Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Convergence Analysis of the Subgradient Method Beyond Weak Convexity

Shunki Arata

arata-shunki147@g.ecc.u-tokyo.ac.jp

Graduate School of Information Science and Technology, The University of Tokyo

Research Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11478381/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Convergence Analysis of the Subgradient Method Beyond Weak Convexity

Shunki Arata[1]

[1]Department of Creative Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan.

Contributing authors: arata-shunki147@g.ecc.u-tokyo.ac.jp;

Abstract

The subgradient method and its variants are widely used to solve non-smooth minimization problems in machine learning and frequently achieve strong performance in practice. Nevertheless, from a theoretical viewpoint, non-asymptotic convergence analyses of the subgradient method typically rely on weak convexity, a condition that is often violated in realistic machine learning applications. To bridge this gap, we introduce a class of functions that strictly contains the class of weakly convex functions and establish convergence guarantees for the subgradient method within this broader setting. Building on these theoretical results, we reveal two key insights: (i) a polynomially decaying step size is a robust choice for ensuring convergence; (ii) when the loss function exhibits additional structure, such as local smoothness or sharpness, a polynomially decaying step size can be suboptimal compared to constant or geometrically decaying step sizes. These theoretical findings align with empirical observations from prior work and are further validated by our numerical experiments.

Overall, our results expand the theoretical applicability of first-order optimization algorithms to a wider class of non-smooth loss functions, bringing mathematical guarantees closer to practical machine learning scenarios.

Keywords: Non-smooth optimization, Subgradient method, Weak convexity, Step-size scheduling

1 Introduction

We study the subgradient method for minimizing non-smooth but locally Lipschitz loss functions. Such problems naturally arise in modern machine learning due to the use of non-smooth activation functions. To compute a subgradient of the loss function, automatic differentiation [1] is commonly used even when the function is not differentiable. It is known that, under mild assumptions, automatic differentiation returns an element of the Clarke subdifferential of the loss function [2].

Definition 1 (Clarke Subdifferential [3]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz. The Clarke subdifferential of f at $x \in \mathbb{R}^d$ is defined as*

$$\partial f(x) = \operatorname{conv}\left\{ \lim_{k\to\infty} \nabla f(x_k) \;\middle|\; x_k \to x, \;\; f \text{ is differentiable at } x_k \right\}.$$

Consequently, the subgradient method implemented in machine learning libraries such as PyTorch [4] and TensorFlow [5] for minimizing a loss function $f : \mathbb{R}^d \to \mathbb{R}$ can be formulated as

$$x_{t+1} = x_t - \eta_t g_t, \qquad g_t \in \partial f(x_t), \tag{1}$$

where $x_0 \in \mathbb{R}^d$ is an initial point and $(\eta_t)_{t\geq 0}$ is a positive step size sequence.

Classical convergence analyses of the subgradient method typically assume that the loss function is convex [6, 7]. More recent studies [8–12] have relaxed this requirement to weak convexity—that is, requiring $x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2$ to be convex for some $\rho > 0$. Nevertheless, weak convexity remains an overly restrictive condition for many loss functions in machine learning, such as mean squared error losses in deep neural networks [8, 11, 13]. This observation reveals a fundamental gap between existing convergence theory and the practical effectiveness of subgradient methods.

To establish a theoretical foundation for the subgradient method beyond weak convexity, we introduce a new class of loss functions characterized by the existence of a smooth surrogate in Definition 2. Importantly, this function class strictly contains the class of weakly convex functions.

Definition 2 (Smooth Surrogate). *Let $k \in [1, 2]$ and $C, L > 0$. A locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ is said to admit a (k, C, L)-smooth surrogate $\phi : \mathbb{R}^d \to \mathbb{R}$ if ϕ is an L-smooth function satisfying*

$$\langle g_x, \nabla\phi(x)\rangle \geq C\|\nabla\phi(x)\|^k, \qquad \forall x \in \mathbb{R}^d, \; \forall g_x \in \partial f(x) \tag{2}$$

and

$$\nabla\phi(x) = 0 \implies 0 \in \partial f(x). \tag{3}$$

Table 1 Comparison of the convergence results established in this paper. Each theorem assumes that the loss function f admits a smooth surrogate in the sense of Definition 2. The structural assumptions listed in the table are those unique to each theorem. Please refer to the individual theorems for the precise assumptions.

Theorem	Structural assumption	Step size η_t	Convergence result
2	-	$\eta/(1+t)$	$\lim_{T\to\infty} \|\nabla\phi(x_T)\| = 0$
4	locally smooth	η	$f(x_T) - f^* = \mathcal{O}(q^T),\ q \in (0,1)$
6	locally sharp	$\eta\gamma^{\lceil \frac{t-N+1}{N} \rceil}$	$f(x_T) - f^* = \mathcal{O}(\gamma^{T/N})$

Condition (2) requires the gradient of the smooth surrogate ϕ to be positively aligned with every element in the Clarke subdifferential of f. Condition (3) further requires ϕ to preserve the stationary points of f. Consequently, although the objective function f itself may be non-smooth and non-convex, the progress of the subgradient method can be quantified through this smooth surrogate. This perspective generalizes the analytical framework for weakly convex optimization established in [8, 12], in which the Moreau envelope [14] serves as a smooth surrogate. Our definition abstracts the essential properties required for convergence analysis, thereby extending the applicability of this framework beyond weakly convex functions.

For this broader class, we establish several convergence guarantees for the subgradient method, three of which are summarized in Table 1. The locally smooth setting is characterized by the condition

$$\sup_{g\in\partial f(x)} \|g\|^2 \le \theta\|\nabla\phi(x)\|^2$$

for some $\theta > 0$. Since ϕ is L-smooth, this condition implies that every element in the Clarke subdifferential of f vanishes as x approaches a minimizer of ϕ. Conversely, the locally sharp setting is characterized by the condition (2) with $k = 1$. This condition implies that $\sup_{g\in\partial f(x)} \|g\|$ remains bounded away from zero whenever $\nabla\phi(x) \neq 0$.

Taken together, these results provide a unified perspective on step-size selection:

(i) Because Theorem 2 does not require any structural assumptions, a polynomially decaying step size is a robust choice for ensuring convergence across the considered class of functions.

(ii) When additional local smoothness or sharpness properties are present, Theorems 4 and 6 establish linear convergence in the optimality gap of f. Specifically, in both settings, there exist instances where a polynomially decaying step size fails to achieve linear convergence.

(iii) Theorem 6 provides theoretical insight into the empirical observation reported in [15] that geometrically decaying step sizes with a longer first stage, defined as

$$\eta_t = \begin{cases} \eta & \text{if } t < T_0 \\ \eta \cdot \gamma^{\lceil \frac{t+1-T_0}{N} \rceil} & \text{if } t \ge T_0 \end{cases}$$

for $\gamma \in (0,1)$, achieve faster convergence than standard geometrically decaying step sizes corresponding to $T_0 = 0$.

Our convergence guarantees in Theorems 1–6 for loss functions admitting a smooth surrogate in the sense of Definition 2 require additional assumptions. Correspondingly, Corollaries 1–6 specialize these guarantees to weakly convex functions by specifying the respective additional conditions required.

To the best of our knowledge, this is the first non-asymptotic convergence analysis of the subgradient method for a class of functions that strictly contains the class of weakly convex functions. We expect that the analytical framework introduced through Definition 2 may also provide a useful foundation for analyzing other first-order optimization algorithms.

2 Related Work

2.1 Subgradient Method Under Weak Convexity

For non-smooth, non-convex optimization, weak convexity has served as the canonical structural framework for establishing non-asymptotic convergence guarantees. The foundational work of Davis and Drusvyatskiy [8] established that standard and stochastic subgradient methods achieve an $O(k^{-1/4})$ non-asymptotic convergence rate for Lipschitz weakly convex functions in terms of the Moreau envelope gradient norm $\|\nabla f_\lambda(x_k)\|$. Subsequent research significantly expanded this theoretical line along several key directions. In particular, Li et al. [12] addressed the restrictive assumption of global Lipschitz continuity.Specifically, they established a non-asymptotic rate of $\min_{t\in[1,T]} \|\nabla f_\lambda(x_t)\|^2 = \mathcal{O}\left(1/\sqrt{T}\right)$ under a constant step length, alongside an asymptotic convergence guarantee $\lim_{T\to\infty} \|\nabla f_\lambda(x_T)\| = 0$ for not summable, but square summable step length. Analogous non-asymptotic and asymptotic convergence guarantees are established in Theorems 1 and 2 in our work, respectively, under similar settings. The primary distinction between the work of Li et al. [12] and ours lies in the theoretical bottleneck addressed by each paper. While Li et al. [12] focus on relaxing global Lipschitz continuity—an auxiliary assumption often adopted for technical convenience—our work resolves the far more fundamental bottleneck of weak convexity itself.

To obtain faster convergence rates for the subgradient method, Davis et al. [9] analyzed weakly convex functions satisfying a sharpness condition, defined as

$$f(x) - f^* \geq \sigma \cdot \inf_{x^* \in X^*} \|x - x^*\|, \qquad \forall x \in \mathbb{R}^d$$

for some $\sigma > 0$, where $f^* := \min_{x\in\mathbb{R}^d} f(x)$ and $X^* := \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. This standard sharpness condition has been widely adopted across non-smooth optimization [9, 16, 17]. Under this setting, Davis et al. [9] established local linear convergence for the subgradient method using a normalized, geometrically decaying step size $\eta_t = \frac{\eta}{\|g_t\|}\gamma^t$ for $\gamma \in (0,1)$. Although this result shares similarities with Theorem 6 in our work, Theorem 6 offers four distinct advantages: (i) it permits an arbitrary stage length N,

whereas [9] restricts $N = 1$; (ii) it dispenses with step-size normalization, a technique rarely used in practical machine learning applications; (iii) it allows the initial point to be chosen arbitrarily; and (iv) most importantly, it holds beyond the class of weakly convex functions.

2.2 Subgradient Method Beyond Weak Convexity

Davis et al. [11] demonstrated that the stochastic subgradient method guarantees that iterate accumulation points are almost surely Clarke stationary for locally Lipschitz objectives with Whitney stratifiable graphs, without assuming weak convexity. Their proof machinery models discrete iterate trajectories via continuous-time differential inclusions. Building on this dynamical perspective, Bolte and Pauwels [18] introduced conservative set-valued fields as a generalization of Clarke subdifferentials, providing a rigorous mathematical foundation for generalized derivatives computed via automatic differentiation. For mini-batch stochastic subgradient methods applied to finite sums of definable, locally Lipschitz functions, they proved that under standard step-size rules and bounded iterate assumptions, every accumulation point is almost surely critical with respect to the underlying conservative field, with the objective function attaining a constant value across the entire accumulation set.

