Proceedings of Machine Learning Research vol vvv:1–26, 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-Łojasiewic 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-Łojasiewic 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-Lojasiewic* inequality or *Restricted secant inequality* (Polyak, 1963; Zhang and Yin, 2013; Karimi et al., 2016; 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. (2023). 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 classes: *weakly convex* (Davis and Drusvyatskiy, 2019). The 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 subregular) 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 range 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 notation 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 establishes 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 Appendices A to F.

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 $\min_x f(x)$, 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., Bubeck et al., 2015, Theorem 3.3. If the function $f(x)$ is strongly convex, GD achieves a global linear convergence (Bubeck et al., 2015, Theorem 3.10).

However, the assumption of strongly convexity is too strong. It is known that some alternative weaker assumptions are already sufficient for linear convergences. We here introduce two notions: *restricted secant inequality* (RSI) (Zhang and Yin, 2013), and *Polyak-Lojasiewicz* (PL) inequality (Polyak, 1963). A smooth function $f(x)$ satisfies RSI if there exists $\mu > 0$ such that

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

and it satisfies the PL inequality if there exists $\beta > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \beta(f(x) - f^\star), \ \forall x \in \mathbb{R}^n. \tag{3}$$

Note that both RSI (2) and PL (3) imply that any stationary point of $f(x)$ is a global minimum. However, they do not imply the uniqueness of stationary points or convexity of the function. One can think that RSI (2) (resp. PL (3)) requires that the gradient $\nabla f(x)$ grows faster than that of quadratic functions 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 Guille-Escuret et al., 2022, Proposition 1 and Zhang, 2020, Proposition 1. We summarize a simple version below.

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

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

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

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 particular, the RSI (2) leads to a quick proof of (4a), and the PL (3) allows for a simple proof of (4b). Indeed, 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 Karimi et al., 2016, Theorem 2, 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 $f(x) + \frac{\rho}{2}\|x\|^2$ 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 $f(x) : \mathbb{R}^n \to \overline{\mathbb{R}}$ 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., Li et al., 2020, Page 27):

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

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

Let the optimal solution set of $f(x)$ be $S = \operatorname{argmin}_x f(x)$, and we assume $S \neq \emptyset$. Let $\nu > 0$ and consider the following five regularity conditions:

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

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

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

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

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

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

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

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

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 be a proper closed ρ-weakly convex function. The following relationship holds*

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

Furthermore, if any of the following two conditions is satisfied

- *$f(x)$ is convex (i.e., $\rho = 0$),*

1. Error bound is closely related to *metrically 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-Łojasiewic, which is usually used for smooth functions. The property (PL) is actually a special case of *Kurdyka-Łojasiewicz inequality* $\varphi'(f(x) - f^\star)\operatorname{dist}(0, \hat{\partial} f(x)) \geq 1$ with $\varphi(s) = cs^{1/2}$ and $c > 0$.

- *the (QG) coefficient satisfies* $\mu_{\mathrm{q}} > \frac{\rho}{2}$,

then the following equivalence holds

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

Theorem 3.1 includes Karimi et al., 2016, Theorem 2 and Zhang, 2020, Theorem 1 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 smooth functions, Theorem 3.1 is more general than Karimi et al., 2016, Theorem 2 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 $\mu_{\mathrm{q}} > \rho/2$ is new and does not mean that $f(x)$ is convex (see Appendix C.2 for an example).

Our proof details for Theorem 3.1 are presented in Appendix C. 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 Drusvyatskiy et al., 2015, Lemma 2.5. Alternative proofs based *subgradient flows* (Bolte et al., 2017) are also possible. Indeed, one key step in the proof of Karimi et al., 2016, Theorem 2 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 Karimi et al., 2016, Theorem 2 for a nice summary. For nonsmooth convex functions, the equivalency between (EB) and (QG) has been recognized in Drusvyatskiy and Lewis, 2018, Theorem 3.3 and Artacho and Geoffroy, 2008, Theorem 3.3, and the equivalency between (PL) and (QG) is established in Bolte et al., 2017, Theorem 5. Thus, (EB), (PL), and (QG) are equivalent for the class of nonsmooth convex functions (Ye et al., 2021, Proposition 2); also see 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 Drusvyatskiy et al., 2021, Theo. 3.7, Prop. 3.8, and Coro. 5.7 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 one of the two conditions 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 $0 < \mu_{\mathrm{s}} \leq 1$, (RSI) holds with $0 < \mu_{\mathrm{r}} \leq 2$, (EB) holds with $\mu_{\mathrm{e}} \geq 1/2$, (PL) holds with $0 < \mu_{\mathrm{p}} \leq 4$, and (QG) holds with $0 < \mu_{\mathrm{q}} \leq 1$. 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, 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 $\mu_{\mathrm{q}} > \frac{\rho}{2}$: $f(x) = -x^2 + 1$ 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 $\mu_{\mathrm{q}} = 3 > 2/2 = 1$. In this case, Theorem 3.1 guarantees that (RSI), (EB) and (PL) also hold (see Appendix C.2 and Figure 2 for additional details in the appendix).

4. Proximal point method for convex optimization

In this section, we will utilize the regularity conditions in Section 3 to derive fast linear convergence guarantees 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), 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

$$f^\star = \min_x \ f(x), \tag{7}$$

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 $x \in X$, otherwise $\bar{f}(x) = \infty$. Let $S = \operatorname{argmin}_x \ f(x)$. We define the *proximal mapping* as

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

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 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 and clean 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 also available in Taylor et al., 2017, Theorem 4.1 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 $\lim_{k\to\infty} \sum_{t=0}^{k-1} c_t = \infty$, 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:

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

This fact (11) is not difficult to establish. For completeness, we provide proof details in Appendix D. 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, $f(x)$ does not need to be L-smooth, and it can also be nonsmooth. 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 $f(x)$ satisfies additional regularity conditions, the PPM enjoys linear convergence. 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, and $S \neq \emptyset$. 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 $\{c_k\}_{k\ge 0}$ 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

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

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

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

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 $\theta_k \le 1/\sqrt{2c_k\mu_{\rm q} + 1}$. By definition, we have

$$f(\Pi_S(x_k)) = f^\star, \quad \text{and} \quad \|\Pi_S(x_k) - x_k\|^2 = \operatorname{dist}^2(x_k, S).$$

Since $f(x) + \frac{1}{2c_k}\|x - x_k\|^2$ is $1/c_k$ strongly convex, its first-order lower bound at x_{k+1} is

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

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 $\|\Pi_S(x_k) - x_{k+1}\| \geq \operatorname{dist}(x_{k+1}, S)$, leading to

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

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

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

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

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 very simple via (PL), while the proof of (12b) is also clean via (QG) and (EB), which are simpler than typical proofs. In addition, the regularity conditions are weaker than Rockafellar (1976b); Luque (1984). Linear convergence for $\operatorname{dist}(x_k, S)$ was first established in Rockafellar (1976b) with one restrictive assumptions: the inverse of the subdifferential $(\partial f)^{-1}$ is locally Lipschitz at 0. This assumption in turn implies the optimal solution is unique, 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 as ∂f being metrically subregular in Cui et al. (2016); Leventhal (2009). It is known that for convex functions, ∂f is metrically subregular if and only if $f(x)$ satisfies quadratic growth (cf. Theorem 3.1). Indeed, our proof in Theorem 4.2 is based on quadratic growth, which is a more intuitive, geometric property. Our main idea in the proof above is motivated by a recent result for the linear convergence of the spectral bundle method in Ding and Grimmer (2023).

