms also enjoy linear convergences under various weaker regularity conditions. While the relationship among different conditions for convex and smooth functions is well understood, it is not the case for the nonsmooth setting. In this paper, we go beyond convexity and smoothness, and clarify the connections among common regularity conditions (including *strong convexity, restricted secant inequality, subdifferential error bound, Polyak-Łojasiewicz inequality, and quadratic growth*) in the class of weakly convex functions. In addition, we present a simple and modular proof for the linear convergence of the *proximal point method* (PPM) for convex (possibly nonsmooth) optimization using these regularity conditions. The linear convergence also holds when the subproblems of PPM are solved inexactly with a proper control of inexactness.

Keywords: Error bound, Polyak-Łojasiewicz inequality, quadratic growth, proximal point method.

1. Introduction

Machine learning has shown impressive performance on a wide range of applications. Behind these successes, (sub)gradient-based methods and their variants are the workhorse algorithms. Many studies have investigated the theoretical foundations of these first-order iterative algorithms. For smooth and/or convex cases, (sub)gradient methods are most well-understood (Nesterov, 2018). It is well-known that the basic gradient descent algorithm achieves linear convergence for minimizing smooth and strongly convex functions. However, strong convexity is a very strong assumption, and many fundamental models in machine learning lack this good property (Agarwal et al., 2010).

Alternative regularity conditions that are weaker than strong convexity have been revealed in the past. For example, gradient descent also converges linearly under *Polyak-Łojasiewicz* inequality or *restricted secant inequality* (Polyak, 1963; Zhang and Yin, 2013; Guille-Escuret et al., 2022). These two conditions can even hold for nonconvex functions. The classical bundle method converges linearly under *quadratic growth* for smooth convex functions (Díaz and Grimmer, 2023). This linear convergence result has recently been extended for general semidefinite optimization (a very broad class of conic programs) in Ding and Grimmer (2023); Liao et al. (2023b). While smooth and convex (but not strongly convex) problems cover a variety of applications, modern machine learning practice routinely deals with problems lacking both qualities (e.g., training nonsmooth and nonconvex deep networks). Recent studies have further identified one amenable problem class: *weakly convex* (Davis and Drusvyatskiy, 2019). This class of problems includes all convex functions, L-smooth functions, certain compositions of convex functions with smooth functions, and many cost functions in modern machine learning (Drusvyatskiy and Davis, 2020; Atenas et al., 2023). For

nonsmooth problems, it is known that *subdifferential error bound* (or metrically subregularity) or error bound for *proximal gradient mapping* is sufficient to ensure linear convergence of proximal algorithms (Ye et al., 2021; Drusvyatskiy and Lewis, 2018). Very recently, Atenas et al. (2023) also uses error bound properties to establish linear convergence for proximal-type methods.

While a couple of weaker conditions ensure linear convergence of many first-order algorithms, their relationship remains unclear, especially in the class of weakly convex functions. Recently, it has been revealed that some regularity conditions (such as PL, error bound, and quadratic growth) are equivalent (Drusvyatskiy and Lewis, 2018; Drusvyatskiy et al., 2021; Bolte et al., 2017; Karimi et al., 2016; Ye et al., 2021; Zhu et al., 2023). However, many existing results require smooth and/or convex settings; we postpone a detailed discussion of these results in Remark 1 after introducing relevant notations and our main results. In this paper, we have two main contributions: 1) we first clarify the relationship among common regularity conditions in the class of weakly convex functions (Theorem 3.1); 2) we present a simple and modular proof for linear convergence of the classical proximal point method (PPM) (Rockafellar, 1976b) under these regularity conditions (Theorem 4.2). These linear convergence results hold for inexact PPM when controlling stopping criteria properly (Theorem 5.3). We remark that our convergence results require weaker conditions with simpler proofs. We expect their applications to sparse and large-scale conic optimization (Zheng et al., 2021).

The rest of this paper is structured as follows. Section 2 presents a motivation and revisits linear convergence of gradient descent. Section 3 presents the relationship among different regularity conditions. Sections 4 and 5 focus on the (inexact) PPM and establish the sublinear and linear convergences. Section 6 presents three numerical experiments, and Section 7 concludes this paper. Some extra discussions and technical proofs are provided in our technical report (Liao et al., 2023a).

Notation. We use $\mathbb{R}^n$ to denote n-dimensional Euclidean space and $\overline{\mathbb{R}}$ to denote the extended real line, i.e., $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. The notations $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ stand for standard inner product and ℓ_2 norm in $\mathbb{R}^n$. For a closed set $S \subseteq \mathbb{R}^n$, the distance of a point $x \in \mathbb{R}^n$ to S is defined as $\mathrm{dist}(x, S) := \min_{y \in S} \|x - y\|$ and the projection of x onto S is denoted as $\Pi_S(x) = \mathrm{argmin}_{y \in S} \|x - y\|$. The symbol $[f \leq \nu] := \{x \in \mathbb{R}^n \mid f(x) \leq \nu\}$ denotes the ν-sublevel set of f.

2. Motivation: Linear convergence of gradient descent algorithms

To motivate our discussion, consider a smooth convex optimization problem $\min_x f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and L-smooth function, i.e., its gradient is L-Lipschitz satisfying $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, $\forall x, y \in \mathbb{R}^n$. Let $S := \mathrm{argmin}\ f(x)$ be the set of optimal solutions. Assume $S \neq \emptyset$ and denote $f^\star = \min_x f(x)$. The standard *gradient descent* (GD) follows the update

$$x_{k+1} = x_k - t_k \nabla f(x_k), \tag{1}$$

where $t_k > 0$ is the step size. A textbook result says that when choosing a constant step size $t_k = \frac{1}{L}$, the GD iterates converge to $f^\star$ with a *sublinear* rate (precisely, $f(x_k) - f^\star \leq L\mathrm{dist}^2(x_0, S)/(2k)$); see e.g., Corollary 2.1.2 of Nesterov (2018). If the function f is strongly convex, GD achieves a global linear convergence (Nesterov, 2018, Theorem 2.1.15).