While these continuous-time frameworks operate outside weak convexity, their structural requirements are incomparable to it—meaning they fail to encompass classical convergence guarantees for weakly convex functions. Moreover, a fundamental limitation of these differential inclusion approaches is that they yield purely asymptotic convergence guarantees. Consequently, these frameworks offer little actionable guidance for step-size scheduling in practical machine learning applications.

2.3 Extended Methods Beyond Weak Convexity

To establish non-asymptotic complexity bounds for non-smooth, non-convex functions beyond the weakly convex regime, a growing line of research has turned to alternative criteria of stationarity. First, the Goldstein Approximate Stationarity (GAS) framework seeks a point x satisfying $\mathrm{dist}(0, \partial_\delta f(x)) \leq \varepsilon$ for a target tolerance $\varepsilon > 0$ and radius $\delta > 0$, where $\partial_\delta f(x) = \mathrm{conv}\{\nabla f(y) : y \in B(x, \delta) \cap D_f\}$ denotes the Goldstein δ-subdifferential and D_f denotes the domain where f is differentiable. Recent works [19–21] establish non-asymptotic oracle complexities of $O(\delta^{-1}\varepsilon^{-3})$ for randomized sampling algorithms, whereas deterministic methods obtain an $O(\delta^{-1}\varepsilon^{-4})$ complexity [22]. Second, addressing the limitation that GAS permits convex combinations of subgradients across non-smooth kinks, Tian et al. [23] introduced Near-Approximate Stationarity (NAS), which aims to find a point x such that $\mathrm{dist}(0, \bigcup_{y \in B(x,\delta)} \partial f(y)) \leq \varepsilon$, where $\partial f(y)$ denotes the Clarke subdifferential. By demanding a small subgradient at an actual point within the neighborhood rather than taking a convex hull, NAS provides a strictly stronger notion of stationarity than GAS while achieving a matching $O(\delta^{-1}\varepsilon^{-3})$ non-asymptotic complexity bound under subdifferential regularity conditions. Third, to capture algorithmic stability under local perturbations, recent studies [24, 25] analyze distributional criteria such as (δ, ε)-stationarity, defined via $P_\delta f(x) \leq \varepsilon$, where $P_\delta f(x)$ represents a perturbed envelope function evaluated

at smoothing parameter $\delta > 0$. Notably, Kornowski [25] established that achieving (δ, ε)-stationarity requires a fundamental first-order oracle complexity of $\Theta(\delta^{-1}\varepsilon^{-3})$, proving a tight lower bound matching Goldstein stationarity bounds.

While these alternative criteria enable explicit non-asymptotic rates for general non-weakly convex functions, finding points that satisfy GAS, NAS, or (δ, ε)-stationarity fundamentally requires specialized algorithms equipped with local subgradient sampling, inner loops, or randomized perturbations. Consequently, these frameworks fail to explain the empirical success of the standard subgradient method in machine learning applications.

3 Notation and Preliminaries

Throughout, we consider the Euclidean space $\mathbb{R}^d$, equipped with the inner product $\langle \cdot, \cdot \rangle$ and the induced Euclidean norm $\|x\| := \sqrt{\langle x, x \rangle}$.

We now recall several notions and their properties used throughout the paper.

Definition 3 (L-smooth)**.** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall x, y \in \mathbb{R}^d.$$

For an L-smooth function, Lemma 1 is commonly used in the convergence analysis of gradient methods.

Lemma 1 ([26])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and bounded below. Then,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d$$

and

$$f(x) - \min_{x \in \mathbb{R}^d} f(x) \geq \frac{1}{2L}\|\nabla f(x)\|^2, \qquad \forall x \in \mathbb{R}^d.$$

Definition 4 (Locally PL)**.** *A locally Lipschitz function f is said to be locally PL if f attains a global minimum $\min_{x \in \mathbb{R}^d} f(x) > -\infty$ and for any compact set $S \subset \mathbb{R}^d$, there exists $\mu > 0$ such that*

$$\frac{1}{2} \inf_{g \in \partial f(x)} \|g\|^2 \geq \mu \left(f(x) - \min_{x \in \mathbb{R}^d} f(x) \right), \qquad \forall x \in S.$$

Definition 5 (Coercive)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be coercive if its sublevel set $\{x \in \mathbb{R}^d \mid f(x) - \min_{x \in \mathbb{R}^d} f(x) \leq \delta\}$ is bounded for any $\delta > 0$.*

Definition 6 (ρ-weakly Convex)**.** *A locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex, or equivalently, if*

1. *The subgradient inequality holds:*

$$f(y) \geq f(x) + \langle g_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d, \forall g_x \in \partial f(x). \tag{4}$$

2. *The subdifferential map is hypomonotone:*

$$\langle g_x - g_y,\ x - y\rangle \geq -\rho\|x - y\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ \forall g_x \in \partial f(x),\ \forall g_y \in \partial f(y). \tag{5}$$

The equivalence above follows from [27, Theorem 3.1] and [8, Lemma 2.1].

Definition 7 (Moreau Envelope)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a ρ-weakly convex function and let $\lambda \in (0, 1/\rho)$. The Moreau envelope of f is*

$$f_\lambda(x) := \inf_{y \in \mathbb{R}^d} \left(f(y) + \frac{1}{2\lambda}\|y - x\|^2 \right)$$

and the proximal mapping is

$$p_\lambda(x) := \operatorname*{argmin}_{y \in \mathbb{R}^d} \left(f(y) + \frac{1}{2\lambda}\|y - x\|^2 \right).$$

Lemma 2 summarizes the properties of the Moreau envelope that are used in this paper.

Lemma 2. *Let f be a ρ-weakly convex function and let $\lambda < \frac{1}{\rho}$. Then,*

(i) f_λ *is* $L_\lambda := \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}$*-smooth.*
(ii) $\operatorname{argmin}_{x \in \mathbb{R}^d} f(x) = \operatorname{argmin}_{x \in \mathbb{R}^d} f_\lambda(x)$.
(iii) $\nabla f_\lambda(x) = \frac{x - p_\lambda(x)}{\lambda} \in \partial f(p_\lambda(x))$.
(iv) If f is coercive, then f_λ is also coercive.
(v) If f is coercive and locally PL, then f_λ is also locally PL.

Proof (i) follows from [28, Lemma 7.2].

(ii) follows from [28, Proposition 7.3].

(iii) follows from [29, Theorem 23.5 and Theorem 31.5].

(iv) Fix $\delta > 0$. We show that

$$X_\delta := \left\{ x \in \mathbb{R}^d \mid f_\lambda(x) - \min_{x \in \mathbb{R}^d} f_\lambda(x) \leq \delta \right\}$$

is bounded. Let $f^* := \min_{x \in \mathbb{R}^d} f(x)$. First, using (ii), we have

$$\|p_\lambda(x) - x\|^2 = 2\lambda(f_\lambda(x) - f(p_\lambda(x))) \leq 2\lambda\left(f_\lambda(x) - f^*\right) \leq 2\lambda\delta, \qquad \forall x \in X_\delta.$$

Second, we have

$$f(p_\lambda(x)) - f^* \leq f_\lambda(x) - f^* \leq \delta, \qquad \forall x \in X_\delta.$$

Since f is coercive, there exists $R > 0$ such that

$$\|p_\lambda(x)\| \leq R, \qquad \forall x \in X_\delta.$$

Combining the above estimates, we obtain

$$\|x\| \leq \|p_\lambda(x)\| + \|p_\lambda(x) - x\| \leq R + \sqrt{2\lambda\delta}, \qquad \forall x \in X_\delta.$$

Hence, X_δ is bounded.

(v) Let $X \subset \mathbb{R}^d$ be an arbitrary compact set. Since f is coercive,

$$\hat{X} := \left\{ x \in \mathbb{R}^d \mid f(x) \leq \max_{x \in X} f(x) \right\} \supset X$$

is also compact. Since f is locally PL, there exists $\mu_f > 0$ such that f is μ_f-PL on $\hat{X}$. Since $p_\lambda(x) \in \hat{X}$ if $x \in \hat{X}$,

$$f(p_\lambda(x)) - \min_{x \in \mathbb{R}^d} f(x) \leq \frac{1}{2\mu_f} \inf_{g \in \partial f(p_\lambda(x))} \|g\|^2.$$

Using the definition of the Moreau envelope $f_\lambda(x) = f(p_\lambda(x)) + \frac{1}{2\lambda}\|p_\lambda(x) - x\|^2$ and $\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} f_\lambda(x)$, which follows from (ii),

$$f_\lambda(x) - \min_{x \in \mathbb{R}^d} f_\lambda(x) \leq \frac{1}{2\mu_f} \inf_{g \in \partial f(p_\lambda(x))} \|g\|^2 + \frac{1}{2\lambda}\|p_\lambda(x) - x\|^2.$$

Since $\nabla f_\lambda(x) \in \partial f(p_\lambda(x))$ by (iii), it follows that $\inf_{g \in \partial f(p_\lambda(x))} \|g\|^2 \leq \|\nabla f_\lambda(x)\|^2$. Moreover, $\|p_\lambda(x) - x\| = \lambda\|\nabla f_\lambda(x)\|$ by (iii). Combining these observations,

$$f_\lambda(x) - \min_{x \in \mathbb{R}^d} f_\lambda(x) \leq \left(\frac{1}{2\mu_f} + \frac{\lambda}{2} \right) \|\nabla f_\lambda(x)\|^2 = \frac{1 + \lambda\mu_f}{2\mu_f} \|\nabla f_\lambda(x)\|^2.$$

Rearranging, we obtain

$$\frac{1}{2}\|\nabla f_\lambda(x)\|^2 \geq \frac{\mu_f}{1 + \lambda\mu_f} \left(f_\lambda(x) - \min_{x \in \mathbb{R}^d} f_\lambda(x) \right), \qquad \forall x \in \hat{X}.$$

Since $X \subset \hat{X}$, f_λ is locally PL and the proof is complete.

□

4 Relationship between Smooth Surrogate and Weak Convexity

In this section, we show that the class of functions admitting a smooth surrogate in the sense of Definition 2 is strictly larger than the class of weakly convex functions. First, Proposition 1 shows that every ρ-weakly convex function admits such a surrogate with $k = 2$.

Proposition 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be ρ-weakly convex. Then, for any $\lambda < \frac{1}{\rho}$, f admits a $\left(2, 1 - \lambda\rho, \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}\right)$-smooth surrogate f_λ.*

Proof By Lemma 2 (iii), the condition given in (3) is satisfied. By Lemma 2 (i), f_λ is $\max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}$-smooth.