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

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

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

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

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

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

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

3, two implementable alternatives that imply (A) and (B), respectively, are

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

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

where $H_k(x) = \partial f(x) + (x - x_k)/c_k$ is the subdifferential of $f(x) + \|x - x_k\|^2/(2c_k)$ (since $f(x)$ is convex by assumption). Note that (A) and (B) only require the inexact update x_{k+1} to stay close enough to $\operatorname{prox}_{c_k,f}(x_k)$ 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 convergences 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 technically invovled.

Theorem 5.1 (Asymptotic convergence of iterates (Rockafellar, 1976b, Theo. 1)) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function. Consider any sequence $\{x_k\}_{k\ge 0}$ generated by (16) under (A) with a positive sequence $\{c_k\}_{k\ge 0}$ bounded away from zero. Then, we have 1) the sequence $\{x_k\}_{k\ge 0}$ 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 $\{x_k\}_{k\ge 0}$ converges to an optimal point $x_\infty \in S$ asymptotically, i.e., $\lim x_k = x_\infty$.*

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 Lu and Yang, 2023, Theorem 3. We provide some details in Appendix E.

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 $\{c_k\}_{k\ge 0}$ bounded away from zero converges to $x^\star \in S$ asymptotically, and the cost value gaps converge as*

$$\min_{j=0,\dots,k} f(x_j) - f^\star \le \frac{\operatorname{dist}^2(x_0, S) + 2D\sum_{j=0}^{k-1} \epsilon_j}{2\sum_{j=0}^{k-1} c_j},$$

where D is the diameter of iterates $\{x_k\}$ 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 the appendix); 2) the convergence of cost values in Theorem 5.2 relies on the boundedness of iterates in Theorem 5.1 whose proof is technically involved from 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.

We now introduce our final technical result, 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, and $S \neq \emptyset$. Suppose f satisfies (EB) (or (QG), (PL)) over the sublevel set $[f \le f^\star + \nu]$. Let $\{x_k\}$ be any sequence generated by iPPM (16) under (A) and (B) with parameters $\{c_k\}_{k\ge 0}$ bounded away from zero. There exists a nonnegative $\theta_k < 1$ and a large $\bar{k} > 0$ such that for all $k \ge \bar{k}$, we have*

$$\operatorname{dist}(x_{k+1}, S) \le \hat{\theta}_k \operatorname{dist}(x_k, S), \text{ where } \hat{\theta}_k = \frac{\theta_k + 2\delta_k}{1 - \delta_k} \text{ and } \lim_{k\to\infty} \hat{\theta}_k = \theta_k < 1. \tag{17}$$

Proof Thanks to (A), the asymptotic convergence of $\{x_k\}$ is guaranteed by Theorem 5.1. Thus, there exists a k_1 such that the iterate enters the QG region, i.e., (QG) holds. Theorem 4.2 implies

$$\mathrm{dist}(\mathrm{prox}_{c_k,f}(x_k), S) \leq \theta_k \mathrm{dist}(x_k, S), \ \forall k \geq k_1, \tag{18}$$

where $\theta_k = 1/\sqrt{2c_k\mu_{\mathrm{q}} + 1} < 1$. On the other hand, thanks to (B), there exists a $k_2 > 0$ such that

$$(1 - \delta_k)\mathrm{dist}(x_{k+1}, S) \leq 2\delta_k \mathrm{dist}(x_k, S) + \mathrm{dist}(\mathrm{prox}_{c_k,f}(x_k), S), \ \forall k \geq k_2. \tag{19}$$

This inequality (19) quantifies the quality between the next iterate x_{k+1} and the true proximal point $\mathrm{prox}_{c_k,f}(x_k)$, first appeared in Luque, 1984, Equation 2.7. We provide some details of (19) in Appendix E.2. Choosing $k \geq \bar{k} = \max\{k_1, k_2\}$ and combining (18) with (19) directly leads to the desired rate (17). This completes the proof. ■

In Theorem 5.3, criterion (A) is to guarantee that the iterates can reach the QG region (cf. the asymptotic convergence from Theorem 5.1). If (QG) holds globally (i.e., $\nu = \infty$), the same linear convergence result holds with only (B). Note that the convergence proof in Theorem 5.3 is very modular. Indeed, thanks to the regularity conditions (i.e., (QG)) in Theorem 3.1, 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 are shown in Figure 1, which confirms the fast linear convergence of the PPM. The details of the data set and the choices of parameters can be found in Appendix F.

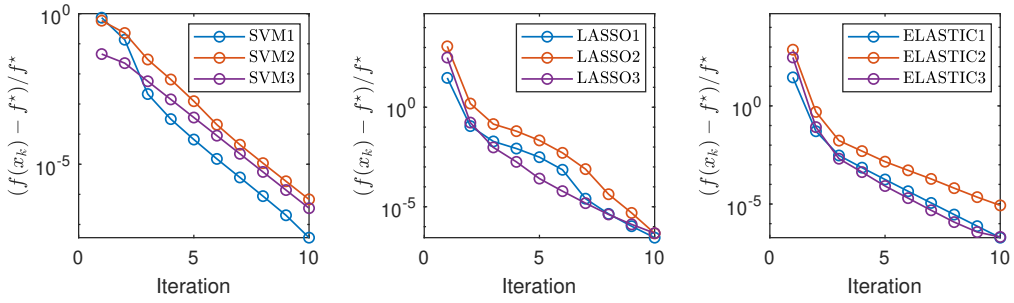

Figure 1: Linear convergences of cost value gaps for linear SVM (left), lasso (middle), and elastic-net (right).

7. Conclusion

In this paper, we have established the relationship between different favorite regularity conditions under the class of ρ-weakly convex functions. This result is beneficial in analysis and design of various first-order algorithm. 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 development 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 and in part by NSF CMMI-2320697.

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.

MOSEK ApS. *MOSEK Optimization Toolbox for MATLAB 10.1.20*, 2019. URL https://docs.mosek.com/latest/toolbox/index.html.

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.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

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.

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

Dmitriy Drusvyatskiy. Slope and geometry in variational mathematics. 2013.

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

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, 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.

Alexander D Ioffe. Variational analysis of regular mappings. *Springer Monographs in Mathematics. Springer, Cham*, 2017.

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. An overview and comparison of spectral bundle methods for primal and dual semidefinite programs. *arXiv preprint arXiv:2307.07651*, 2023.

J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. In *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

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.

Jean-Paul Penot. *Calculus without derivatives*, volume 266. Springer, 2013.

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.

R Tyrrell Rockafellar. Characterizing firm nonexpansiveness of prox mappings both locally and globally. *Journal of Nonlinear and convex Analysis*, 22(5), 2021.

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.

Appendix A. Subdifferential characterization of weakly convex functions

For self-completeness, we review a useful subdifferential characterization of weakly convex functions. Recall that $\overline{\mathbb{R}}$ denotes the extended real line, i.e., $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$.

We say that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper if $f(x) < \infty$ for at least one $x \in \mathbb{R}^n$ and $-\infty < f(x)$ for all $x \in \mathbb{R}^n$. We say that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed if the epigraph $\text{epi} f : \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$ is closed.

Definition 1 *For a closed function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, we define the Fréchet subdifferential as*

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

Definition 2 *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called ρ-weakly convex if $f(x) + \frac{\rho}{2}\|x\|^2$ is convex.*

For this class of ρ-weakly convex functions, the Fréchet subdifferential always exists for any point $x \in \mathbb{R}^n$. This is because the Fréchet subdifferential is the same as the Clarke subdifferential for ρ-weakly convex functions and the Clarke subdifferential always exists (Li et al., 2020, Section 3 and fact 5). The class of ρ-weakly convex functions also has the following nice characterizations. These characterizations offer favorable properties for ρ-weakly convex functions, which can be viewed as extensions from strongly convex functions.