However, the assumption of strong convexity is often not satisfied in practice. It is known that some alternative weaker assumptions are sufficient for linear convergences. We here introduce two notions: *restricted secant inequality* (RSI) (Zhang and Yin, 2013), and *Polyak-Łojasiewicz* (PL)

inequality (Polyak, 1963). A differentiable function f satisfies RSI if there exists $\mu_{\rm r} > 0$ such that

$$\langle \nabla f(x), x - \Pi_S(x) \rangle \geq \mu_{\rm r} \|x - \Pi_S(x)\|^2 = \mu_{\rm r} \cdot \mathrm{dist}^2(x, S), \ \forall x \in \mathbb{R}^n \tag{2}$$

and it satisfies the PL inequality if there exists $\mu_{\rm p} > 0$ such that

$$\|\nabla f(x)\|^2 \geq 2\mu_{\rm p}(f(x) - f^\star), \ \forall x \in \mathbb{R}^n. \tag{3}$$

Note that both RSI (2) and PL (3) imply that any stationary point of f is a global minimum. However, they do not imply the uniqueness of stationary points or the convexity of the function. One can think that RSI (2) (resp. PL (3)) requires that the gradient $\nabla f(x)$ grows faster than a quadratic function when moving away from the solution set S (resp. the optimal value $f^\star$). Linear convergence of GD under the PL inequality was first proved in Polyak (1963), and linear convergence under RSI was discussed in Proposition 1 of Guille-Escuret et al. (2022) and Proposition 1 of Zhang (2020). We summarize a simple version below.

Theorem 2.1 (Linear convergence of GD) *Consider the problem $\min_x f(x)$, where f is an L-smooth (possibly nonconvex) function. Suppose its solution set S is nonempty. If RSI (2) holds with $\mu_{\rm r} > 0$ and PL inequality (3) holds with $\mu_{\rm p} > 0$, then the GD algorithm (1) with a constant stepsize $t_k = \frac{\mu_{\rm r}}{L^2}$ has a global linear convergence rate for iterates and function values, i.e.,*

$$\mathrm{dist}(x_{k+1}, S) \leq \omega_1 \cdot \mathrm{dist}(x_k, S), \quad \textit{where} \quad \omega_1 = \sqrt{1 - \mu_{\rm r}^2/L^2} \in (0, 1), \tag{4a}$$

$$f(x_{k+1}) - f^\star \leq \omega_2 \cdot (f(x_k) - f^\star), \quad \textit{where} \quad \omega_2 = (L^3 - 2\mu_{\rm r} L \mu_{\rm p} + \mu_{\rm r}^2 \mu_{\rm p})/L^3 \in (0, 1). \tag{4b}$$

Thanks to RSI (2) and PL inequality (3), the proof of Theorem 2.1 is very elegant and only takes a few lines. We provide a simple proof and some additional discussions in Appendix B in Liao et al. (2023a). In particular, the RSI (2) leads to a quick proof of (4a), and the PL (3) allows for a simple proof of (4b). It is known that for L-smooth convex functions, the two conditions RSI (2) and PL (3) are equivalent (cf. Karimi et al., 2016, Theorem 2). Some recent studies, such as Bolte et al. (2017); Necoara et al. (2019); Zhang (2017), have explored the relationships among different regularity conditions for linear convergences. A nice summary appeared in Theorem 2 of Karimi et al. (2016), but it only works in the context of L-smooth functions. In this paper, we aim to characterize the relationships among different regularity conditions for *nonsmooth and nonconvex* functions (Section 3), and apply them to derive *simple and clean* proofs for linear convergences of (inexact) proximal point methods for convex (possibly nonsmooth) optimization (Sections 4 and 5).

3. Relationships between regularity conditions

In this section, we move away from convex and smooth functions and expand our view to the class of *weakly convex (potentially nonsmooth)* functions. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called ρ-weakly convex if the function $f + \frac{\rho}{2}\|\cdot\|^2$ is convex. The class of weakly convex functions is very rich: it includes all convex functions, L-smooth functions, certain compositions of convex functions with smooth functions, and many cost functions in modern machine learning applications; we refer interested readers to Drusvyatskiy and Davis (2020); Atenas et al. (2023) for more details.

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, closed, ρ-weakly convex function. For this function class, gradients may not always exist. We define the Fréchet subdifferential (see e.g., Page 27 in Li et al. (2020)):

$$\hat{\partial} f(x) = \left\{ s \in \mathbb{R}^n \mid \liminf_{y \to x} \frac{f(y) - f(x) - \langle s, y - x \rangle}{\|y - x\|} \geq 0 \right\}.$$

If f is convex, Fréchet subdifferential $\hat{\partial} f$ is the same as the usual convex subdifferential, i.e., $\hat{\partial} f(x) = \partial f(x) = \{s \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle s, y - x \rangle, \forall y \in \mathbb{R}^n\}, \forall x \in \mathbb{R}^n$; if f is differentiable, Fréchet subdifferential reduces to the usual gradient, i.e., $\hat{\partial} f(x) = \{\nabla f(x)\}, \forall x \in \mathbb{R}^n$.