By the hypomonotonicity of the subdifferential map for weakly convex functions (5),

$$\langle g_x - g_y,\ x - y\rangle \geq -\rho\|x-y\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ \forall g_x \in \partial f(x),\ \forall g_y \in \partial f(y).$$

Let $y := p_\lambda(x)$. Then, by Lemma 2 (iii), $\frac{x-y}{\lambda} \in \partial f(y)$. Substituting $g_y = \frac{x-y}{\lambda}$ into the above inequality yields

$$\langle g_x,\ x - y\rangle \geq \left(\frac{1}{\lambda} - \rho\right)\|x-y\|^2, \qquad \forall x \in \mathbb{R}^d,\ \forall g_x \in \partial f(x).$$

Substituting $x - y = \lambda\nabla f_\lambda(x)$, which follows from Lemma 2 (iii), we obtain

$$\langle g_x, \nabla f_\lambda(x)\rangle \geq (1-\lambda\rho)\|\nabla f_\lambda(x)\|^2, \qquad \forall x \in \mathbb{R}^d,\ \forall g_x \in \partial f(x).$$

Thus, the proof is complete. □

The converse, however, does not hold, i.e., Proposition 2 provides an example of a function that is not weakly convex but still admits a smooth surrogate.

Proposition 2. *The function* $f(x) := (\max\{\alpha x, x\} - 1)^2$ *with* $\alpha \in (0,1)$ *is not weakly convex, but admits a* $(2, \alpha^2, 2)$*-smooth surrogate* $\phi(x) = (x-1)^2$.

Proof First, we show that f is not weakly convex, i.e., for any $\rho > 0$, $f_\rho(x) := f(x) + \frac{\rho}{2}x^2$ is not convex. Considering f_ρ in a neighborhood of $x = 0$, we have

$$f_\rho(0) = 1, \tag{6}$$

$$f_\rho(-\delta) = (-\alpha\delta - 1)^2 + \frac{\rho}{2}\delta^2, \tag{7}$$

$$f_\rho(\delta) = (\delta - 1)^2 + \frac{\rho}{2}\delta^2. \tag{8}$$

Then,

$$\frac{1}{2}f_\rho(-\delta) + \frac{1}{2}f_\rho(\delta) = 1 + \delta(\alpha - 1) + \delta^2\left(\frac{1}{2}(1+\alpha^2) + \frac{\rho}{2}\right).$$

Since $\alpha < 1$, for any $\rho > 0$, there exists $\delta > 0$ such that

$$\frac{1}{2}f_\rho(-\delta) + \frac{1}{2}f_\rho(\delta) < 1 = f_\rho(0) = f_\rho\left(-\frac{1}{2}\delta + \frac{1}{2}\delta\right),$$

which contradicts the convexity of f_ρ. Therefore, f is not ρ-weakly convex for any $\rho > 0$.

Next, we show that f admits a $(2, \alpha^2, 2)$-smooth surrogate $\phi(x) = (x-1)^2$. Clearly, ϕ is 2-smooth. Moreover, f and ϕ share the unique stationary point $x^* = 1$ and therefore condition (3) holds. For $x > 0$, we have $f(x) = \phi(x)$, and thus

$$\langle \nabla f(x), \nabla\phi(x)\rangle = \|\nabla\phi(x)\|^2.$$

For $x < 0$, we have $\nabla f(x) = 2\alpha(\alpha x - 1)$. Hence,

$$\begin{aligned}\langle \nabla f(x), \nabla\phi(x)\rangle - \alpha^2\|\nabla\phi(x)\|^2 &= 4\alpha(\alpha x - 1)(x-1) - 4\alpha^2(x-1)^2 \\ &= 4\alpha(\alpha - 1)(x-1) \\ &\geq 0.\end{aligned}$$

At $x = 0$, $\partial f(0) = [-2, -2\alpha]$. Since every $g_0 \in \partial f(0)$ can be written as $g_0 = -2\bar{\alpha}$ for some $\bar{\alpha} \in [\alpha, 1]$,

$$\langle g_0, \nabla\phi(0)\rangle - \alpha\|\nabla\phi(0)\|^2 = 4\bar{\alpha} - 4\alpha \geq 0 \qquad \forall g_0 \in \partial f(0).$$

Combining the three cases above, we obtain

$$\langle g_x, \nabla\phi(x)\rangle \geq \min\{\alpha^2, \alpha\}\|\nabla\phi(x)\|^2 = \alpha^2\|\nabla\phi(x)\|^2 \qquad \forall x \in \mathbb{R}^d, \forall g_x \in \partial f(x).$$

Thus, the proof is complete. □

The function $x \mapsto (\max\{\alpha x, x\} - 1)^2$ in Proposition 2 can be interpreted as the squared loss of a neural network with one input unit, one hidden unit, and one output unit, equipped with a Leaky ReLU activation parameterized by α and a fixed output-layer weight. Specifically, the network is trained on a single input-label pair $(1, 1)$. Although this example is deliberately simple, it suggests that many loss functions arising from practical neural networks employing the ReLU family of activation functions may fail to be weakly convex while still admitting smooth surrogates.

Beyond weakly convex functions, we are not aware of any systematic procedure for constructing smooth surrogates as in Definition 2. Nevertheless, the subgradient method does not require knowledge of the smooth surrogate, which is used solely as an analytical tool in the convergence analysis.

5 Convergence Analysis

We consider three structural settings and establish convergence bounds by adaptively choosing proper step sizes in each setting.

5.1 General Case

In this section, we assume that the loss function f is G-Lipschitz continuous. Equivalently, the norm of every element of the Clarke subdifferential of f is bounded by G [3], i.e.,

$$\sup_{g \in \partial f(x)} \|g\| \leq G, \qquad \forall x \in \mathbb{R}^d.$$

Theorem 1 establishes a non-asymptotic convergence guarantee for the subgradient method with a constant step size.

Theorem 1. *Suppose that f is G-Lipschitz and admits (k, C, L)-smooth surrogate ϕ that attains a global minimum $\min_{x\in\mathbb{R}^d} \phi(x) > -\infty$. Then, the subgradient method with step size $\eta_t = \eta/\sqrt{1+T}$ satisfies*

$$\min_{t\in[0,T]} \|\nabla\phi(x_t)\|^2 \leq \left(\frac{2\left(\phi(x_0) - \min_{x\in\mathbb{R}^d}\phi(x)\right) + LG^2\eta^2}{2C\eta\sqrt{1+T}}\right)^{2/k}.$$

Proof Since the subgradient method updates the sequence as $x_{t+1} = x_t - \eta_t g_t$ for $g_t \in \partial f(x_t)$,

$$\phi(x_{t+1}) - \phi(x_t) \leq -\eta_t \langle \nabla\phi(x_t), g_t\rangle + \frac{\eta_t^2 L}{2}\|g_t\|^2 \leq -\eta_t C\|\nabla\phi(x_t)\|^k + \frac{\eta_t^2 LG^2}{2},$$

where the first inequality follows from L-smoothness of ϕ and the second inequality follows from the assumption that f is G-Lipschitz and admits (k, C, L)-smooth surrogate ϕ. Taking the telescoping sum from $t = 0$ to $t = T$,

$$\phi(x_{T+1}) - \phi(x_0) \leq -C \sum_{t=0}^{T} \eta_t \|\nabla \phi(x_t)\|^k + \frac{LG^2}{2} \sum_{t=0}^{T} \eta_t^2.$$

Rearranging this using $\phi(x_{T+1}) \geq \min_{x \in \mathbb{R}^d} \phi(x)$,

$$\min_{t \in [0,T]} \|\nabla \phi(x_t)\|^k \leq \frac{2\left(\phi(x_0) - \min_{x \in \mathbb{R}^d} \phi(x)\right) + LG^2 \sum_{t=0}^{T} \eta_t^2}{2C \sum_{t=0}^{T} \eta_t}.$$

Substituting $\eta_t = \eta/\sqrt{1+T}$,

$$\min_{t \in [0,T]} \|\nabla \phi(x_t)\|^k \leq \frac{2\left(\phi(x_0) - \min_{x \in \mathbb{R}^d} \phi(x)\right) + LG^2 \eta^2}{2C\eta\sqrt{1+T}}.$$

Raising both sides to the power $2/k$ completes the proof. □

Theorem 1 establishes that the subgradient method finds a point at which the gradient norm of the smooth surrogate ϕ is sufficiently small. Corollary 1 specializes this convergence guarantee to Lipschitz weakly convex functions.

Corollary 1. *Suppose that f is G-Lipschitz, ρ-weakly convex, and attains a global minimum $\min_{x \in \mathbb{R}^d} f(x) > -\infty$. Then, for any $\lambda \in (0, 1/\rho)$, the subgradient method with step size $\eta_t = \eta/\sqrt{1+T}$ satisfies*

$$\min_{t \in [0,T]} \|\nabla f_\lambda(x_t)\|^2 \leq \frac{2\left(f_\lambda(x_0) - \min_{x \in \mathbb{R}^d} f_\lambda(x)\right) + \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\} G^2 \eta^2}{2(1-\lambda\rho)\eta\sqrt{1+T}}.$$

Proof By Proposition 1, for any $\lambda < \frac{1}{\rho}$, f admits $\left(2, 1-\lambda\rho, \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}\right)$-smooth surrogate f_λ. Hence, Theorem 1 yields the desired result. □

Corollary 1 almost recovers the result of [12, Theorem 2], which demonstrates that the subgradient method with step size $\eta_t = \eta/\left(\|g_t\|\sqrt{1+T}\right)$ achieves $\min_{t \in [0,T]} \|\nabla f_\lambda(x_t)\|^2 = \mathcal{O}(1/\sqrt{T})$ for locally Lipschitz weakly convex functions. The key distinction is that Corollary 1 omits step-size normalization, at the expense of assuming the loss function is globally Lipschitz continuous.

Theorem 2 establishes an asymptotic convergence guarantee for the subgradient method with a non-summable but square-summable step size.