Lemma 1 (Subdifferential characterization (Davis and Drusvyatskiy, 2019, Lemma 2.1)) *The following statements are equivalent for any closed function $f(x)$.*

- *The function f is ρ-weakly convex.*
- *The approximate secant inequality holds:*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) + \frac{\rho\lambda(1-\lambda)}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n, \lambda \in [0, 1].$$

- *The subgradient inequality holds:*

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n, v \in \hat{\partial} f(x). \tag{20}$$

- *The subdifferential map is hypomonotone:*

$$\langle v - w, x - y \rangle \geq -\rho\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n, v \in \hat{\partial} f(x), \text{ and } w \in \hat{\partial} f(y).$$

If f is $\mathcal{C}^2$-smooth, then the four properties above are all equivalent to

$$\nabla^2 f(x) \succeq -\rho I, \forall x \in \mathbb{R}^n.$$

The subgradient inequality (20) is particularly useful in our proof for Theorem 3.1 in Appendix C. We will repeatedly apply (20) to establish the connection between subdifferential and cost value gap.

Appendix B. Proof of Theorem 2.1

Recall that we say a differentialable function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if it satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \tag{21}$$

A differentialable L-smooth function f also satisfy the following useful inequalities

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{22}$$

Assume that RSI (2) holds with constant $\mu > 0$. We first note that $\mu \leq L$. Indeed, this can be seen from the following fact: let $x \in \mathbb{R}^n, x \notin S$, we have

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

where the first inequality uses RSI (2), and the last inequality applies Cauchy-Schwarz and (21).

Recall that for a closed set $S \subseteq \mathbb{R}^n$, we denote the distance of a point $x \in \mathbb{R}^n$ to S as $\operatorname{dist}(x, S) := \min_{y\in S} \|x - y\|$ and the projection of x onto S as $\Pi_S(x) = \operatorname{argmin}_{y\in S} \|x - y\|$. In the following, S denotes the set of minimizers of $f(x)$. We first the following inequality in terms of the distance to the solution set

$$\begin{aligned}\operatorname{dist}^2(x_{k+1}, S) &\leq \|x_k - t_k\nabla f(x_k) - \Pi_S(x_k)\|^2 && \text{(23a)}\\ &= \|x_k - \Pi_S(x_k)\|^2 - 2t_k\langle \nabla f(x_k), x_k - \Pi_S(x_k)\rangle + t_k^2\|\nabla f(x_k)\|^2 && \text{(23b)}\\ &\leq \operatorname{dist}^2(x_k, S) - 2t_k\mu\operatorname{dist}^2(x_k, S) + t_k^2L^2\operatorname{dist}^2(x_k, S) && \text{(23c)}\\ &= \left(1 - 2t_k\mu + t_k^2L^2\right)\operatorname{dist}^2(x_k, S), && \text{(23d)}\end{aligned}$$

where (23a) plugs in definition of x_{k+1} and uses the fact $\operatorname{dist}(x_{k+1}, S) = \|x_{k+1} - \Pi_S(x_{k+1})\| \leq \|x_{k+1} - \Pi_S(x_k)\|$, (23b) simply expends the square, and (23c) uses (2) to upper bound the inner product term and applies the L-smoothness inequality (21) to $\|\nabla f(x_k)\| = \|\nabla f(x_k) - \nabla f(\Pi_S(x_k))\| \leq L\|x_k - \Pi_S(x_k)\|$ (noting that $\nabla f(\Pi_S(x_k)) = 0$).

On the other hand, suppose the PL property (3) holds with constant $\beta > 0$. First, we have

$$\begin{aligned}f(x_{k+1}) - f(x_k) &\leq \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= -t_k\langle \nabla f(x_k), \nabla f(x_k)\rangle + \frac{Lt_k^2}{2}\|\nabla f(x_k)\|^2 \\ &= \left(\frac{-2t_k + Lt_k^2}{2}\right)\|\nabla f(x_k)\|^2.\end{aligned}$$

where the first inequality applies (22). If the step size t_k is chosen such that $0 < t_k < \frac{2}{L}$, then $\frac{-2t_k+Lt_k^2}{2} < 0$. Then, applying PL (3) to the above inequality and subtracting f^* from both sides, we arrive at the linear convergence in the cost value gap

$$f(x_{k+1}) - f^\star \leq (1 + \left(-2t_k + Lt_k^2\right)\beta)(f(x_k) - f^\star). \tag{24}$$

To summarize, if we choose the step size t_k such that the constants in (23d) and (24) less than one (equivalently, $0 \leq 1 - 2t_k\mu + t_k^2L^2 < 1$ and $0 < t_k < \frac{2}{L}$), linear convergence for both the distance and the cost value gap is guaranteed. In particular, choosing $t_k = \frac{\mu}{L^2}$, we have $1-2t_k\mu+t_k^2L^2 = 1-\frac{\mu}{L} < 1$ and $t_k \leq \frac{L}{L^2} < \frac{2}{L}$ as $\mu \leq L$, leading to the desired linear convergence:

$$\begin{aligned} \mathrm{dist}^2(x_{k+1}, S) &\leq (1 - \mu/L)\mathrm{dist}^2(x_k, S), \\ f(x_{k+1}) - f^\star &\leq \frac{L^3 + (-2\mu L + \mu^2)\beta}{L^3}(f(x_k) - f^\star). \end{aligned}$$

Note that if (3) holds with $\beta > 0$, then $\beta \leq \frac{L^3}{2\mu L - \mu^2}$. Otherwise, the above cost value decrease would contradict with the fact that the cost value gap is always nonnegative. This choice of step size $t_k = \frac{\mu}{L^2}$ is not new, which was used in (Guille-Escuret et al., 2022, Proposition 1) or implicitly (Zhang, 2020, Proposition 1).

Remark 1 *Note that Theorem 2.1 is consistent with the linear convergence result in (Karimi et al., 2016, Theorem 1) since the step size $t_k = \frac{1}{L}$ also satisfies $0 < t_k < \frac{2}{L}$. Despite choosing $t_k = \frac{1}{L}$ renders a cleaner reduction constant $(1 - \frac{\mu}{L})$ in (Karimi et al., 2016, Theorem 1), it is unclear if the distance also converges linearly under this stepsize choice (i.e., it is not guaranteed that $t_k = \frac{1}{L}$ leads to the constant $1 - 2t_k\mu + t_k^2L^2 = 2 - 2\frac{\mu}{L} < 1$).*

Appendix C. Details of Theorem 3.1

C.1. Proofs of Theorem 3.1

Our proofs are adapted from the techniques in (Drusvyatskiy et al., 2021, Theorem 3.7, Proposition 3.8, and Corollary 5.7). We first start by introducing the necessary ingredients to build up the proofs. We define the notion of *slope* as (Drusvyatskiy et al., 2021, Section 2).

Definition 3 (Slope) *Let $\mathcal{X}$ be a complete metric space with the metric $d(\cdot, \cdot)$. Consider a closed function $f : \mathcal{X} \to \overline{\mathbb{R}}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. The slope of f at $\bar{x}$ is its maximal instantaneous rate of decrease:*

$$|\nabla f|(\bar{x}) := \limsup_{x \to \bar{x}} \frac{(f(\bar{x}) - f(x))^+}{d(x, \bar{x})},$$

where we use the notation $r^+ = \max\{0, r\}$.