Let S the optimal solution set of f, i.e., $S = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$, and we assume $S \neq \emptyset$. Let $f^\star = \min_{x \in \mathbb{R}^n} f(x)$ and $\nu > 0$. Consider the following five regularity conditions:

1. **Local Strong Convexity (SC)**: there exists a positive constant $\mu_{\mathrm{s}} > 0$ such that

$$f(x) + \langle g, y - x \rangle + \frac{\mu_{\mathrm{s}}}{2} \cdot \|y - x\|^2 \leq f(y), \quad \forall x, y \in [f \leq f^\star + \nu] \text{ and } g \in \hat{\partial} f(x). \tag{SC}$$

2. **Restricted Secant Inequality (RSI)**: there exists a positive constant $\mu_{\mathrm{r}} > 0$ such that

$$\mu_{\mathrm{r}} \cdot \operatorname{dist}^2(x, S) \leq \langle g, x - \hat{x} \rangle, \quad \forall x \in [f \leq f^\star + \nu],\ g \in \hat{\partial} f(x),\ \hat{x} \in \Pi_S(x). \tag{RSI}$$

3. **Error bound (EB)**[1]: there exists a constant $\mu_{\mathrm{e}} > 0$ such that

$$\operatorname{dist}(x, S) \leq \mu_{\mathrm{e}} \cdot \operatorname{dist}(0, \hat{\partial} f(x)), \quad \forall x \in [f \leq f^\star + \nu]. \tag{EB}$$

4. **Polyak-Łojasiewicz (PL) inequality**[2]: there exists a constant $\mu_{\mathrm{p}} > 0$ such that

$$2\mu_{\mathrm{p}} \cdot (f(x) - f^\star) \leq \operatorname{dist}^2(0, \hat{\partial} f(x)), \quad \forall x \in [f \leq f^\star + \nu]. \tag{PL}$$

5. **Quadratic Growth (QG)**: there exists a constant $\mu_{\mathrm{q}} > 0$ such that

$$\frac{\mu_{\mathrm{q}}}{2} \cdot \operatorname{dist}^2(x, S) \leq f(x) - f^\star, \quad \forall x \in [f \leq f^\star + \nu]. \tag{QG}$$

All the regularity conditions above are defined over a sublevel set $[f \leq f^\star + \nu]$. If $v = +\infty$, then they are global. In particular, (SC) imposes a quadratic lower bound for every point in the sublevel set. On the other hand, (RSI), (EB) and (PL) all require a certain growth of the subdifferential $\hat{\partial} f(x)$ when moving away from its solution set S or optimal value $f^\star$. It is easy to see that (RSI), (EB) and (PL) all imply that every stationary point $0 \in \hat{\partial} f(x)$ in the sublevel set $[f \leq f^\star + \nu]$ is a global minimum (but they do not imply the uniqueness of stationary points). Finally, (QG) shows that $f(x)$ grows at least quadratically when moving away from the solution set S.

Our first technical result summarizes the relationships among the five regularity conditions.

Theorem 3.1 *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed ρ-weakly convex function with $f^\star = \min_x f(x)$ and $S = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$. Suppose $S \neq \emptyset$ and let $\nu > 0$ be the same constant throughout (SC), (RSI), (EB), (PL), and (QG). The following relationship holds*

$$\text{(SC)} \to \text{(RSI)} \to \text{(EB)} \equiv \text{(PL)} \to \text{(QG)}. \tag{5}$$

Furthermore, if the coefficient of (QG) satisfies $\mu_{\mathrm{q}} > \rho$ (including the function f is convex), then the following equivalence holds

$$\text{(RSI)} \equiv \text{(EB)} \equiv \text{(PL)} \equiv \text{(QG)}. \tag{6}$$

1. Error bound is closely related to *metric subregularity* at $x^\star$ for 0 (Artacho and Geoffroy, 2008, Def. 2.3): there exist a constant $a > 0$ and a set $\mathcal{U}$ containing $x^\star$ such that $\operatorname{dist}(x, (\partial f)^{-1}(0)) \leq a \cdot \operatorname{dist}(0, \partial f(x)), \forall x \in \mathcal{U}$.
2. To be consistent with the smooth case in Karimi et al. (2016), we call the property (PL) Polyak-Łojasiewicz, which is usually used for smooth functions. The property (PL) is actually a special case of the *Kurdyka-Łojasiewicz inequality* $\varphi'(f(x) - f^\star)\operatorname{dist}(0, \hat{\partial} f(x)) \geq 1$ with $\varphi(s) = cs^{1/2}$ and $c > 0$.

Note that if f is a convex function and also satisfies (QG) with $\mu_{\rm q} > 0$, then (6) holds naturally (since f is 0-weakly convex). Theorem 3.1 includes Theorem 2 of Karimi et al. (2016) and Theorem 1 of Zhang (2020) as a special case, in which only L-smooth functions are considered. It is easy to see that all L-smooth functions are also L-weakly convex. Even for differentiable functions, Theorem 3.1 is more general than Theorem 2 of Karimi et al. (2016) in the sense that 1) we require no Lipschitz constant L for gradients to ensure the equivalency among (EB), (PL), and (QG) in the convex case; 2) the condition $\mu_{\rm q} > \rho$ is new and does not mean that f is convex (see Appendix C.2 in Liao et al. (2023a) for an example).