Theorem 2. *Suppose that f is G-Lipschitz and admits (k, C, L)-smooth surrogate ϕ that attains a global minimum $\min_{x \in \mathbb{R}^d} \phi(x) > -\infty$. Then, the subgradient method with step size η_t satisfying $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$ satisfies*

$$\lim_{T \to \infty} \|\nabla \phi(x_T)\| = 0.$$

Proof Following the argument in the proof of Theorem 1,

$$\phi(x_{T+1}) - \phi(x_0) \leq -C \sum_{t=0}^{T} \eta_t \|\nabla \phi(x_t)\|^k + \frac{LG^2}{2} \sum_{t=0}^{T} \eta_t^2. \tag{9}$$

Rearranging this using $\phi(x_{T+1}) \geq \min_{x \in \mathbb{R}^d} \phi(x)$,

$$\sum_{t=0}^{T} \eta_t \|\nabla \phi(x_t)\|^k \leq \frac{\phi(x_0) - \min_{x \in \mathbb{R}^d} \phi(x)}{C} + \frac{LG^2}{2C} \sum_{t=0}^{T} \eta_t^2.$$

Since $\sum_{t=0}^{\infty} \eta_t^2 < \infty$,

$$\sum_{t=0}^{\infty} \eta_t \|\nabla \phi(x_t)\|^k < \infty. \tag{10}$$

We show $\lim_{T \to \infty} \|\nabla \phi(x_T)\| \to 0$ by following the argument in [12, Theorem 3]. For any $\varepsilon > 0$, let $N_\varepsilon := \{j \in \mathbb{N} : \|\nabla \phi(x_j)\| \leq \varepsilon\}$. It immediately follows from (10) and $\sum_{t=0}^{\infty} \eta_t = \infty$ that $\liminf_{j \to \infty} \|\nabla \phi(x_j)\| = 0$, which implies that N_ε is an infinite set. Let $\bar{N}_\varepsilon$ be the complementary set of N_ε in $\mathbb{N}$. Then, we have

$$\varepsilon^k \sum_{j \in \bar{N}_\varepsilon} \eta_j \leq \sum_{j \in \bar{N}_\varepsilon} \eta_j \|\nabla \phi(x_j)\|^k \leq \sum_{j \in \mathbb{N}} \eta_j \|\nabla \phi(x_j)\|^k < \infty,$$

where the last inequality is from (10). Hence, for any $\delta > 0$, there exists an integer $\mathrm{n}(\delta)$ such that

$$\sum_{\ell \geq \mathrm{n}(\delta),\, \ell \in \bar{N}_\varepsilon} \eta_\ell \leq \delta.$$

Note that ϕ is L-smooth for some $L > 0$. Next, we take an arbitrary $\gamma > 0$ and set $\varepsilon = \gamma/2, \delta = \gamma/(2LG)$. For all $j \geq \mathrm{n}(\delta)$, if $j \in N_\varepsilon$, then we have $\|\nabla \phi(x_j)\| \leq \varepsilon < \gamma$; otherwise $j \in \bar{N}_\varepsilon$. Let m be the smallest element in the set $\{\ell \in N_\varepsilon : \ell \geq j\}$. Clearly, m is finite since N_ε is an infinite set. Without loss of generality, we can assume that $m > j$. Then, by the L-Lipschitz continuity of $\nabla \phi$, (32), and the definitions of ε, δ, and m, we can compute

$$\begin{aligned}
\|\nabla \phi(x_j)\| &\leq \|\nabla \phi(x_j) - \nabla \phi(x_m)\| + \|\nabla \phi(x_m)\| \\
&\leq L \sum_{\ell=j}^{m-1} \|x_{\ell+1} - x_\ell\| + \varepsilon \\
&\leq LG \sum_{\ell=j}^{m-1} \eta_\ell + \frac{\gamma}{2} \\
&= LG \sum_{j \leq \ell < m,\, \ell \in \bar{N}_\varepsilon} \eta_\ell + \frac{\gamma}{2} \\
&\leq LG \sum_{\ell \geq \mathrm{n}(\delta),\, \ell \in \bar{N}_\varepsilon} \eta_\ell + \frac{\gamma}{2} \\
&\leq \gamma.
\end{aligned}$$

Here, the equality is due to the fact that for all ℓ with $j \leq \ell < m$, we have $\ell \in \bar{N}_\varepsilon$ by the choice of m. Since $\gamma > 0$ is arbitrary, we have $\|\nabla \phi(x_j)\| \to 0$ as $j \to \infty$. □

From an asymptotic perspective, a non-summable but square-summable step-size sequence—i.e., $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$—is sufficient to guarantee convergence in terms of the gradient norm of the smooth surrogate ϕ. For instance, a polynomially

decaying step size of the form $\eta_t = \eta/(1+t)^\alpha$ with $\alpha \in (1/2, 1]$ satisfies these conditions. Theorem 2 requires only the Lipschitz continuity of f alongside the existence of a smooth surrogate, without requiring the locally smooth or sharp structures discussed in Sections 5.2 and 5.3. Consequently, a polynomially decaying step size can be regarded as one of the most robust choices for ensuring convergence in the asymptotic sense.

Corollary 2 specializes this convergence guarantee to Lipschitz weakly convex functions.

Corollary 2. *Suppose that f is G-Lipschitz, ρ-weakly convex, and attains a global minimum $\min_{x\in\mathbb{R}^d} f(x) > -\infty$.. Then, for any $\lambda \in (0, 1/\rho)$, the subgradient method with step size η_t satisfying $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$ satisfies*

$$\lim_{T\to\infty} \|\nabla f_\lambda(x_T)\| = 0.$$

Proof By Proposition 1, for any $\lambda < \frac{1}{\rho}$, f admits $\left(2, 1-\lambda\rho, \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}\right)$-smooth surrogate f_λ. Hence, Theorem 2 yields the desired result. □

Corollary 2 almost recovers the result of [12, Theorem 3], which demonstrates that the subgradient method with step size $\eta_t = \beta_t/\|g_t\|$ satisfying $\sum_{t=0}^{\infty} \beta_t = \infty$ and $\sum_{t=0}^{\infty} \beta_t^2 < \infty$ achieves $\lim_{T\to\infty} \|\nabla f_\lambda(x_T)\| = 0$ for locally Lipschitz weakly convex functions. As in the discussion regarding Corollary 1, the key distinction is that Corollary 2 omits step-size normalization, at the expense of assuming global Lipschitz continuity of the loss function.

5.2 Locally Smooth Case

In this section, we assume that the loss function f possesses the following locally smooth structure.

Assumption 1 (Locally Smooth)**.** *For any $\delta > 0$ there exists $\theta > 0$ such that*

$$\sup_{g\in\partial f(x)} \|g\|^2 \le \theta \|\nabla\phi(x)\|^2 \qquad \forall x \in \mathbb{R}^d \text{ such that } \phi(x) - \min_{x\in\mathbb{R}^d} \phi(x) \le \delta. \tag{11}$$

Since ϕ is L-smooth, the locally smooth condition implies that every element of Clarke subdifferential of f tends to zero near the minimizer of ϕ. Theorem 3 shows the convergence bound for the gradient norm of f, whereas Theorem 1 concerns the gradient norm of the surrogate.

Theorem 3. *Suppose that f admits a $(2, C, L)$-smooth surrogate ϕ that is bounded below and satisfies Assumption 1. Then, for any initial point x_0, the subgradient method with a sufficiently small constant step size η satisfies*

$$\min_{t\in[0,T-1]} \sup_{g\in\partial f(x_t)} \|g\|^2 \le \frac{q_1}{T\eta}$$

for all $T > 0$, where $q_1 > 0$ is a constant independent of T and η.

Proof Suppose that $\eta > 0$ is sufficiently small so that $C - \frac{\eta L \theta}{2} \geq C/2$. In the proof of Theorem 1, we showed

$$\phi(x_{t+1}) - \phi(x_t) \leq -\eta C \|\nabla \phi(x_t)\|^2 + \frac{\eta^2 L}{2} \|g_t\|^2.$$

If $\phi(x_t) \leq \phi(x_0)$, Assumption 1 gives

$$\phi(x_{t+1}) - \phi(x_t) \leq -\eta \left(C - \frac{\eta L \theta}{2} \right) \|\nabla \phi(x_t)\|^2 \leq -\frac{\eta C}{2} \|\nabla \phi(x_t)\|^2.$$

Thus, $\phi(x_t) \leq \phi(x_0)$ holds for all $t \geq 0$ and the above inequality is valid for all $t \geq 0$. Taking the telescoping sum,

$$\phi(x_T) - \phi(x_0) \leq -\frac{\eta C}{2} \sum_{t=0}^{T-1} \|\nabla \phi(x_t)\|^2.$$

Rearranging this using $\phi(x_T) \geq \min_{x \in \mathbb{R}^d} \phi(x)$,

$$\min_{t \in [0, T-1]} \|\nabla \phi(x_t)\|^2 \leq \frac{2 \left(\phi(x_0) - \min_{x \in \mathbb{R}^d} \phi(x) \right)}{T \eta C}.$$

Finally, applying (11) and letting $q_1 := \frac{2\theta \left(\phi(x_0) - \min_{x \in \mathbb{R}^d} \phi(x) \right)}{C}$ complete the proof. □

Moreover, if f and ϕ satisfy additional conditions, then Theorem 4 establishes a convergence bound for the optimality gap of f.

Theorem 4. *Suppose that the assumptions in Theorem 3 are satisfied. Moreover, if f and ϕ are locally PL and ϕ is coercive, then for any initial point x_0, the subgradient method with a sufficiently small constant step size η satisfies*

$$f(x_T) - \min_{x \in \mathbb{R}^d} f(x) \leq q_2 q_3(\eta)^T$$

for all $T > 0$, where $q_2 \geq 0$ is a constant independent of T, and $q_3(\eta) = 1 - \Theta(\eta) \in (0, 1)$.

Proof Let

$$X := \left\{ x \in \mathbb{R}^d \mid \phi(x) \leq \phi(x_0) \right\}.$$

In the proof of Theorem 3, we showed that $\{\phi(x_t)\}$ is non-increasing for sufficiently small η. Hence, $x_t \in X$ for all $t \geq 0$. Since ϕ is coercive, X is compact. Since f and ϕ are locally PL and X is compact, there exist μ_ϕ and μ_f such that

$$\frac{1}{2} \|\nabla \phi(x_t)\|^2 \geq \mu_\phi \left(\phi(x_t) - \min_{x \in \mathbb{R}^d} \phi(x) \right), \qquad \forall t \geq 0, \tag{12}$$

$$\frac{1}{2} \inf_{g \in \partial f(x_t)} \|g\|^2 \geq \mu_f \left(f(x_t) - \min_{x \in \mathbb{R}^d} f(x) \right), \qquad \forall t \geq 0. \tag{13}$$

Substituting (12) into $\phi(x_{t+1}) - \phi(x_t) \leq -\eta\left(C - \frac{\eta L\theta}{2}\right)\|\nabla\phi(x_t)\|^2$, shown in the proof of Theorem 3,

$$\phi(x_{t+1}) - \min_{x\in\mathbb{R}^d}\phi(x) \leq \left(1 - \eta\left(2C - \eta L\theta\right)\mu_\phi\right)\left(\phi(x_t) - \min_{x\in\mathbb{R}^d}\phi(x)\right).$$

Hence,

$$\phi(x_T) - \min_{x\in\mathbb{R}^d}\phi(x) \leq q_3(\eta)^T\left(\phi(x_0) - \min_{x\in\mathbb{R}^d}\phi(x)\right),$$

where $q_3(\eta) := 1 - \eta\left(2C - \eta L\theta\right)\mu_\phi = 1 - \Theta(\eta) \in (0,1)$.