If the function f is smooth on a Euclidean space, the slope $|\nabla f|(\bar{x})$ is simply the norm of the gradient $\|\nabla f(\bar{x})\|$. If $f(x)$ is convex, the slope $|\nabla f|(\bar{x})$ is equivalent to the length of the minimal norm element in the subdifferential, i.e., $|\nabla f|(\bar{x}) = \mathrm{dist}(0, \partial f(\bar{x}))$. If it is ρ-weakly convex, the slope $|\nabla f|(\bar{x})$ is the same as the length of the minimal element in the Fréchet subdifferential, i.e., $|\nabla f|(\bar{x}) = \mathrm{dist}(0, \hat{\partial} f(\bar{x}))$ (see Ioffe (2017); Drusvyatskiy (2013) for more details). We summarize these useful properties into a lemma below:

Lemma 2 *Consider a closed function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\bar{x} \in \mathbb{R}^n$ with $f(\bar{x})$ finite. The following statements for the slope hold.*

1. *If f is smooth, then $|\nabla f|(\bar{x}) = \|\nabla f(\bar{x})\|$;*

2. *If f is convex, then* $|\nabla f|(\bar{x}) = \operatorname{dist}(0, \partial f(\bar{x}))$;

3. *If f is ρ-weakly convex, then* $|\nabla f|(\bar{x}) = \operatorname{dist}(0, \hat{\partial} f(\bar{x}))$.

Proof We note that the following inequality holds for any $\bar{x} \in \mathbb{R}^n$ with a finite function value $f(\bar{x})$ (Drusvyatskiy et al., 2021, Section 2)

$$|\nabla f|(\bar{x}) \leq \operatorname{dist}(0, \hat{\partial} f(\bar{x})). \tag{25}$$

As the slope lacks basic lower-semicontinuity properties, it is important to define the *limiting slope*

$$\overline{|\nabla f|}(\bar{x}) := \liminf_{x \to \bar{x}, f(x) \to f(\bar{x})} |\nabla f|(x). \tag{26}$$

The result from (Ioffe, 2017, Proposition 8.5) has shown that

$$\overline{|\nabla f|}(\bar{x}) = \operatorname{dist}(0, \partial f(\bar{x})), \tag{27}$$

where $\partial f(\bar{x})$ denotes the *limiting subdifferential* (defined in Equation 18 Li et al. (2020)) of f evaluated at $\bar{x}$. Then we claim that for the class of functions whose Fréchet subdifferential and limiting subdifferential coincide, we have $|\nabla f|(\bar{x}) = \operatorname{dist}(0, \hat{\partial}(\bar{x}))$. Indeed, suppose $\hat{\partial} f(\bar{x}) = \partial f(\bar{x})$, using (27) and (25), we have

$$\overline{|\nabla f|}(\bar{x}) = \operatorname{dist}(0, \partial f(\bar{x})) = \operatorname{dist}(0, \hat{\partial} f(\bar{x})) \geq |\nabla f|(\bar{x}).$$

On the other hand, $\overline{|\nabla f|}(\bar{x}) \leq |\nabla f|(\bar{x})$ always holds true by defintion (26). Thus, $|\nabla f|(\bar{x}) = \operatorname{dist}(0, \hat{\partial} f(\bar{x}))$. One sufficient condition for Fréchet subdifferential and limiting subdifferential being the same (i.e., $\hat{\partial} f(\bar{x}) = \partial f(\bar{x})$) is f being subdifferentially regular at $\bar{x}$ (Li et al., 2020, Definition 2 and fact 5). For smooth, convex, and weakly convex functions, they are subdifferentially regular at every point (Li et al., 2020, Page 24). Thus, the result follows. ■

We then introduce a key result from (Drusvyatskiy et al., 2015, Lemma 2.5) that will help us estimate the distance to the optimal solution set.

Theorem 2 (Lemma 2.5 Drusvyatskiy et al. (2015)) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed function. Suppose for some point $x \in \operatorname{dom}(f)$, there are constants $\alpha < f(x)$ and $r > 0, K > 0$ such that $f(x) - \alpha < Kr$ and*

$$|\nabla f|(u) \geq r, \qquad \forall u \in [\alpha < f \leq f(x)] \cap (\operatorname{dist}(u, x) \leq K).$$

Then the sublevel set $[f \leq \alpha]$ is nonempty and $\operatorname{dist}(x, [f \leq \alpha]) \leq (f(x) - \alpha)/r$.

Note that if $\alpha = f^\star$ in Theorem 2, we can estimate the distance $\operatorname{dist}(x, [f = f^\star])$.

Theorem 3 (Ekeland's variational principle Ekeland (1974)) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed function. Suppose that for some $\epsilon > 0$ and $z \in [f \leq \inf f + \epsilon]$. Then for any $\rho > 0$ there exists $y \in \mathbb{R}^n$ such that*

$$\operatorname{dist}(z, y) \leq \epsilon/\rho, \tag{28a}$$

$$f(y) \leq f(z), \tag{28b}$$

$$f(x) + \rho \cdot \operatorname{dist}(x, y) > f(y), \quad \forall x \in \mathbb{R}^n / \{y\}. \tag{28c}$$

Note that (28a) means that the distance between y and z is bounded, (28b) tells us that y is also in the sublevel set, and (28c) shows that y uniquely minimizes the function $f(\cdot) + \rho \cdot \mathrm{dist}(\cdot, y)$. Following from the definition of slope (Definition 3), one immediate consequence of Theorem 3 is that the slope of y is also bounded by ρ as

$$|\nabla f|(y) = \limsup_{x \to y} \frac{(f(y) - f(x))^+}{\mathrm{dist}(x, y)} \leq \limsup_{x \to y} \frac{\rho \cdot \mathrm{dist}(x, y)}{\mathrm{dist}(x, y)} = \rho. \tag{29}$$

We are ready to prove the main result in Theorem 3.1. In particular, Theorems 2 and 3 will play an important role in the proof of (PL) $\Rightarrow$ (EB) $\Rightarrow$ (QG).

Poof of Theorem 3.1 Recall that S denotes the set of minimizers of $f(x)$, which is non-empty.

- (SC) $\Rightarrow$ (RSI): Suppose f satisfies (SC) with constant $\mu_{\mathrm{s}} > 0$. For any $x \in [f \leq f^\star + \nu]$ and $g \in \partial f(x)$, we let $y = \Pi_S(x)$. It follows that

$$\begin{aligned} & f(x) + \langle g, \Pi_S(x) - x \rangle + \mu_{\mathrm{s}} \cdot \mathrm{dist}^2(x, S) \leq f(\Pi_S(x)) \\ \Longrightarrow \ & \mu_{\mathrm{s}} \cdot \mathrm{dist}^2(x, S) \leq f(\Pi_S(x)) - f(x) + \langle g, x - \Pi_S(x) \rangle \\ & \leq \langle g, x - \Pi_S(x) \rangle , \end{aligned}$$

  where the last inequality comes from the fact that $f(\Pi_S(x)) - f(x) \leq 0$. This inequality also implies that $\mu_{\mathrm{r}} \geq \mu_{\mathrm{s}}$.