The associated coefficients for different conditions and the proof details for Theorem 3.1 are provided in Table 1 and Appendix C.1 of Liao et al. (2023a) respectively. The proof of Theorem 3.1 relies heavily on the notion of *slope* defined in Drusvyatskiy et al. (2021), Ekeland's variational principle (Ekeland, 1974), and a technical result in Lemma 2.5 of Drusvyatskiy et al. (2015). Alternative proofs based on *subgradient flows* (Bolte et al., 2017) are also possible. Indeed, one key step in the proof of Theorem 2 in Karimi et al. (2016) is based on *gradient flows* for smooth functions, which is as a special case of *subgradient flows* for nonsmooth cases.

Remark 1 *The regularity conditions (EB), (QG) and (PL) have been discussed for different function classes in the literature. For the smooth case, we refer to Theorem 2 of Karimi et al. (2016) for a nice summary; also see Guille-Escuret et al. (2021) for related discussions. For nonsmooth convex functions, the equivalence between (EB) and (QG) has been recognized in Theorem 3.3 in Drusvyatskiy and Lewis (2018) and Theorem 3.3 in Artacho and Geoffroy (2008), and the equivalence between (PL) and (QG) is established in Theorem 5 of Bolte et al. (2017). Thus, (EB), (PL), and (QG) are equivalent for the class of nonsmooth convex functions (Ye et al., 2021, Proposition 2); see also Zhu et al. (2023) for a recent discussion. Our Theorem 3.1 extends these results to ρ-weakly convex functions. The most closely related work is Drusvyatskiy et al. (2021) which focuses on nonsmooth optimization using Taylor-like models. Indeed, we specialize the proof in Theorem 3.7, Proposition 3.8, and Corollary 5.7 of Drusvyatskiy et al. (2021) in our setting and prove the relationship in (5) and (6) directly using the slope technique. We note that the implication from (QG) to (EB)/(PL) is not true in general. Yet, with the condition $\mu_{\rm q} > \rho$ in Theorem 3.1, all four regularity conditions (RSI), (EB), (PL) and (QG) are equivalent.* □

We conclude this section with a few simple instances. In principle, all the five properties in Theorem 3.1 are generalizations of quadratic functions to non-quadratic, nonconvex, and even nonsmooth cases. For illustration, let us first consider the simplest quadratic function $f(x) = x^2$, which is convex and differentiable. It is clear that $\hat{\partial} f(x) = \{2x\}$ and $S = \{0\}$. It is also immediate to verify that (SC) holds with $0 < \mu_{\rm s} \leq 2$, (RSI) holds with $0 < \mu_{\rm r} \leq 2$, (EB) holds with $\mu_{\rm e} \geq 1/2$, (PL) holds with $0 < \mu_{\rm p} \leq 2$, and (QG) holds with $0 < \mu_{\rm q} \leq 2$. Consider another simple convex function $f(x) = x^2$, if $|x| \leq 1$, and $f(x) = \frac{1}{2}x^4 + \frac{1}{2}$ otherwise. All the five properties hold for this function, but it is not L-smooth globally. Let us move away from convex functions, and consider $f(x) = x^2 + 6\sin^2(x)$. It is clear that this function satisfies (QG) globally, however, there exist suboptimal stationary points and consequently (EB) and (PL) do not hold globally. Thus, in this case ($\nu = +\infty$), the relationship (5) is strict, and (QG) is more general than the other conditions. Finally, we consider a ρ-weakly convex function with a QG constant $\mu_{\rm q} > \rho$: $f(x) = -x^2 + 1$ if $-1 < x < -0.5$, and $f(x) = 3(x+1)^2$ otherwise. The function is not convex but 2-weakly convex with the QG constant $\mu_{\rm q} = 6 > 2 = \rho$. In this case, Theorem 3.1 guarantees that (RSI), (EB) and (PL) also hold (see Appendix C.2 in Liao et al. (2023a) for more details).

4. Proximal point method for convex optimization

In this section, we will utilize the regularity conditions in Section 3 to derive linear convergence of the classical proximal point method (PPM) (Rockafellar, 1976a) for convex (potentially nonsmooth) optimization. PPM is a conceptually simple algorithm, which has been historically used for guiding algorithm design and analysis, such as proximal bundle methods (Lemarechal et al., 1981) and augmented Lagrangian methods (Rockafellar, 1976a). It has recently found increasing applications in modern machine learning; see Drusvyatskiy (2017).

4.1. Proximal point method

Consider the optimization problem

$$f^\star = \min_{x\in\mathbb{R}^n} \; f(x), \tag{7}$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper closed convex function. Note that (7) is also an abstract model for constrained optimization, since given a closed convex set X, we can define $\bar{f}(x) = f(x)$ if $x \in X$, and $\bar{f}(x) = \infty$ otherwise. Let $S = \operatorname{argmin}_x \; f(x)$. We define the *proximal mapping* as

$$\operatorname{prox}_{\alpha f}(x_k) := \operatorname*{argmin}_{x\in\mathbb{R}^n} \; f(x) + \frac{1}{2\alpha}\left\|x - x_k\right\|^2, \tag{8}$$

where $\alpha > 0$. Starting with any initial point x_0, the PPM generates a sequence of points as follows

$$x_{k+1} = \operatorname{prox}_{c_k f}(x_k), \quad k = 0, 1, 2, \ldots \tag{9}$$

where $\{c_k\}_{k\geq 0}$ is a sequence of positive real numbers. The quadratic term in (8) makes the objective function strongly convex and always admits a unique solution. The iterates (9) are thus well-defined.