Meanwhile, for all $t \geq 0$, we have

$$f(x_t) - \min_{x\in\mathbb{R}^d} f(x) \overset{(13)}{\leq} \frac{1}{2\mu_f}\inf_{g\in\partial f(x_t)}\|g\|^2 \overset{(11)}{\leq} \frac{\theta}{2\mu_f}\|\nabla\phi(x_t)\|^2 \leq \frac{L\theta}{\mu_f}\left(\phi(x_t) - \min_{x\in\mathbb{R}^d}\phi(x)\right),$$

where the last inequality follows from Lemma 1.

Thus, we obtain

$$f(x_T) - \min_{x\in\mathbb{R}^d} f(x) \leq \frac{L\theta}{\mu_f}q_3(\eta)^T\left(\phi(x_0) - \min_{x\in\mathbb{R}^d}\phi(x)\right).$$

Letting $q_2 := \frac{L\theta}{\mu_f}\left(\phi(x_0) - \min_{x\in\mathbb{R}^d}\phi(x)\right)$ completes the proof. □

In Theorems 3 and 4, Assumption 1 plays an essential role in their proof. We next show that this condition is satisfied when f is weakly convex, coercive, and satisfies Assumption 2. As a consequence, Theorem 3 can be specialized to weakly convex functions in Corollary 3.

Assumption 2. *f has a unique stationary point x^*, which is also a minimizer. Moreover, $f \in C^2$ on $\{x \in \mathbb{R} \mid \|x - x^*\| \leq R\}$ for some $R > 0$.*

Corollary 3. *Suppose that f is ρ-weakly convex and coercive, and satisfies Assumption 2. Then, for any initial point x_0, the subgradient method with a sufficiently small constant step size η satisfies*

$$\min_{t\in[0,T-1]}\sup_{g\in\partial f(x_t)}\|g\|^2 \leq \frac{q_1}{T\eta} \tag{14}$$

for all $T > 0$, where $q_1 > 0$ is a constant independent of T and η.

Proof By Proposition 1, f admits $\left(2, 1 - \lambda\rho, \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}\right)$-smooth surrogate f_λ. Therefore, by Theorem 3, (14) follows if Assumption 1 is satisfied. Specifically, we define

$$X_\delta := \left\{x \in \mathbb{R}^d \mid f_\lambda(x) \leq \delta\right\}.$$

and show that for any $\delta > \min_{x\in\mathbb{R}^d} f_\lambda(x)$, there exists $\theta > 0$ such that

$$\sup_{g\in\partial f(x)}\|g\|^2 \leq \theta\|\nabla f_\lambda(x)\|^2, \qquad \forall x \in X_\delta. \tag{15}$$

For $x \in \mathbb{R}$ and $R \geq 0$, $B(x,R)$ denote open ball of radius R, i.e., $B(x,R) := \{y \in \mathbb{R} \mid \|y - x\| < R\}$. Moreover, define

$$S(\varepsilon) := \left\{x \in \mathbb{R}^d \mid f(x) - f(x^*) \leq \varepsilon\right\}.$$

We begin by claiming that for $R > 0$ specified by Assumption 2, there exists $r > 0$ and $\varepsilon > 0$ such that

$$B(x^*, r) \subset S(\varepsilon) \subset B(x^*, R). \tag{16}$$

We first show the second inclusion. Suppose, to the contrary, that no such $\varepsilon > 0$ exists. Then, for every $n \in \mathbb{N}$,

$$S\left(\frac{1}{n}\right) \not\subset B(x^*, R).$$

Hence, there exists $y_n \notin B(x^*, R)$ satisfying

$$f(y_n) - f(x^*) \leq \frac{1}{n}.$$

Since $\{y_n\} \subset S(1)$ and $S(1)$ is compact by assumption, $\{y_n\}$ has a convergent subsequence $\{z_n\}$ whose limit is denoted by $\bar{z}$. Since f is continuous,

$$f(\bar{z}) = \lim_{n\to\infty} f(z_n) = f(x^*).$$

Moreover, since x^* is the unique minimizer of f by Assumption 2, $\bar{z} = x^*$. On the other hand, since $y_n \notin B(x^*, R)$ for every n, we have $\bar{z} \neq x^*$, which proves the contradiction. Therefore, the second inclusion is proved. The first inclusion follows for arbitrary ε from the continuity of f. Thus, (16) is proved.

We now consider two exhaustive cases.

Case 1. $x \in B(x^*, r)$. It follows from the definition of the Moreau envelope that

$$f(p_\lambda(x)) = f_\lambda(x) - \frac{1}{2\lambda}\|p_\lambda(x) - x\|^2 \leq f_\lambda(x) \leq f(x).$$

By (16), we have

$$x, p_\lambda(x) \in S(\varepsilon) \subset B(x^*, R).$$

Since $B(x^*, R)$ is convex, the line segment joining x and $p_\lambda(x)$ lies entirely in $B(x^*, R)$. Moreover, by Assumption 2, there exists $L_f \geq 0$ such that $\|\nabla^2 f(x)\|_{op} \leq L_f$ for all $x \in B(x^*, R)$. Hence, the integral form of Taylor's theorem yields

$$\begin{aligned}\|\nabla f(x)\| &= \left\|\nabla f(p_\lambda(x)) + \left(\int_0^1 \nabla^2 f\big(p_\lambda(x) + t\big(x - p_\lambda(x)\big)\big)\, dt\right)\big(x - p_\lambda(x)\big)\right\| \\ &\leq \|\nabla f(p_\lambda(x))\| + L_f\|x - p_\lambda(x)\|.\end{aligned}$$

Since f is differentiable at $x \in B(x^*, r) \subset B(x^*, R)$, $\partial f(x) = \{\nabla f(x)\}$. Hence, using the identities $\nabla f_\lambda(x) = \nabla f(p_\lambda(x)) = \frac{1}{\lambda}(x - p_\lambda(x))$, which follow from Lemma 2 (iii), we obtain

$$\sup_{g\in\partial f(x)} \|g\|^2 \leq (1 + \lambda L_f)^2\|\nabla f_\lambda(x)\|^2, \qquad \forall x \in B(x^*, r). \tag{17}$$

Case 2. $x \in X_\delta \setminus B(x^*, r)$. By Lemma 2 (iv), X_δ is compact. Since f is locally Lipschitz, there exists $G_1 > 0$ such that

$$\sup_{g\in\partial f(x)} \|g\|^2 \leq G_1 < \infty, \qquad \forall x \in X_\delta \setminus B(x^*, r).$$

We then show that there exists $G_2 > 0$ such that

$$\|\nabla f_\lambda(x)\|^2 > G_2, \qquad \forall x \in X_\delta \setminus B(x^*, r).$$

Since $X_\delta \setminus B(x^*, r)$ is compact and ∇f_λ is continuous, $\min_{x\in X_\delta\setminus B(x^*,r)} \|\nabla f_\lambda(x)\|^2$ exists. Suppose, to the contrary, that $\min_{x\in X_\delta\setminus B(x^*,r)} \|\nabla f_\lambda(x)\|^2 = 0$. Then, there exists $\bar{x} \in X_\delta \setminus B(x^*, r)$ such that $\nabla f_\lambda(\bar{x}) = 0$. Since $0 = \nabla f_\lambda(\bar{x}) \in \partial f(p_\lambda(\bar{x}))$ holds by Lemma 2 (iii),

$p_\lambda(\bar{x})$ is a stationary point of f. Since x^* is the unique stationary point of f by Assumption 2, we obtain $p_\lambda(\bar{x}) = x^*$. Consequently,

$$\nabla f_\lambda(\bar{x}) = \frac{1}{\lambda}(\bar{x} - p_\lambda(\bar{x})) = \frac{1}{\lambda}(\bar{x} - x^*) = 0,$$

which implies $\bar{x} = x^*$, contradicting $\bar{x} \in X_\delta \setminus B(x^*, r)$. Thus, such $G_2 > 0$ exists. Therefore, we obtain

$$\sup_{g \in \partial f(x)} \|g\|^2 \leq \frac{G_1}{G_2} \|\nabla f_\lambda(x)\|^2 \qquad \forall x \in X_\delta \setminus B(x^*, r). \tag{18}$$

Combining (17) and (18), (15) holds with

$$\theta = \max\left\{(1 + \lambda L_f)^2, \frac{G_1}{G_2}\right\}.$$

□

Similarly, Theorem 4 can be specialized to weakly convex functions in Corollary 4.

Corollary 4. *Suppose that f is ρ-weakly convex, coercive, and locally PL, and satisfies Assumption 2. Then, for any initial point $x_0 \in \mathbb{R}^d$, the subgradient method with a sufficiently small constant step size η satisfies*

$$f(x_T) - \min_{x \in \mathbb{R}^d} f(x) \leq q_2 q_3(\eta)^T$$

for all $T > 0$, where $q_2 > 0$ is a constant independent of T, and $q_3(\eta) = 1 - \Theta(\eta) \in (0, 1)$.

Proof By Lemma 2 (iv) and (v), f_λ is coercive and locally PL. Hence, the claim of the corollary follows immediately from Theorem 4 and Corollary 3. □

Remark 1. *Corollary 4 shows that a constant step size can achieve linear convergence for weakly convex functions. In contrast, a polynomially decaying step size can fail to achieve linear convergence for locally smooth weakly convex functions. For example, consider the subgradient method for $f(x) = \frac{1}{2}x^2 : \mathbb{R} \to \mathbb{R}$. Its sequence is written explicitly as*

$$x_T = x_0 \cdot \Pi_{t=0}^{T-1}(1 - \eta_t).$$

Since $\eta_t \to 0$ under a polynomially decaying step size, the resulting rate is slower than linear.

5.3 Locally Sharp Case

In this section, we consider the locally sharp setting, in which f admits a $(1, C, L)$-smooth surrogate ϕ, i.e.,

$$\langle g_x, \nabla\phi(x)\rangle \geq C\|\nabla\phi(x)\|, \qquad \forall x \in \mathbb{R}^d,\ \forall g_x \in \partial f(x).$$

This condition implies that $\sup_{g \in \partial f(x)} \|g\|$ is bounded away from zero whenever $\nabla\phi(x) \neq 0$.