- (RSI) $\Rightarrow$ (EB): Suppose f satisfies (RSI) with constant $\mu_{\mathrm{r}} > 0$. Let $x \in [f \leq f^\star + \nu]$ and g be the minimal norm element in $\hat{\partial} f(x)$. Then, by definition of (RSI), we have

$$\langle g, x - \Pi_S(x) \rangle \geq \mu_{\mathrm{r}} \cdot \mathrm{dist}^2(x, S).$$

  Applying Cauchy-Schwarz on the left side yields

$$\mathrm{dist}(0, \hat{\partial} f(x)) \geq \mu_{\mathrm{r}} \cdot \mathrm{dist}(x, S).$$

  This also implies that $\mu_{\mathrm{e}} \leq \frac{1}{\mu_{\mathrm{r}}}$.

- (EB) $\Rightarrow$ (PL): Suppose f satisfies (EB) with constant $\mu_{\mathrm{e}} > 0$. From the subdifferential property for ρ-weakly convex function (20), we have

$$f^\star \geq f(x) + \langle v, \Pi_S(x) - x \rangle - \frac{\rho}{2} \|\Pi_S(x) - x\|^2,$$

  where $v \in \hat{\partial} f(x)$. Choosing v as the minimal norm element of $\hat{\partial} f(x)$, we deduce

$$\begin{aligned} f(x) - f^\star & \leq \mathrm{dist}(0, \hat{\partial} f(x)) \mathrm{dist}(x, S) + \frac{\rho}{2} \mathrm{dist}^2(x, S) \\ & \leq \mu_{\mathrm{e}} \cdot \mathrm{dist}^2(0, \hat{\partial} f(x)) + \frac{\rho \mu_{\mathrm{e}}^2}{2} \mathrm{dist}^2(0, \hat{\partial} f(x)) \\ & = \left(\frac{2\mu_{\mathrm{e}} + \rho \mu_{\mathrm{e}}^2}{2} \right) \mathrm{dist}^2(0, \hat{\partial} f(x)). \end{aligned}$$

  Dividing both sides by $(2\mu_{\mathrm{e}} + \rho \mu_{\mathrm{e}}^2)/2$ gets the desired constant.

- (PL) $\Rightarrow$ (EB): Suppose f satisfies (PL) with constant $\mu_{\mathrm{p}} > 0$. As f is ρ-weakly convex, applying Lemma 2 on (PL) shows that

$$|\nabla f|^2(x) \geq \mu_{\mathrm{p}} \cdot (f(x) - f^\star), \quad \forall x \in [f \leq f^\star + \nu]. \tag{30}$$

Let $g(x) = (f(x) - f^\star)^{1/2}$. Note that from (30) and the chain rule of limiting subdifferential (Penot, 2013, Proposition 6.19), for all $\bar{x} \in [0 < g \leq \sqrt{\nu}]$, we have the following inequalities

$$\begin{aligned}
|\nabla g|(\bar{x}) &\overset{(a)}{\geq} \overline{|\nabla g|}(\bar{x}) \overset{(b)}{=} \operatorname{dist}(0, \partial g(\bar{x})) \overset{(c)}{\geq} \operatorname{dist}(0, \frac{1}{2}(f(\bar{x}) - f^\star)^{-1/2}\partial f(\bar{x})) \\
&\overset{(d)}{=} \operatorname{dist}(0, \frac{1}{2}(f(\bar{x}) - f^\star)^{-1/2}\hat{\partial} f(\bar{x})) = \frac{1}{2}(f(\bar{x}) - f^\star)^{-1/2}\operatorname{dist}(0, \hat{\partial} f(\bar{x})) \\
&\overset{(e)}{=} \frac{1}{2}(f(\bar{x}) - f^\star)^{-1/2}|\nabla f|(\bar{x}) \\
&\overset{(f)}{\geq} \frac{1}{2}(f(\bar{x}) - f^\star)^{-1/2} \cdot \sqrt{\mu_{\mathrm{p}}}(f(\bar{x}) - f^\star)^{1/2} \\
&= \frac{\sqrt{\mu_{\mathrm{p}}}}{2},
\end{aligned}$$

where (a) follows the definition (26), (b) comes from (27), (c) uses the fact that $\partial g \subseteq \frac{1}{2}(f(\bar{x}) - f^\star)^{-1/2}\partial f(\bar{x})$ (Penot, 2013, Proposition 6.19), (d) uses the fact that f is weakly convex so it is subdifferentially regular at every point, implying $\hat{\partial} f(\bar{x}) = \partial f(\bar{x})$ (Li et al., 2020, Page 24), (e) uses Lemma 2, and (f) comes from (30).

Fix a point $x \in [0 < g \leq \sqrt{\nu}]$. We choose $K > g(x)\frac{2}{\sqrt{\mu_{\mathrm{p}}}}$. It is obvious we have

$$\begin{aligned}
&|\nabla g|(u) \geq \frac{\sqrt{\mu_{\mathrm{p}}}}{2}, \quad \forall u \in [0 < g \leq g(x)] \cap (\operatorname{dist}(u, x) \leq K), \\
&g(x) < K\frac{\sqrt{\mu_{\mathrm{p}}}}{2}.
\end{aligned}$$

Choosing $r = \frac{\sqrt{\mu_{\mathrm{p}}}}{2}, \alpha = 0$ in Theorem 2, we have the following bounds

$$\begin{aligned}
\operatorname{dist}(x, S) = \operatorname{dist}(x, [g \leq 0]) &\leq \frac{2}{\sqrt{\mu_{\mathrm{p}}}}g(x) = \frac{2}{\sqrt{\mu_{\mathrm{p}}}}(f(x) - f^\star)^{1/2} \\
&\leq \frac{2}{\mu_{\mathrm{p}}}|\nabla f|(x) = \frac{2}{\mu_{\mathrm{p}}}\operatorname{dist}(0, \hat{\partial} f(x)), \quad \forall x \in [f \leq f^\star + \nu],
\end{aligned}$$

where the first inequality is from Theorem 2, the second inequality comes from (30), and the third inequality applies Lemma 2. This completes the proof of (PL) $\Rightarrow$ (EB).

- (EB) $\Rightarrow$ (QG): Suppose f satisfies (EB) with $\mu_{\mathrm{e}} > 0$. Fixing any $x \in [f^\star < f \leq f^\star + \nu]$, we have $x \in [f^\star < f \leq f^\star + \epsilon]$ with $\epsilon = f(x) - f^\star$. Choosing $\rho = \sqrt{\epsilon/\mu_{\mathrm{e}}}$, Theorem 3 and (29) ensure that there exists a y such that $y \in [f \leq f^\star + \epsilon]$ and $|\nabla f|(y) \leq \rho$. Thus,

$$\begin{aligned}
\mu_{\mathrm{e}}\rho \geq \mu_{\mathrm{e}} \cdot |\nabla f|(y) &\overset{(a)}{\geq} \operatorname{dist}(y, S) \\
&\overset{(b)}{\geq} \operatorname{dist}(x, S) - \operatorname{dist}(x, y) \\
&\overset{(c)}{\geq} \operatorname{dist}(x, S) - \epsilon/\rho,
\end{aligned}$$

where (a) uses Lemma 2 (i.e., we have $|\nabla f|(y) = \operatorname{dist}(0, \hat{\partial} f(y))$ for weakly convex functions) and (EB), (b) applies the triangle inequality for a point to a set, and (c) comes from (28a). Substituting $\epsilon = f(x) - f^\star$ and $\rho = \sqrt{\frac{\epsilon}{\mu_{\mathrm{e}}}}$ and rearranging the above inequality yields

$$2\sqrt{\mu_{\mathrm{e}}}(f(x) - f^\star)^{1/2} \geq \operatorname{dist}(x, S).$$

Squaring both sides completes the proof of (EB) $\Rightarrow$ (QG).