The convergence of PPM (9) for (nonsmooth) convex optimization has been studied since the 1970s (Rockafellar, 1976b). The sublinear convergence is relatively easy to establish, and many different assumptions exist for linear convergences of (9); see Rockafellar (1976b); Luque (1984); Leventhal (2009); Cui et al. (2016); Drusvyatskiy and Lewis (2018). However, as we will highlight later, some assumptions are restrictive and the corresponding proofs are sophisticated and nontransparent. We aim to provide simple proofs under the general regularity conditions in Section 3.

4.2. (Sub)linear convergences of PPM

Under a very general setup, the PPM (9) converges at a sublinear rate for cost value gaps, and the iterates converge asymptotically, as summarized in Theorem 4.1. This result is classical (Güler, 1991, Theorem 2.1), and a new bound with a constant 4 is available in Theorem 4.1 in Taylor et al. (2017) using the performance estimation technique.

Theorem 4.1 (Sublinear convergence (Güler, 1991, Theorem 2.1)) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function and $S \neq \emptyset$. Then, the iterates (9) with a positive sequence $\{c_k\}_{k\geq 0}$ satisfy*

$$f(x_k) - f^\star \leq \operatorname{dist}^2(x_0, S)/(2\textstyle\sum_{t=0}^{k-1} c_t). \tag{10}$$

If we further have $\lim_{k\to\infty} \sum_{t=0}^{k-1} c_t = \infty$, then the iterates converge to an optimal solution $\bar{x}$ asymptotically, i.e., $\lim_{k\to\infty} x_k = \bar{x}$, where $\bar{x} \in S$.

The proof of (10) is immediate from a telescope sum via the following one-step improvement:

$$2c_k(f(x_{k+1}) - f(x^\star)) \le \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2, \quad \forall c_k > 0, x^\star \in S. \tag{11}$$

This fact is not difficult to establish. For completeness, we provide proof details in Appendix D.1 in Liao et al. (2023a). Note that choosing any constant step size $c_k = c > 0$ in (10) directly implies the common sublinear rate $\mathcal{O}(1/k)$. In Theorem 4.1, f does not need to be L-smooth, and it can also be non-differentiable. Thus, the guarantees in Theorem 4.1 are much stronger than those by (sub)gradient methods. This is because the proximal mapping (8) is a stronger oracle than simple (sub)gradient updates.

Similar to GD in Section 2, when f satisfies additional regularity conditions, the PPM enjoys linear convergence. With the convexity assumption, our next main technical result establishes linear convergences of the PPM under the general regularity conditions in Theorem 3.1.

Theorem 4.2 (Linear convergence) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function, $S \neq \emptyset$, and $\nu > 0$. Suppose f satisfies (PL) (or (EB), (RSI), (QG)) over the sublevel set $[f \le f^\star + \nu]$. Then, for all $k \ge k_0$ steps, the iterates (9) with a positive sequence $\{c_k\}_{k\ge 0}$ bounded away from zero enjoy linear convergence rates, i.e.,*

$$f(x_{k+1}) - f^\star \le \omega_k \cdot (f(x_k) - f^\star), \tag{12a}$$
$$\operatorname{dist}(x_{k+1}, S) \le \theta_k \cdot \operatorname{dist}(x_k, S), \tag{12b}$$

where the constants are

$$\omega_k = \frac{1}{1 + c_k\mu_{\rm p}} < 1, \quad \theta_k = \min\left\{\frac{1}{\sqrt{c_k\mu_{\rm q} + 1}}, \frac{1}{\sqrt{c_k^2/\mu_{\rm e}^2 + 1}}\right\} < 1, \quad k_0 = \frac{\operatorname{dist}^2(x_0, S)}{2\nu \inf_{k\ge 0} c_k}.$$

Proof The sublinear convergence in Theorem 4.1 ensures that the iterate x_k reaches $[f \le f^\star + \nu]$ after at most k_0 iterations. Once x_k is within $[f \le f^\star + \nu]$, all the properties (EB), (PL), (RSI), and (QG) are equivalent by Theorem 3.1. For the analysis below, we assume $x_k \in [f \le f^\star + \nu]$.

We next show that (PL) gives a simple proof of (12a), and (QG) together with (EB) leads to a clean proof of (12b). Recall that the optimality condition of (9) directly implies

$$-(x_{k+1} - x_k)/c_k \in \partial f(x_{k+1}). \tag{13}$$

Then, the following inequalities hold

$$f(x_k) - f(x_{k+1}) \overset{(a)}{\ge} \frac{1}{2c_k}\|x_{k+1} - x_k\|^2 \overset{(b)}{\ge} \frac{c_k}{2}\operatorname{dist}^2(0, \partial f(x_{k+1})) \overset{(c)}{\ge} c_k\mu_{\rm p}(f(x_{k+1}) - f^\star), \tag{14}$$

where (a) applies the fact that x_k is a suboptimal solution to (9), (b) comes from the optimality (13), and (c) applies (PL). Re-arranging and subtracting f^* from both sides of (14) lead to the desired linear convergence result in (12a).