Lemma 3 shows the property of the sequence of the optimality gap generated by the subgradient method.

Lemma 3. *Suppose that f admits a $(1, C, L)$-smooth surrogate ϕ that is coercive and locally PL. For any initial point x_0, there exists $\hat{\eta} > 0$ such that the subgradient method with step size sequence $\{\eta_t\}$ satisfying $\sup_t \eta_t \leq \hat{\eta}$ satisfies for all $t \geq 0$,*

$$E_{t+1} \leq E_t - \alpha\eta_t\sqrt{E_t} + \beta\eta_t^2 \tag{19}$$

and

$$\begin{cases} E_{t+1} \leq E_t & \text{if } E_t \geq \frac{\beta^2}{\alpha^2}\eta_t^2 \\ E_{t+1} \leq \eta_t^2 \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\} & \text{if } E_t < \frac{\beta^2}{\alpha^2}\eta_t^2 \end{cases} \tag{20}$$

for some $\alpha, \beta > 0$, where $E_t := \phi(x_t) - \min_{x\in\mathbb{R}^d} \phi(x)$.

Proof Since ϕ is continuous and coercive, $\min_{x\in\mathbb{R}^d} \phi(x)$ exists. Fix $\varepsilon > 0$ and define

$$X := \left\{x \in \mathbb{R}^d \mid \phi(x) - \min_{x\in\mathbb{R}^d} \phi(x) \leq \max\{E_0, \varepsilon\}\right\}.$$

Since f is locally Lipschitz and X is compact, there exists $G > 0$ such that

$$\sup_{g\in\partial f(x)} \|g\| \leq G \qquad \forall x \in X. \tag{21}$$

Since ϕ is locally PL, there exists $\mu > 0$ such that

$$\frac{1}{2}\|\nabla\phi(x)\|^2 \geq \mu\left(\phi(x) - \min_{x\in\mathbb{R}^d} \phi(x)\right) \qquad \forall x \in X. \tag{22}$$

We first derive (19) assuming $x_t \in X$. Since ϕ is L-smooth,

$$\begin{aligned} \phi(x_{t+1}) - \phi(x_t) &\leq -\eta_t \langle\nabla\phi(x_t), g_t\rangle + \frac{\eta_t^2 L}{2}\|g_t\|^2 \\ &\leq -\eta_t C\|\nabla\phi(x_t)\| + \frac{\eta_t^2 L}{2}\|g_t\|^2 \\ &\leq -\eta_t\sqrt{2\mu}C\sqrt{\phi(x_t) - \min_{x\in\mathbb{R}^d} \phi(x)} + \frac{\eta_t^2 LG^2}{2}, \end{aligned}$$

where the second inequality follows from the assumption that f admits a $(1, C, L)$-smooth surrogate ϕ and the third inequality follows from $x_t \in X$, (21), and (22). Defining $E_t := \phi(x_t) - \min_{x\in\mathbb{R}^d} \phi(x)$, the above inequality can be rewritten as

$$E_{t+1} \leq E_t - \alpha\eta_t\sqrt{E_t} + \beta\eta_t^2, \tag{23}$$

where $\alpha = \sqrt{2\mu}C$ and $\beta = \frac{LG^2}{2}$.

We now prove by induction that $x_t \in X$ for every $t \geq 0$. The claim is trivial for $t = 0$. We then suppose that $x_t \in X$ for some $t \geq 0$ and show that $x_{t+1} \in X$. Define

$$E_t^* := \frac{\beta^2}{\alpha^2}\eta_t^2.$$

When $E_t > E_t^*$, $E_{t+1} < E_t$ by (23), and hence $x_{t+1} \in X$ by induction hypothesis.

When $E_t \le E_t^*$, since $E_t \mapsto E_t - \alpha\eta_t\sqrt{E_t} + \beta\eta_t^2$ is a convex function, its maximum over $E_t \in [0, E_t^*]$ is achieved at one of the endpoints. Hence,

$$E_{t+1} \le \max\{\beta\eta_t^2, E_t^*\} = \eta_t^2 \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\} \le \sup_t \eta_t^2 \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\}.$$

Let

$$\hat{\eta} = \sqrt{\frac{\max\{E_0, \varepsilon\}}{\max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\}}} > 0.$$

Then $\sup_t \eta_t \le \hat{\eta}$ implies $E_{t+1} \le \max\{E_0, \varepsilon\}$, equivalently, $x_{t+1} \in X$. Hence, by induction, $x_t \in X$ for every $t \ge 0$. Therefore, (19) holds for every $t \ge 0$. This completes the proof. □

Theorem 5 shows that, in the locally sharp setting, a constant step size η achieves linear convergence to a threshold of magnitude η. Moreover, the coefficient of the linear rate gets worse as the step size decreases, revealing a tradeoff between the guaranteed threshold on the optimality gap and the convergence rate.

Theorem 5. *Suppose that f is locally PL and admits a $(1, C, L)$-smooth surrogate ϕ that is coercive and locally PL. For any initial point x_0, the subgradient method with a sufficiently small constant step size η satisfies*

$$f(x_T) - \min_{x \in \mathbb{R}^d} f(x) \le q_1 \left(q_2^T + \eta\right)$$

for all $T \ge 0$, where $q_1 > 0, q_2 \in (0, 1)$ are constants independent of T.

Proof We first show that $E_T := \phi(x_T) - \min_{x \in \mathbb{R}^d} \phi(x)$ is bounded from above as

$$E_T \le \tilde{q}_1 \left(\tilde{q}_2^T + \eta^2\right)$$

for some $\tilde{q}_1 > 0$ and $\tilde{q}_2 \in (0, 1)$.

By Lemma 3, we have

$$E_{t+1} \le E_t - \alpha\eta\sqrt{E_t} + \beta\eta^2$$

for some $\alpha, \beta > 0$. Define

$$E^* := \frac{\beta^2}{\alpha^2}\eta^2.$$

We now consider two exhaustive cases.

Case 1. $E_t > E^*$ holds for all $t \le T$. Subtracting $E^* := \frac{\beta^2}{\alpha^2}\eta^2$ from both sides, for all $t < T$,

$$\begin{aligned} E_{t+1} - E^* &\le E_t - E^* - \alpha\eta\left(\sqrt{E_t} - \sqrt{E^*}\right) \\ &\le \left(1 - \frac{\alpha\eta}{\sqrt{E_t} + \sqrt{E^*}}\right)(E_t - E^*) \\ &\le \left(1 - \frac{\alpha\eta}{\sqrt{E_0} + \sqrt{E^*}}\right)(E_t - E^*), \end{aligned}$$

where the last inequality follows from monotonic decrease of $\{E_t\}_{t \ge 0}$ ensured by Lemma 3. Introducing

$$\tilde{q}_2 := 1 - \frac{\alpha\eta}{\sqrt{E_0} + \sqrt{E^*}} = 1 - \frac{\alpha\eta}{\sqrt{E_0} + \frac{\beta}{\alpha}\eta},$$

we obtain

$$E_T \leq E^* + \tilde{q}_2^T \left(E_0 - E^*\right).$$

Note that $\tilde{q}_2 \in (0,1)$. $\tilde{q}_2 < 1$ follows immediately from $\frac{\alpha\eta}{\sqrt{E_0}+\frac{\beta}{\alpha}\eta} > 0$. $\tilde{q}_2 > 0$ follows from $E_{t+1} - E^*, E_t - E^* > 0$ for all $t < T$.

Case 2. $E_t \leq E^*$ holds at some $t \leq T$. Since $E_t \mapsto E_t - \alpha\eta\sqrt{E_t} + \beta\eta^2$ is a convex function, its maximum over $E_t \in [0, E^*]$ is achieved at one of the endpoints. Hence,

$$E_{t+1} \leq \max\{\beta\eta^2, E^*\}.$$

Since $E_{s+1} \leq E_s$ when $E_s \geq E^*$ for all $s \geq 0$,

$$E_T \leq \max\{E_{T-1}, \beta\eta^2, E^*\} \leq \max\{E_t, \beta\eta^2, E^*\} = \max\{\beta\eta^2, E^*\}.$$

Combining the two cases yields

$$\begin{aligned} E_T &\leq \max\left\{E^* + \tilde{q}_2^T \left(E_0 - E^*\right), \beta\eta^2, E^*\right\} \\ &\leq \max\left\{E^* + \tilde{q}_2^T \left|E_0 - E^*\right|, \beta\eta^2\right\} \\ &\leq \tilde{q}_2^T \left|E_0 - \frac{\beta^2}{\alpha^2}\eta^2\right| + \max\left\{\frac{\beta^2}{\alpha^2}, \beta\right\}\eta^2. \end{aligned}$$

Finally, introducing $\tilde{q}_1 := \max\left\{\left|E_0 - \frac{\beta^2}{\alpha^2}\eta^2\right|, \frac{\beta^2}{\alpha^2}, \beta\right\}$, we obtain

$$E_T \leq \tilde{q}_1 \left(\tilde{q}_2^T + \eta^2\right). \tag{24}$$

Let

$$\begin{aligned} X &:= \left\{x \in \mathbb{R}^d \mid \phi(x) - \min_{x\in\mathbb{R}^d} \phi(x) \leq \tilde{q}_1 \left(1 + \eta^2\right)\right\}, \\ X_\phi^* &:= \left\{x \in \mathbb{R}^d \mid \phi(x) = \min_{x\in\mathbb{R}^d} \phi(x)\right\}, \\ X_f^* &:= \left\{x \in \mathbb{R}^d \mid f(x) = \min_{x\in\mathbb{R}^d} f(x)\right\}. \end{aligned}$$