Summarizing the relationship above leads to the first part of Theorem 3.1 in (5). We now prove the second part in (6).

- (QG) with f being convex or $\mu_{\mathrm{q}} > \frac{\rho}{2} \Rightarrow$ (RSI): We prove the later as the former follows the exact argument by setting $\rho = 0$ (which means that f is convex).

  Let $x \in [f \leq f^\star + \nu]$ and $g \in \hat{\partial} f(x)$. From the assumption of (QG) and the property of subgradient of weakly convex function in (20), we have

  $$\mu_{\mathrm{q}} \cdot \operatorname{dist}^2(x, S) \leq f(x) - f^\star \leq \langle g, x - \Pi_S(x) \rangle + \frac{\rho}{2} \operatorname{dist}^2(x, S).$$

  Rearranging terms yields

  $$\left(\mu_{\mathrm{q}} - \frac{\rho}{2}\right) \cdot \operatorname{dist}^2(x, S) \leq \langle g, x - \Pi_S(x) \rangle .$$

This completes our proof of Theorem 3.1.

C.2. An example of ρ-weakly convex function with $\mu_{\mathrm{q}} > \frac{\rho}{2}$

We give a detailed examination of the example showing that even if (QG) holds with the constant $\mu_{\mathrm{q}} > \frac{\rho}{2}$, the function is still nonconvex while all the (RSI), (PL), (EB), and (QG) all hold true. Consider the function

$$f(x) = \begin{cases} -x^2 + 1 & \text{if } -1 < x < -0.5, \\ 3(x+1)^2 & \text{otherwise.} \end{cases}$$

It can be verified that the function is not convex (due to the part $-1 < x < -0.5$) but 2-weakly convex with a unique minimizer $x^\star = -1$ and $f^\star = 0$. See Figure 2 for an illustration of these properties. The Fréchet subdifferential can be calculated as

$$\hat{\partial} f(x) = \begin{cases} -2x & \text{if } -1 < x < -0.5, \\ 6(x+1) & \text{if } x < -1 \text{ or } x > -0.5, \\ [0, 2] & \text{if } x = -1, \\ [1, 3] & \text{if } x = -0.5. \end{cases}$$

The length of the minimal element is then

$$\operatorname{dist}(0, \hat{\partial} f(x)) = \begin{cases} |2x| & \text{if } -1 < x < -0.5, \\ |6(x+1)| & \text{if } x < -1 \text{ or } x > -0.5, \\ 0 & \text{if } x = -1, \\ 1 & \text{if } x = -0.5. \end{cases}$$

One can see that (QG) holds for $0 < \mu_{\mathrm{q}} \leq 3$, (EB) holds for $\mu_{\mathrm{e}} \geq 1/2$, and (PL) holds for $0 < \mu_{\mathrm{p}} \leq 4/3$. These can also be observed in Figure 2.

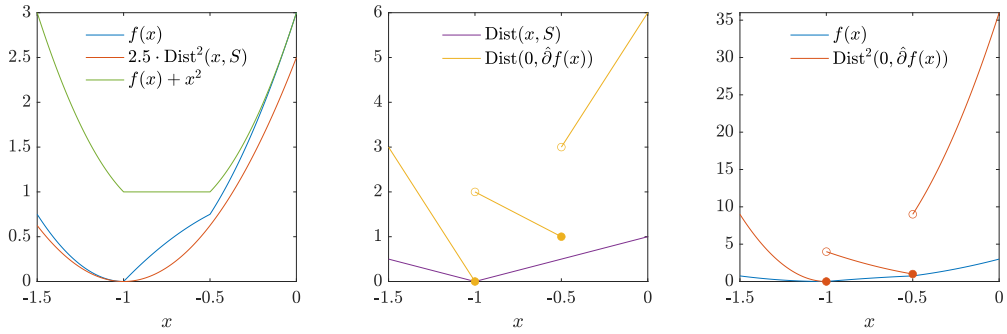

Figure 2: A nonconvex function with $f^\star = 0$ that satisfies the equivalency (RSI) $\equiv$ (EB) $\equiv$ (PL) $\equiv$ (QG). Left: the function f (blue line) is ρ-weakly convex (confirmed by the green line), and also satisfies (QG) with $\mu_{\mathrm{q}} = 2.5$ (confirmed by the red line); Middle: f satisfies (EB) as the value of $\mathrm{dist}(0, \hat{\partial} f(x))$ (yellow curve) is lower bounded by $\mathrm{dist}(x, S)$ (purple curve). Right: f satisfies (PL) as the value of $\mathrm{dist}^2(0, \hat{\partial} f(x))$ (red curve) is lower bounded by $f(x)$ (blue curve).

Appendix D. Convergence proofs of the PPM in Section 4.2

D.1. Proof for Theorem 4.1

One important step to prove the sublinear convergence Theorem 4.1 is to establish the one-step improvement (11). Specifically, from the optimality condition of (9), the iterate x_{k+1} satisfies

$$-1/c_k(x_{k+1} - x_k) \in \partial f(x_{k+1}). \tag{31}$$

By the definition of subdifferential for convex functions, we know

$$f(x^\star) \geq f(x_{k+1}) + \left\langle -\frac{1}{c_k}(x_{k+1} - x_k), x^\star - x_{k+1} \right\rangle. \tag{32}$$

On the other hand, we have

$$\begin{aligned} \|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star\|^2 - \|x_k - x_{k+1}\|^2 + 2\left\langle x_{k+1} - x_k, x_{k+1} - x^\star \right\rangle \\ &\leq \|x_k - x^\star\|^2 + 2\left\langle x_{k+1} - x_k, x_{k+1} - x^\star \right\rangle, \end{aligned} \tag{33}$$

where the first identity follows from simple algebraic manipulations $\|x_k - x^\star\|^2 = \|x_k - x_{k+1} + x_{k+1} - x^\star\|^2 = \|x_k - x_{k+1}\|^2 - 2\langle x_{k+1} - x_k, x_{k+1} - x^\star\rangle + \|x_{k+1} - x^\star\|^2$, and the second inequality drops a nonnegative term $\|x_k - x_{k+1}\|^2$. Combining (33) with (32) leads to

$$\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 - 2c_k(f(x_{k+1}) - f(x^\star)) \tag{34}$$

which is the same as (11). The rest of the proof follows from a standard telescope argument. Let $x^\star$ be a minimizer. Summing (34) for $t = 0, \ldots, k-1$, we have a telescope sum as follows

$$\sum_{t=0}^{k-1} c_t(f(x_{k+1}) - f(x^\star)) \leq \frac{1}{2}\sum_{t=0}^{k-1}(\|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2) \leq \frac{1}{2}\|x_0 - x^\star\|^2.$$

Considering the fact that the function values $\{f(x_k)\}_{k\geq 0}$ are nonincreasing, we deduce that

$$(f(x_k)-f(x^\star))\sum_{t=0}^{k-1}c_t \leq \sum_{t=0}^{k-1}c_t(f(x_{t+1})-f(x^\star)) \leq \frac{1}{2}\|x_0-x^\star\|^2.$$

Dividing both sides by $\sum_{t=0}^{k-1}c_t$ and taking $x^\star$ as the closed point to x_0 in S (since $x^\star$ can be any minimizer), we get the desired inequality in (10). Finally, we show that the sequence of iterates $\{x_k\}_{k\geq 0}$ converges to a minimizer. From (11), we know $\{\|x_k-x^\star\|\}_{k\geq 0}$ is nonincreasing and the sequence $\{x_k\}_{k\geq 0}$ is bounded. Hence, the sequence $\{x_k\}_{k\geq 0}$ has at least one accumulation point $\bar{x}$. Due to the convergence of $f(x_k)$, we know that $\bar{x}$ must be a minimizer, i.e., $f(\bar{x})=f^\star$ and $\bar{x}\in S$. We let $x^\star=\bar{x}$ in (11). This implies that $\|x_k-\bar{x}\|$ is nonincreasing and bounded by an infimum 0, thus $\lim_{k\to\infty}\|x_k-\bar{x}\|=0$, i.e., the whole sequence of iterates x_k converge to $\bar{x}$ asymptotically.