We next use (QG) to prove (12b) with coefficient $\theta_k \le 1/\sqrt{c_k\mu_{\rm q} + 1}$. By definition, we have $f(\Pi_S(x_k)) = f^\star$ and $\|\Pi_S(x_k) - x_k\|^2 = \operatorname{dist}^2(x_k, S)$. Since $f + \frac{1}{2c_k}\|\cdot - x_k\|^2$ is $1/c_k$ strongly convex, its first-order lower bound at x_{k+1} is

$$\begin{aligned} & f^\star + \frac{1}{2c_k}\|\Pi_S(x_k) - x_k\|^2 = f(\Pi_S(x_k)) + \frac{1}{2c_k}\operatorname{dist}^2(x_k, S) \\ \ge\ & f(x_{k+1}) + \frac{1}{2c_k}\|x_{k+1} - x_k\|^2 + \langle 0, \Pi_S(x_k) - x_{k+1}\rangle + \frac{1}{2c_k}\|\Pi_S(x_k) - x_{k+1}\|^2, \end{aligned} \tag{15}$$

where we also applied the fact that x_{k+1} minimizes (9) so 0 is a subgradient. From (15), we drop the positive term $\|x_{k+1} - x_k\|^2$ and use the fact that $\|\Pi_S(x_k) - x_{k+1}\| \geq \text{dist}(x_{k+1}, S)$, leading to

$$f^\star - f(x_{k+1}) + \text{dist}^2(x_k, S)/(2c_k) \geq \text{dist}^2(x_{k+1}, S)/(2c_k).$$

Combining this inequality with (QG) and simple re-arranging leads to the desired linear rate

$$\text{dist}^2(x_{k+1}, S) \leq 1/\sqrt{c_k\mu_{\text{q}} + 1} \cdot \text{dist}^2(x_k, S).$$

Simple arguments based on (EB) can establish (12b) with coefficient $\theta_k \leq (c_k^2/\mu_{\text{e}}^2 + 1)^{-1/2}$. We provide some details in Appendix D.2 in Liao et al. (2023a). This completes the proof. ■

It is possible to extend Theorem 4.2 to the class of weakly convex functions when a proper initialization is given. We provided this extension in Appendix D.3 of Liao et al. (2023a). Two nice features of Theorem 4.2 are 1) the simplicity of its proofs and 2) the generality of its conditions. Indeed, the proof of (12a) is simple via (PL), and the proof of (12b) is clean via (QG) and (EB), which are simpler than typical proofs. In addition, the regularity conditions are weaker than Rockafellar (1976b) and Luque (1984). Linear convergence for $\text{dist}(x_k, S)$ was first established in Rockafellar (1976b) with one restrictive assumption: the inverse of the subdifferential $(\partial f)^{-1}$ is locally Lipschitz at 0, which implies a unique optimal solution, i.e., S is a singleton. The uniqueness assumption is lifted in Luque (1984), which allows an unbounded solution set. More recently, this assumption is further relaxed to ∂f being metrically subregular in Cui et al. (2016) and Leventhal (2009). It is known that for convex functions, ∂f is metrically subregular if and only if f satisfies quadratic growth (cf. Theorem 3.1). Indeed, our proof in Theorem 4.2 is based on quadratic growth, which is a more intuitive geometrical property. Our main proof idea above is motivated by a result for the linear convergence of the spectral bundle method in Ding and Grimmer (2023); also see Liao et al. (2023b).

5. Inexact proximal point method (iPPM) and its convergence

In Section 4, each subproblem (8) is solved exactly. This may not be practical since one still needs an iterative solver to solve (8), where stopping criteria naturally introduce errors. We here discuss an inexact version of PPM (iPPM, Rockafellar, 1976b) where the subproblem (9) is solved inexactly. The regularity conditions in Theorem 3.1 also allow us to establish linear convergences of iPPM.

5.1. iPPM and stopping criteria

We replace the exact update (9) with an inexact update

$$x_{k+1} \approx \text{prox}_{c_k f}(x_k). \tag{16}$$

Two classical criteria suggested in Rockafellar's seminal work (Rockafellar, 1976b) are

$$\|x_{k+1} - \text{prox}_{c_k f}(x_k)\| \leq \epsilon_k, \quad \textstyle\sum_{k=0}^\infty \epsilon_k < \infty, \tag{A}$$

$$\|x_{k+1} - \text{prox}_{c_k f}(x_k)\| \leq \delta_k\|x_{k+1} - x_k\|, \quad \textstyle\sum_{k=0}^\infty \delta_k < \infty. \tag{B}$$

The inexact update (16) with (A) or (B) is called iPPM. The two criteria are not directly implementable as the value of $\text{prox}_{c_k f}(x_k)$ is unknown. As discussed in Proposition 3 of Rockafellar (1976b), two implementable alternatives that imply (A) and (B), respectively, are

$$\text{dist}(0, H_k(x_{k+1})) \leq \epsilon_k/c_k, \quad \textstyle\sum_{k=0}^\infty \epsilon_k < \infty, \tag{A$'$}$$

$$\text{dist}(0, H_k(x_{k+1})) \leq (\delta_k/c_k)\|x_{k+1} - x_k\|, \quad \textstyle\sum_{k=0}^\infty \delta_k < \infty, \tag{B$'$}$$

where $H_k(x) = \partial f(x) + (x - x_k)/c_k$ is the subdifferential of $f + \|\cdot - x_k\|^2/(2c_k)$ at x (since f is convex by assumption). Note that (A) and (B) only require the inexact update x_{k+1} to stay close enough to $\text{prox}_{c_k f}(x_k)$ with respect to the Euclidean distance, but they do not require the inexact update x_{k+1} to be within the domain of f, i.e., $f(x_{k+1})$ might be infinity. However, the stopping criteria (A′) and (B′) require that x_{k+1} is in the domain of f.

5.2. (Sub)linear convergence of iPPM

The seminal work (Rockafellar, 1976b) has established the asymptotic convergence of iterates for iPPM under a general setup. We state the results below whose proof is very technical.