For $x^* \in X_f^*$, it follows $0 \in \partial f(x^*)$. Since f admits $(1, C, L)$-smooth surrogate, we have $0 \geq C\|\nabla\phi(x^*)\|$, implying that $\|\nabla\phi(x^*)\| = 0$. Moreover, since ϕ is locally PL, we have $\phi(x^*) - \min_{x\in\mathbb{R}^d} \phi(x) \leq \frac{1}{2\mu}\|\nabla\phi(x^*)\|^2$, implying that $\phi(x^*) = \min_{x\in\mathbb{R}^d} \phi(x)$. Thus, we obtain $X_f^* \subset X_\phi^*$. Moreover, for $x^* \in X_\phi^*$, it follows $\nabla\phi(x^*) = 0$. By condition given in (3), $0 \in \partial f(x^*)$. Since f is locally PL, we have $f(x^*) - \min_{x\in\mathbb{R}^d} f(x) \leq \frac{1}{2\mu}\|g\|^2$ for any $g \in \partial f(x^*)$, implying that $f(x^*) = \min_{x\in\mathbb{R}^d} f(x)$. Thus, we obtain

$$X_f^* = X_\phi^*. \tag{25}$$

Since ϕ is coercive, $\mathrm{conv}(X)$ is compact. Since f is locally Lipschitz, there exists $G \geq 0$ such that

$$\sup_{g\in\partial f(x)} \|g\| \leq G \qquad \forall x \in \mathrm{conv}(X).$$

By convexity of $\mathrm{conv}(X)$,

$$f(x) - f(y) \leq \int_0^1 \sup_{g\in\partial f(x+t(y-x))} \|g\|\|x - y\|dt \leq G\|x - y\|, \quad \forall x, y \in X. \tag{26}$$

Moreover, since ϕ is μ-PL on $\mathrm{conv}(X)$ for some $\mu > 0$, the argument in the proof of [30, Theorem 2] yields

$$\inf_{x^* \in X_\phi^*} \|x - x^*\|^2 \leq \frac{2}{\mu}\left(\phi(x) - \min_{x \in \mathbb{R}^d} \phi(x)\right), \quad \forall x \in X. \tag{27}$$

Thus, for any $x \in X$,

$$\begin{aligned} f(x) - \min_{x \in \mathbb{R}^d} f(x) &\overset{(26)}{\leq} \leq G \inf_{x^* \in X_f^*} \|x - x^*\| \\ &\overset{(25)}{=} G \inf_{x^* \in X_\phi^*} \|x - x^*\| \\ &\overset{(27)}{\leq} G\sqrt{\frac{2}{\mu}\left(\phi(x) - \min_{x \in \mathbb{R}^d} \phi(x)\right)}. \end{aligned}$$

Finally, $x_T \in X$ for all $T \geq 0$ yields

$$f(x_T) - \min_{x \in \mathbb{R}^d} f(x) \leq G\sqrt{\frac{2}{\mu}\left(\phi(x_T) - \min_{x \in \mathbb{R}^d} \phi(x)\right)}, \quad \forall T \geq 0.$$

Hence, substituting (24) into the above, we obtain

$$f(x_T) - \min_{x \in \mathbb{R}^d} f(x) \leq G\sqrt{\frac{2\tilde{q}_1}{\mu}\left(\tilde{q}_2^T + \eta^2\right)} \leq G\sqrt{\frac{2\tilde{q}_1}{\mu}}\left(\tilde{q}_2^{T/2} + \eta\right).$$

Redefining $q_1 := G\sqrt{\frac{2\tilde{q}_1}{\mu}}$ and $q_2 := \sqrt{\tilde{q}_2}$ yields the desired result. □

Theorem 6 establishes linear convergence of the optimality gap of f to zero by employing a geometrically decaying step size.

Theorem 6. *Suppose that f is locally PL and admits a $(1, C, L)$-smooth surrogate ϕ that is coercive and locally PL. For any initial point $x_0 \in \mathbb{R}^d$ and any $N \in \mathbb{N}$, the subgradient method with a geometrically decaying step size $\eta_t = \eta \cdot \gamma^{\left\lceil \frac{t-N+1}{N} \right\rceil}$ for a sufficiently small $\eta > 0$ and γ sufficiently close to 1 satisfies*

$$f(x_{T \cdot N}) - \min_{x \in \mathbb{R}^d} f(x) \leq q\gamma^{T-1}$$

for all $T \geq 1$, where $q > 0$ is a constant independent of T.

Proof We use constants $\alpha, \beta, \tilde{q}_1, \tilde{q}_2$ defined in the proof of Lemma 3 and Theorem 5. Let η sufficiently small so that the assumption in Lemma 3 is satisfied. Using these constants, define

$$\zeta := \max\left\{\frac{\tilde{q}_1\left(\tilde{q}_2 + \eta^2\right)}{\eta^2}, \ \frac{\beta^2}{\alpha^2} + 1, \ \beta\right\}.$$

We show, by induction, that there exists $\hat{\gamma} \in (0, 1)$ such that for any $\gamma \in (\hat{\gamma}, 1)$ and for all $T \geq 1$,

$$E_{T \cdot N} \leq \zeta\left(\gamma^{T-1}\eta\right)^2. \tag{28}$$

For $T = 1$, Theorem 5 yields

$$E_N \leq \tilde{q}_1 \left(\tilde{q}_2 + \eta^2 \right).$$

Hence, by the definition of ζ, (28) is true for $T = 1$.

Suppose that (28) holds for $T = i$ and show (28) holds for $T = i+1$. For $t \in [i \cdot N, (i+1) \cdot N]$, by Lemma 3,

$$E_{t+1} \leq E_t - \alpha \gamma^i \eta \sqrt{E_t} + \beta \left(\gamma^i \eta \right)^2. \tag{29}$$

Hence, by using

$$E^*_{i+1} := \frac{\beta^2}{\alpha^2} \left(\gamma^i \eta \right)^2,$$

Lemma 3 ensures

$$\begin{cases} E_{t+1} \leq E_t & \text{if } E_t \geq E^*_{i+1} \\ E_{t+1} \leq \max\left\{ \beta, \frac{\beta^2}{\alpha^2} \right\} \left(\gamma^i \eta \right)^2 & \text{if } E_t < E^*_{i+1} \end{cases}. \tag{30}$$

for all $t \in [i \cdot N, (i+1) \cdot N]$.

We now consider two exhaustive cases.

Case 1. $E_t > E^*_{i+1}$ holds for all $t \in [i \cdot N,\ (i+1) \cdot N]$. Subtracting E^*_{i+1} from both sides of (29), for any $t \in [i \cdot N, (i+1) \cdot N - 1]$,

$$\begin{aligned} E_{t+1} - E^*_{i+1} &\leq \left(1 - \frac{\gamma^i \eta \alpha}{\sqrt{E_t} + \sqrt{E^*_{i+1}}} \right) (E_t - E^*_{i+1}) \\ &\overset{(30)}{\leq} \left(1 - \frac{\gamma^i \eta \alpha}{\sqrt{E_{i \cdot N}} + \sqrt{E^*_{i+1}}} \right) (E_t - E^*_{i+1}). \end{aligned}$$

Substituting the induction hypothesis $E_{i \cdot N} \leq \zeta \left(\gamma^{i-1} \eta \right)^2$ and $E^*_{i+1} = \frac{\beta^2}{\alpha^2} \left(\gamma^i \eta \right)^2$,

$$E_{t+1} - \frac{\beta^2}{\alpha^2} \left(\gamma^i \eta \right)^2 \leq \left(1 - \frac{\gamma \alpha}{\sqrt{\zeta} + \frac{\beta}{\alpha} \gamma} \right) \left(E_t - \frac{\beta^2}{\alpha^2} \left(\gamma^i \eta \right)^2 \right).$$

Introducing $q(\gamma) := 1 - \frac{\gamma \alpha}{\sqrt{\zeta} + \frac{\beta}{\alpha} \gamma}$, we obtain

$$\begin{aligned} E_{(i+1) \cdot N} &\leq \frac{\beta^2}{\alpha^2} \left(\gamma^i \eta \right)^2 + q(\gamma)^N \left(E_{i \cdot N} - \frac{\beta^2}{\alpha^2} \left(\gamma^i \eta \right)^2 \right) \\ &\leq \left(\left(1 - q(\gamma)^N \right) \frac{\beta^2}{\alpha^2} + \zeta \frac{q(\gamma)^N}{\gamma^2} \right) \left(\gamma^i \eta \right)^2, \end{aligned}$$

where the last inequality follows from the induction hypothesis $E_{i \cdot N} \leq \zeta \left(\gamma^{i-1} \eta \right)^2$. Note that $q(\gamma) \in (0, 1)$. $q(\gamma) < 1$ follows immediately from $\frac{\gamma \alpha}{\sqrt{\zeta} + \frac{\beta}{\alpha} \gamma} > 0$. $q(\gamma) > 0$ follows from $E_{t+1} - E^*_{i+1}, E_t - E^*_{i+1} > 0$ for all $t \in [i \cdot N, (i+1) \cdot N - 1]$.

Therefore, if

$$h : \gamma \mapsto \left(1 - q(\gamma)^N \right) \frac{\beta^2}{\alpha^2} + \zeta \frac{q(\gamma)^N}{\gamma^2} - \zeta$$

satisfies $h(\gamma) \leq 0$, the induction hypothesis is proved. Since $h(\gamma)$ is continuous on $(0, 1]$ and

$$h(1) = \left(1 - q(1)^N \right) \left(\frac{\beta^2}{\alpha^2} - \zeta \right) < 0,$$

there exists $\hat{\gamma} \in (0,1)$ such that $h(\gamma) \leq 0$ for all $\gamma \in (\hat{\gamma}, 1)$.

Case 2. $E_t \leq E^*_{i+1}$ holds at some $t \in [i \cdot N, (i+1) \cdot N]$.

By (30) ,

$$\begin{aligned} E_{(i+1)N} &\leq \max\left\{E_{(i+1)N-1}, \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\}\left(\gamma^i\eta\right)^2\right\} \\ &\leq \max\left\{E_t, \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\}\left(\gamma^i\eta\right)^2\right\} \\ &\leq \max\left\{E^*_{i+1}, \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\}\left(\gamma^i\eta\right)^2\right\} \\ &= \max\left\{\beta, \frac{\beta^2}{\alpha^2}\right\}\left(\gamma^i\eta\right)^2 \\ &\leq \zeta\left(\gamma^i\eta\right)^2. \end{aligned}$$

Thus, (28) holds for $T = i+1$.

Therefore, combining the two cases, we obtain

$$E_{T\cdot N} \leq \zeta\left(\gamma^{T-1}\eta\right)^2$$

for all $T \geq 1$ by induction.

Finally, following exactly the same argument as in the proof of Theorem 5, we obtain

$$\begin{aligned} f(x_{T\cdot N}) - \min_{x\in\mathbb{R}^d} f(x) &\leq G\sqrt{\frac{2}{\mu}\left(\phi(x_{T\cdot N}) - \min_{x\in\mathbb{R}^d}\phi(x)\right)} \\ &\leq G\sqrt{\frac{2\zeta}{\mu}\left(\gamma^{T-1}\eta\right)^2} \\ &= q\gamma^{T-1}, \end{aligned}$$

where $q := G\sqrt{\frac{2\zeta}{\mu}}\eta$. □

Although Theorem 6 chooses γ sufficiently close to 1 after fixing the stage length N, the reverse is also admissible, i.e., for any $\gamma \in (0,1)$, sufficiently large N achieves the same rate in Theorem 6. This is shown by using the following argument

$$h(N) := \left(1 - q(\gamma)^N\right)\frac{\beta^2}{\alpha^2} + \zeta\frac{q(\gamma)^N}{\gamma^2} - \zeta \to \frac{\beta^2}{\alpha^2} - \zeta < 0, \qquad N \to \infty$$

in Case 1 in the proof of Theorem 6.