D.2. A compelete proof for Theorem 4.2

To finish the proof for (12b) with coefficient $\theta_k \leq 1/\sqrt{\mu_{\mathrm{e}}^2/c_k^2+1}$. We first review a crucial but standard result about the proximal mapping.

Lemma 3 ((Leventhal, 2009, Equation 3.1)) *For any $c_k>0$ in (9), the following holds true*

$$\mathrm{dist}^2(x_k,S)\geq \|x_{k+1}-\Pi_S(x_k)\|^2+\|x_k-x_{k+1}\|^2. \tag{35}$$

Proof As shown in Rockafellar (2021), the proximal point operator for a convex function $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ with $\alpha>0$ is *firmly nonexpensive*, i.e.,

$$\|\mathrm{prox}_{\alpha,f}(x)-\mathrm{prox}_{\alpha,f}(y)\|^2 \leq \left\langle x-y,\mathrm{prox}_{\alpha,f}(x)-\mathrm{prox}_{\alpha,f}(y)\right\rangle, \quad \forall x,y\in\mathbb{R}^n. \tag{36}$$

For notational convenience, let $P(x):=\mathrm{prox}_{\alpha,f}(x)$. We have

$$\begin{aligned}
\|x-y\|^2 &= \|P(x)+(x-P(x))-(P(y)+(y-P(y)))\|^2\\
&= \|P(x)-P(y)+(x-P(x))-(y-P(y))\|^2\\
&= \|P(x)-P(y)\|^2+\|(x-P(x))-(y-P(y))\|^2\\
&\quad +2\left\langle P(x)-P(y),(x-P(x))-(y-P(y))\right\rangle\\
&\geq \|P(x)-P(y)\|^2+\|(x-P(x))-(y-P(y))\|^2,
\end{aligned}$$

where the last inequality comes from the fact that the inner product is nonnegative by (36).

Setting $x=x_k$ and $y=\Pi_S(x_k)$, we have $x_{k+1}=P(x_k)$, and $y=P(y)$, which leads to the desired result in (35). ■

With Lemma 3, it follows that

$$\begin{aligned}
\mathrm{dist}^2(x_{k+1},S) &\leq \|x_{k+1}-\Pi_S(x_k)\|^2\\
&\leq \|x_k-\Pi_S(x_k)\|^2-\|x_k-x_{k+1}\|^2\\
&\leq \mathrm{dist}^2(x_k,S)-c_k^2\mathrm{dist}^2(0,\partial f(x_{k+1}))\\
&\leq \mathrm{dist}^2(x_k,S)-\frac{c_k^2}{\mu_{\mathrm{e}}^2}\mathrm{dist}(x_{k+1},S),
\end{aligned}$$

where the second inequality is from (35), the third inequality comes from (13), and the last inequality uses (EB). Rearranging terms, we have

$$\operatorname{dist}^2(x_{k+1}, S) \leq \frac{\mu_{\mathrm{e}}^2}{c_k^2 + \mu_{\mathrm{e}}^2} \operatorname{dist}^2(x_k, S).$$

Taking a square root yields the desired result.

D.3. Weakly convex extension

The linear convergence result in Theorem 4.2 can be extended to the class of ρ-weakly convex functions with proper initialization when the regularity conditions in Theorem 3.1 hold. As shown in theorem below, the coefficient $\{c_k\}$ is implicitly related to the weakly convex constant ρ.

Theorem D.1 (Extension to weakly convex function) *Suppose the function f in (7) is ρ-weakly convex. Let S be the solution set of (7) and assume $S \neq \emptyset$. Suppose f satisfies (QG) with $\mu_{\mathrm{q}} > \frac{\rho}{2} > 0$ over the sublevel set $[f \leq f^\star + \nu]$ with $\nu > 0$. The iterates generated by the PPM (9) with any positive sequence $\{c_k\}$ satisfying $\frac{1}{c_k} > \rho, \forall k \geq 1$, and initialization $x_0 \in [f \leq f^\star + \nu]$ satisfy (12a) and (12b) with constants*

$$\omega_k = \frac{2}{2 + \mu_{\mathrm{p}} c_k} < 1 \quad \textit{and} \quad \min\left\{ \frac{1}{\sqrt{2c_k\beta + 1}}, \frac{1}{\sqrt{\mu_{\mathrm{e}}^2/c_k^2 + 1}} \right\} < 1, \ \textit{where } \beta = \mu_{\mathrm{q}} - \frac{\rho}{2} > 0.$$

Proof The proof of the linear decrease in the cost value gap is identical to that in Theorem 4.2. As $\mu_{\mathrm{q}} > \frac{\rho}{2}$, (EB), (RSI), and (PL) also hold by Theorem 3.1. For the proof of the linear decrease in the distance , the coefficient $\frac{1}{\sqrt{\mu_{\mathrm{e}}^2/c_k^2+1}}$ follows the same argument as in the convex case since c_k is chosen such that $1/c_k > \rho$ so that the function $f(x) + \frac{1}{2c_k}\|x - x_k\|^2$ is a convex function plus a quadratic term and the proximal mapping is still firmly inexpensive; For the coefficient $\frac{1}{\sqrt{2c_k\beta+1}}$, the argument is essential the same as the convex case with a small modification that the subproblem is $(\frac{1}{c_k} - \rho)$-strongly convex. ■

Appendix E. Convergence proofs for the inexact PPM in Section 5

E.1. Proofs for Theorem 5.2

The optimality condition of (16) and criteria (A′) implies that there exist $v_k \in \partial f(x_{k+1})$ and $\theta_k \in \mathbb{R}^n$ such that

$$0 = v_k + \frac{x_{k+1} - x_k + \theta_k}{c_k}, \quad \|\theta_k\| \leq \epsilon_k.$$

Following the same argument in Appendix D.1, we can get a similar version of (11) as

$$\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 - 2c_k(f(x_{k+1}) - f(x^\star)) + 2\epsilon_k\|x^\star - x_{k+1}\|.$$

This improvement holds for all $k \geq 0$. Taking a sum of the above inequality from 0 to $k-1$, we get

$$\begin{aligned} &\|x_k - x^\star\|^2 \leq \|x_0 - x^\star\|^2 - 2\sum_{j=0}^{k-1} c_j(f(x_{j+1}) - f(x^\star)) + \sum_{j=0}^{k-1} 2\epsilon_j\|x^\star - x_{j+1}\| \\ \Longrightarrow\ & 2\sum_{j=0}^{k-1} c_j \min_{j=0,\ldots,k} f(x_j) - f(x^\star) \leq 2\sum_{j=0}^{k-1} c_j(f(x_{j+1}) - f(x^\star)) \leq \|x_0 - x^\star\|^2 + 2D\sum_{j=0}^{k-1}\epsilon_j. \end{aligned}$$

Dividing both sides by $2\sum_{j=0}^{k-1} c_j$ and letting $x^\star = \Pi_S(x^0)$ gets the desired inequality. The convergence of the iterate $\{x_k\}$ and the boundedness of the diameter D is guaranteed by Theorem 5.1.