Theorem 5.1 (Asymptotic convergence of iterates (Rockafellar, 1976b, Theorem 1)) *Consider a proper closed convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. Let $\{x_k\}_{k\geq 0}$ be any sequence generated by (16) under (A) with a positive sequence $\{c_k\}_{k\geq 0}$ bounded away from zero. Then, we have 1) the sequence $\{x_k\}_{k\geq 0}$ is bounded if and only if there exists a solution to $0 \in \partial f(x)$, i.e., $S \neq \emptyset$; 2) if $S \neq \emptyset$, the whole sequence $\{x_k\}_{k\geq 0}$ converges to an optimal point $x_\infty \in S$, i.e., $\lim_{k\to\infty} x_k = x_\infty$.*

We next establish a sublinear convergence of iPPM for cost value gaps. Our simple proof is based on the boundedness of the iterates from Theorem 5.1 and a recent idea in Theorem 3 in Lu and Yang (2023). We provide some details in Appendix E of Liao et al. (2023a).

Theorem 5.2 (Sublinear convergence of iPPM) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function, and $S \neq \emptyset$. The iterates (16) under (A′) with a positive sequence $\{c_k\}_{k\geq 0}$ bounded away from zero converge to $x^\star \in S$ asymptotically, and the cost value gaps converge as*

$$\min_{j=0,\ldots,k} f(x_j) - f^\star \leq (\text{dist}^2(x_0, S) + 2D\sum_{j=0}^{k-1} \epsilon_j)/(2\sum_{j=0}^{k-1} c_j),$$

where D is the diameter of the sequence $\{x_k\}_{k\geq 0}$ which is bounded.

Note that Theorems 5.1 and 5.2 can be viewed as the convergence counterpart for iPPM of Theorem 4.1 with two major differences: 1) Theorem 5.2 deals with the best iterate, unlike the last iterate in Theorem 4.1 (the guarantee for the average $\bar{x}_k = \frac{1}{k}\sum_{j=1}^k x_j$ or weighted average $\tilde{x}_k = (\sum_{j=0}^{k-1} c_j x_{j+1})/(\sum_{j=0}^{k-1} c_j)$ is also straightforward; see Remark 2 in Liao et al. (2023a); 2) the convergence of cost values in Theorem 5.2 relies on the boundedness of iterates in Theorem 5.1 whose proof is very technical (Rockafellar, 1976b, Theorem 1), while the convergence of iterates of exact PPM is established from the sublinear convergence of cost values in Theorem 4.1.

Similar to Theorem 4.2, the linear convergence of iPPM can also be established when suitable regularity conditions are assumed. Note that (SC), (RSI), (EB), (PL), and (QG) in Section 3 are all defined over a sublevel set $[f \leq f^\star + \nu]$ with $\nu > 0$, which is convenient to estimate the number of iterations to enter a certain region, such as the constant k_0 in Theorem 4.2; but this definition is too restrictive in the analysis of iPPM. Indeed, as the stopping criteria (A) and (B) allow infeasible points, it is possible that the sequence $\{x_k\}_{k\geq 0}$ might not be always feasible (i.e., $f(x_k) = +\infty$ for some $k \geq 0$) but the sequence $\{x_k\}_{k\geq 0}$ is approaching to the optimal solution set. In this case, the inequality $\mu_{\text{q}}/2 \cdot \text{dist}(x_k, S) \leq f(x_k) - f^\star$ with any $\mu_{\text{q}} > 0$ still holds automatically. To state the linear convergence result in a general case, we here modify the definition of (QG) from a sublevel set to a neighborhood of the optimal solution. Precisely, we say f satisfies QG, if there is a constant $\mu_{\text{q}} > 0$ and a neighborhood $\mathcal{U} \subseteq \mathbb{R}^n$ containing the optimal solution set S such that

$$\frac{\mu_{\text{q}}}{2} \cdot \text{dist}^2(x, S) \leq f(x) - f^\star, \quad \forall x \in \mathcal{U}. \tag{17}$$

Notice that we can also redefine RSI, EB, and PL using a neighborhood (similar to (17)) and show the equivalence between them. However, the neighborhood $\mathcal{U}$ for different regularity conditions may be different. For ease of presentation, we present our final technical result that shows the linear convergence of iPPM under (17), which is the counterpart of Theorem 4.2.

Theorem 5.3 (Linear convergence of iPPM) *Let $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a proper closed convex function. Suppose $S\neq\emptyset$ and f satisfies (17). Let $\{x_k\}$ be any sequence generated by iPPM (16) under (A) and (B) with parameters $\{c_k\}_{k\geq 0}$ bounded away from zero. Then, there exist a nonnegative $\theta_k<1$ and a large $\bar{k}>0$ such that for all $k\geq\bar{k}$, we have the linear convergence*

$$\operatorname{dist}(x_{k+1},S)\leq\hat{\theta}_k\operatorname{dist}(x_k,S),\quad \textit{where}\quad \hat{\theta}_k=\frac{\theta_k+2\delta_k}{1-\delta_k}<1.$$

We provide the proof of Theorem 5.3 in Appendix E.2 of Liao et al. (2023a). In Theorem 5.3, criterion (A) serves to guarantee that the iterates can reach the neighborhood $\mathcal{U}$ in (17) (cf. the asymptotic convergence from Theorem 5.1). If (17) holds globally (i.e., $\mathcal{U}=\mathbb{R}^n$), the same linear convergence result holds with (B) only. Note that our convergence proof for Theorem 5.3 in Liao et al. (2023a) is very modular, combining a key inequality in (Luque, 1984, Equation 2.7) with Theorem 4.2. Indeed, thanks to the regularity conditions (i.e., (17)), our proof is simpler and less conservative than typical proofs in the literature; see the discussions at the end of Section 4.

6. Applications

In this section, we consider three different applications of convex optimization in machine learning and signal processing: linear support vector machine (SVM) (Zhang and Lin (2015)), lasso (Tibshirani (1996)), and elastic-net (Zou and Hastie (2005)) respectively. For each application, we run the PPM on three different data sets. These problems satisfy the regularity conditions in Theorem 4.2. The numerical results, shown in Figure 1, confirm the linear convergence of the PPM. The details of the data set and the choices of parameters can be found in Appendix F in Liao et al. (2023a).

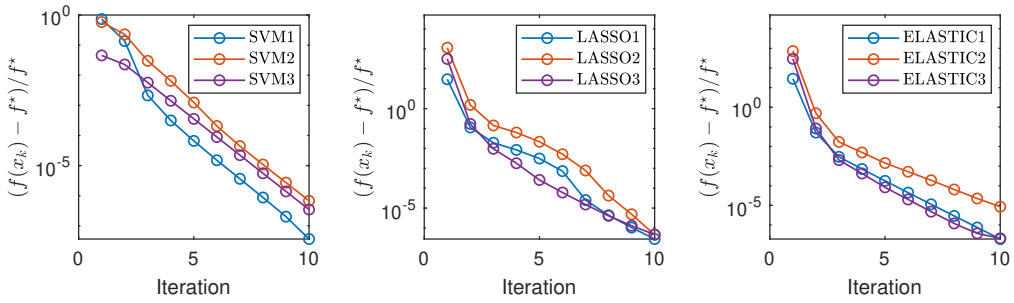

Figure 1: Linear convergences of cost value gaps for linear SVM, lasso, and elastic-net.

7. Conclusion

In this paper, we have established the relationship between different popular regularity conditions under the class of ρ-weakly convex functions. This result is beneficial in the analysis and design of various first-order algorithms. We have also presented simple and clear proofs for the (inexact) PPM which makes the analysis of the (inexact) PPM more accessible to new readers. We believe these results will facilitate algorithm developments in nonsmooth optimization. We are particularly interested in further applications in large-scale conic optimization (Zheng et al., 2021).

Acknowledgments

This work is supported in part by NSF ECCS-2154650, NSF CMMI-2320697, and NSF CAREER-2340713.

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 23, 2010.

FJ Aragón Artacho and Michel H Geoffroy. Characterization of metric regularity of subdifferentials. *Journal of Convex Analysis*, 15(2):365, 2008.

Felipe Atenas, Claudia Sagastizábal, Paulo JS Silva, and Mikhail Solodov. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33(1):89–115, 2023.

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

Ying Cui, Defeng Sun, and Kim-Chuan Toh. On the asymptotic superlinear convergence of the augmented Lagrangian method for semidefinite programming with multiple solutions. *arXiv preprint arXiv:1610.00875*, 2016.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Mateo Díaz and Benjamin Grimmer. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, 33(2):424–454, 2023.

Lijun Ding and Benjamin Grimmer. Revisiting spectral bundle methods: Primal-dual (sub) linear convergence rates. *SIAM Journal on Optimization*, 33(2):1305–1332, 2023.

Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

Dmitriy Drusvyatskiy and Damek Davis. Subgradient methods under weak convexity and tame geometry. *SIAG/OPT Views and News*, 28:1–10, 2020.

Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015.

Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. *Mathematical Programming*, 185:357–383, 2021.

Ivar Ekeland. On the variational principle. *Journal of Mathematical Analysis and Applications*, 47 (2):324–353, 1974.

Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2021.

Charles Guille-Escuret, Adam Ibrahim, Baptiste Goujaud, and Ioannis Mitliagkas. Gradient descent is optimal under lower restricted secant inequality and upper error bound. *Advances in Neural Information Processing Systems*, 35:24893–24904, 2022.

Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM journal on control and optimization*, 29(2):403–419, 1991.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

Claude Lemarechal, Jean-Jacques Strodiot, and André Bihain. On a bundle algorithm for nonsmooth optimization. In *Nonlinear programming 4*, pages 245–282. Elsevier, 1981.

D Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360(2):681–688, 2009.

Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020.

Feng-Yi Liao, Lijun Ding, and Yang Zheng. Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. *arXiv preprint arXiv:2312.16775*, 2023a.

Feng-Yi Liao, Lijun Ding, and Yang Zheng. An overview and comparison of spectral bundle methods for primal and dual semidefinite programs. *arXiv preprint arXiv:2307.07651*, 2023b.

Haihao Lu and Jinwen Yang. On a unified and simplified proof for the ergodic convergence rates of PPM, PDHG and ADMM. *arXiv preprint arXiv:2305.02165*, 2023.

Fernando Javier Luque. Asymptotic convergence analysis of the proximal point algorithm. *SIAM Journal on Control and Optimization*, 22(2):277–293, 1984.

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

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

R Tyrrell Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976a.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976b.

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, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Jane J Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. *Set-Valued and Variational Analysis*, pages 1–35, 2021.

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

Hui Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *Mathematical Programming*, 180(1-2):371–416, 2020.

Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Yuchen Zhang and Xiao Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, pages 353–361. PMLR, 2015.

Yang Zheng, Giovanni Fantuzzi, and Antonis Papachristodoulou. Chordal and factor-width decompositions for scalable semidefinite and polynomial optimization. *Annual Reviews in Control*, 52: 243–279, 2021.

Daoli Zhu, Lei Zhao, and Shuzhong Zhang. A unified analysis for the subgradient methods minimizing composite nonconvex, nonsmooth and non-Lipschitz functions. *arXiv preprint arXiv:2308.16362*, 2023.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.