Remark 2 *Theorem 5.2 provides a convergence guarantee for the best iterate as opposed to that for 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 the weighted average* $\tilde{x}_k = \frac{\sum_{j=0}^{k-1} c_j x_{j+1}}{\sum_{j=0}^{k-1} c_j}$ *can also be shown using the inequality*

$$f(\bar{x}_k) - f^\star = f\left(\sum_{j=1}^{k}\frac{1}{k}x_j\right) - \frac{1}{k}\sum_{j=1}^{k} f^\star \leq \frac{1}{k}\sum_{j=0}^{k-1}(f(x_{j+1}) - f^\star)$$

and

$$f(\tilde{x}_k) - f^\star \leq \frac{1}{\sum_{j=0}^{k-1} c_j}\sum_{j=0}^{k-1} c_j(f(x_{j+1}) - f^\star).$$

The proof in Theorem 5.2 borrows some proof idea from (Lu and Yang, 2023, Theorem 3) with the fact that (A′) implies the existence of $\theta_k \in \mathbb{R}^n$ *such that* $\mathrm{dist}(0, H_k(x_{k+1})) = \|\theta_k\|/c_k \leq \epsilon_k$.

E.2. Proofs for (19)

We here discuss some details for the useful inequality (19). Since this inequality plays an important role in the proof of Theorem 5.3, we put it into a lemma below.

Lemma 4 (Inexact one-step improvement (Luque, 1984, Equation 2.7)) *Let S be the solution set of (7) and assume $S \neq \emptyset$. Let $\{x_k\}$ be any sequence generated by the inexact PPM (16) under (B) with parameters $\{c_k\}_{k\geq 0}$ bounded away from zero. Then there exists a $\hat{k} \geq 0$ such that $\hat{k}$ satisfies $\delta_k < 1, \forall k \geq \hat{k}$ and*

$$(1-\delta_k)\mathrm{dist}(x_{k+1}, S) \leq 2\delta_k\mathrm{dist}(x_k, S) + \mathrm{dist}(\mathrm{prox}_{c_k,f}(x_k), S),\ \forall k \geq \hat{k}, \tag{37}$$

Proof Fix a $x \in \mathbb{R}^n$. From the update (9), we know

$$\begin{aligned} & f(\mathrm{prox}_{c_k,f}(x)) + \frac{1}{2c_k}\|\mathrm{prox}_{c_k,f}(x) - x\|^2 \leq f(\Pi_S(x)) + \frac{1}{2c_k}\|\Pi_S(x) - x\|^2 \\ \Longrightarrow\ & \frac{1}{2c_k}\|\mathrm{prox}_{c_k,f}(x) - x\|^2 \leq f^\star - f(\mathrm{prox}_{c_k,f}(x)) + \frac{1}{2c_k}\|\Pi_S(x) - x\|^2 \\ \Longrightarrow\ & \|\mathrm{prox}_{c_k,f}(x) - x\| \leq \mathrm{dist}(x, S), \end{aligned}$$

where the last implication is due to $f^\star - f(\text{prox}_{c_k,f}(x)) \leq 0$. On the other hand, using triangle inequality, the nonexpensiveness of projection onto a convex set, and the above inequality, we have

$$\begin{aligned}\forall x \in \mathbb{R}^n,\ \|x - \Pi_S(\text{prox}_{c_k,f}(x))\| &\leq \|x - \Pi_S(x)\| + \|\Pi_S(x) - \Pi_S(\text{prox}_{c_k,f}(x))\| \\ &\leq \text{dist}(x,S) + \|x - \text{prox}_{c_k,f}(x)\| \\ &\leq 2\text{dist}(x,S). \end{aligned} \tag{38}$$

Also,

$$\begin{aligned}&\|x_{k+1} - \Pi_S(\text{prox}_{c_k,f}(x))\| \\ &\leq \|x_{k+1} - \text{prox}_{c_k,f}(x_k)\| + \|\text{prox}_{c_k,f}(x_k) - \Pi_S(\text{prox}_{c_k,f}(x))\| \\ &\leq \delta_k\|x_{k+1} - x_k\| + \text{dist}(\text{prox}_{c_k,f}(x_k), S) \\ &\leq \delta_k\|x_{k+1} - \Pi_S(\text{prox}_{c_k,f}(x))\| + \delta_k\|x_k - \Pi_S(\text{prox}_{c_k,f}(x))\| + \text{dist}(\text{prox}_{c_k,f}(x_k), S),\end{aligned}$$

where the second inequality is from the stopping criterion (B). As $\delta_k \to 0$, choosing $\hat{k}$ such that $\delta_k < 1, \forall k \geq \hat{k}$, we have

$$(1-\delta_k)\|x_{k+1} - \Pi_S(\text{prox}_{c_k,f}(x))\| \leq \delta_k\|x_k - \Pi_S(\text{prox}_{c_k,f}(x))\| + \text{dist}(\text{prox}_{c_k,f}(x_k), S).$$

Finally, using the fact $\text{dist}(x_{k+1}, S) \leq \|x_{k+1} - \Pi_S(\text{prox}_{c_k,f}(x))\|$ and (38) for $x = x_k$ completes the proof. ■

Appendix F. Details of numerical experiments

In Section 6, we consider three applications of convex optimization in machine learning and signal processing, including linear support vector machine (SVM) (Zhang and Lin (2015)), Lasso (Tibshirani (1996)), and Elastic-Net (Zou and Hastie (2005)). The problem formulations are listed as

- Linear SVM:
$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - b_i(a_i^\top x)\} + \frac{\rho}{2}\|x\|^2,$$
where $\rho > 0, a_i \in \mathbb{R}^d, b_i \in \{-1, 1\}, i = 1, \ldots, n$.
- Lasso (ℓ_1-regularization):
$$\min_{x\in\mathbb{R}^m} \quad \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1,$$
where $A \in \mathbb{R}^{n\times m}, y \in \mathbb{R}^m$, and $\lambda > 0$.
- Elastic-Net ($\ell_1 - \ell_2^2$-regularization):
$$\min_{x\in\mathbb{R}^m} \quad \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 + \frac{\mu}{2}\|x\|^2,$$
where $A \in \mathbb{R}^{n\times m}, y \in \mathbb{R}^m$, and $\lambda, \mu > 0$.

For linear SVM, we consider three datasets ("a1a", "Australian", "fourclass") from Chang and Lin (2011). The regularization term is set as $\rho = 1$. The step size parameter of the PPM is chosen as $c_k = 1$ for all $k \geq 1$. For Lasso, we randomly generate $A \in \mathbb{R}^{n\times m}$ and $\hat{x} \in \mathbb{R}^m$ with s elements being zero ($s < m$), then construct the vector $y = A\hat{x}$. We consider problems with three different sizes: $(n = 10, m = 40, s = 5), (n = 20, m = 50, s = 10)$, and $(n = 30, m = 60, s = 15)$. The regularization terms are set as $\rho = 10$. The step size parameter of the PPM is chosen as $c_k = 0.16$ for all $k \geq 1$. For Elastic-Net, we consider the same data in Lasso with the regularization terms $\rho = 10$ and $\mu = 1$.

We use the modeling package Yalmip (Löfberg, 2004) to formulate the subproblem (9) and call the conic solver Mosek (ApS, 2019) to solve it. Our code is available at

https://github.com/soc-ucsd/PPM_examples