Remark 2. *Lemma 3 and Theorem 5 and 6 provide theoretical insight into the empirical observation reported in [15] that geometrically decaying step sizes with a longer first stage, given by*

$$\eta_t = \begin{cases} \eta & \text{if } t < T_0 \\ \eta \cdot \gamma^{\left\lceil \frac{t+1-T_0}{N} \right\rceil} & \text{if } t \geq T_0 \end{cases}$$

achieve faster convergence than standard geometrically decaying step sizes. By Lemma 3, $\frac{\beta^2}{\alpha^2}$ does not decrease as $\phi(x_0) - \min_x \phi(x)$ increases. Therefore, for simplicity, suppose that

$$\zeta := \max\left\{\frac{q_1\left(q_2+\eta^2\right)}{\eta^2},\ \frac{\beta^2}{\alpha^2}+1,\ \beta\right\} = \frac{\beta^2}{\alpha^2}+1.$$

Then, we obtain

$$h(\gamma) = \left(1-q(\gamma)^N\right)\frac{\beta^2}{\alpha^2} + \zeta\left(\frac{q(\gamma)^N}{\gamma^2}-1\right) = q(\gamma)^N\left(\frac{\beta^2}{\alpha^2}\left(\frac{1}{\gamma^2}-1\right)+\frac{1}{\gamma^2}\right)-1.$$

Since $\gamma \in (0,1)$, $h(\gamma)$ is monotonically increasing as $\frac{\beta^2}{\alpha^2}$ increases. Consequently, as the initial optimality gap of ϕ increases, the valid interval of γ in Theorem 6 shrinks toward 1, resulting in a worse coefficient of the linear rate. In contrast, by Theorem 5, the longer first stage can be regarded as a way of searching for a good initial point with small optimality gap of ϕ, so that a smaller γ is admissible in Theorem 6.

In Theorems 5 and 6, the assumption that f admits a $(1, C, L)$-smooth surrogate ϕ plays an essential role in their proof. We next show that this condition is satisfied when f is weakly convex, coercive, and locally PL, and satisfies Assumption 3. Specifically, Assumption 3 imposes the standard sharpness condition, which is assumed in [9, 16, 17]. As a consequence, Theorem 5 can be specialized to weakly convex functions satisfying Assumption 3 in Corollary 5.

Assumption 3. *Let $f^* := \min_{x\in\mathbb{R}^d} f(x)$ and $X^* := \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. There exists $\underline{G}, \sigma > 0$ such that*

$$\inf_{g\in\partial f(x)} \|g\| \geq \underline{G}, \qquad \forall x \notin X^* \tag{31}$$

and

$$f(x) - f^* \geq \sigma \cdot \inf_{x^*\in X^*} \|x - x^*\|, \qquad \forall x \in \mathbb{R}^d. \tag{32}$$

Corollary 5. *Suppose that f is ρ-weakly convex, coercive, and locally PL, and satisfies Assumption 3. For any initial point x_0, there exists $q_1 > 0$ such that the subgradient method with a sufficiently small constant step size η satisfies*

$$f(x_T) - \min_{x\in\mathbb{R}^d} f(x) \leq q_1\left(q_2(\eta)^T + \eta\right),$$

for all $T \geq 0$, where $q_2(\eta) = 1 - \Theta(\eta) \in (0,1)$.

Proof By Lemma 2(iv)–(v), the Moreau envelope f_λ is coercive and locally PL. Furthermore, Lemma 2(iii) guarantees that (3) holds, while Lemma 2(i) establishes that f_λ is L_λ-smooth with $L_\lambda := \max\left\{\frac{1}{\lambda}, \frac{\rho}{1-\lambda\rho}\right\}$.

To invoke Theorem 5, it suffices to establish the existence of a constant $C > 0$ such that

$$\langle g_x, \nabla f_\lambda(x) \rangle \geq C \|\nabla f_\lambda(x)\|, \qquad \forall x \in \mathbb{R}^d,\ \forall g_x \in \partial f(x). \tag{33}$$

Once (33) is established, f_λ serves as a $(1, C, L_\lambda)$-smooth surrogate for f that is coercive and locally PL, whereupon the assertion of the corollary follows directly from Theorem 5.

We establish (33) by considering two exhaustive cases regarding the proximal point $p_\lambda(x)$.

Case 1: $p_\lambda(x) \notin X^*$. By Proposition 1 together with the subgradient identity

$$\nabla f_\lambda(x) = \frac{x - p_\lambda(x)}{\lambda} \in \partial f(p_\lambda(x)) \tag{34}$$

from Lemma 2(iii), we have

$$\langle g_x, \nabla f_\lambda(x) \rangle \geq (1 - \lambda\rho) \inf_{g \in \partial f(p_\lambda(x))} \|g\| \|\nabla f_\lambda(x)\|, \qquad \forall g_x \in \partial f(x).$$

Since $p_\lambda(x) \notin X^*$, applying (31) from Assumption 3 yields

$$\langle g_x, \nabla f_\lambda(x) \rangle \geq (1 - \lambda\rho)\underline{G} \|\nabla f_\lambda(x)\|, \qquad \forall g_x \in \partial f(x).$$

Case 2: $p_\lambda(x) \in X^*$. Fix an arbitrary $x^* \in X^*$. By the definition of the Moreau envelope,

$$f(p_\lambda(x)) + \frac{1}{2\lambda}\|x - p_\lambda(x)\|^2 \leq f(x^*) + \frac{1}{2\lambda}\|x - x^*\|^2.$$

Since $p_\lambda(x) \in X^*$, we have $f(p_\lambda(x)) = f(x^*)$, which simplifies the inequality to $\|x - p_\lambda(x)\| \leq \|x - x^*\|$. Taking the infimum over $x^* \in X^*$ yields

$$\|x - p_\lambda(x)\| = \inf_{x^* \in X^*} \|x - x^*\|. \tag{35}$$

Furthermore, coercivity of f implies that the minimizer set X^* is bounded. Because f is locally Lipschitz continuous, subgradients on X^* are bounded, ensuring the existence of $\overline{G} \geq 0$ such that

$$\|x - p_\lambda(x)\| \leq \lambda\overline{G} \tag{36}$$

for all x satisfying $p_\lambda(x) \in X^*$.

Now, invoking the ρ-weak convexity subgradient inequality at x, given by

$$\langle g_x, x - p_\lambda(x) \rangle \geq f(x) - f(p_\lambda(x)) - \frac{\rho}{2}\|x - p_\lambda(x)\|^2, \qquad \forall g_x \in \partial f(x),$$

and substituting $\nabla f_\lambda(x) = \frac{x - p_\lambda(x)}{\lambda}$, which follows from (34), we deduce

$$\begin{aligned}
\langle g_x, \nabla f_\lambda(x) \rangle &\geq \frac{f(x) - f^*}{\lambda} - \frac{\rho}{2\lambda}\|x - p_\lambda(x)\|^2 \\
&\overset{(32)}{\geq} \frac{\sigma}{\lambda} \inf_{x^* \in X^*} \|x - x^*\| - \frac{\rho}{2\lambda}\|x - p_\lambda(x)\|^2 \\
&\overset{(35)}{=} \frac{\sigma}{\lambda}\|x - p_\lambda(x)\| - \frac{\rho}{2\lambda}\|x - p_\lambda(x)\|^2 \\
&\overset{(36)}{\geq} \left(\frac{\sigma}{\lambda} - \frac{\rho\overline{G}}{2}\right)\|x - p_\lambda(x)\| \\
&\overset{(34)}{\geq} \left(\sigma - \frac{\lambda\rho\overline{G}}{2}\right)\|\nabla f_\lambda(x)\|.
\end{aligned}$$

Combining Case 1 and 2, inequality (33) holds for all $x \in \mathbb{R}^d$ and $g_x \in \partial f(x)$ with

$$C = \min\left\{(1 - \lambda\rho)\underline{G},\ \sigma - \frac{\lambda\rho\overline{G}}{2}\right\} > 0,$$

provided $\lambda > 0$ is chosen sufficiently small. This completes the proof. □

Similarly, Theorem 6 is specialized to weakly convex functions in Corollary 6.

Corollary 6. *Suppose that f is ρ-weakly convex, coercive, and locally PL, and satisfies Assumption 3. For any initial point $x_0 \in \mathbb{R}^d$ and any $N \in \mathbb{N}$, the subgradient method with a geometrically decaying step size $\eta_t = \eta \cdot \gamma^{\left\lceil \frac{t-N+1}{N} \right\rceil}$ for a sufficiently small $\eta > 0$ and γ sufficiently close to 1 satisfies*

$$f(x_{T\cdot N}) - \min_{x\in\mathbb{R}^d} f(x) \leq q\gamma^{T-1}$$

for all $T > 0$, where $q > 0$ is a constant independent of T.

Proof The claim immediately follows from Theorem 6 and Corollary 5. □

Assumption 3 impose sharp structure on f in two ways: Condition (31) is about lower bound of gradient norm and condition (32) is about the function value. Specifically, condition (32) is widely known as sharpness and same to that assumed in the prior works [9, 16, 17].

As an example of a locally sharp weakly convex function, consider $f(x) = |x|$. It is easily checked that f admits $(1, 1, 2)$-smooth surrogate $\phi(x) = x^2$ as

$$f'(x)\phi'(x) = 2x\,\mathrm{sign}(x) = 2|x| = |\phi'(x)|, \qquad x \neq 0.$$

Since f satisfies the assumptions in Corollary 6, a geometrically decaying step size achieves linear convergence. In contrast, Proposition 3 shows that the polynomially decaying step size can fail to achieve linear convergence for this loss function. Combining this with the discussion in Remark 1, a polynomially decaying step size can be suboptimal in both the locally smooth and locally sharp settings.

Proposition 3. *Let $f(x) = |x|$. Consider the subgradient method for f with a polynomially decaying step size $\eta_t = \frac{\eta}{(t+1)^\alpha}$ for $\eta > 0$ and $\alpha > 0$. Then there exists an initial point $x_0 \in \mathbb{R}$ such that*

$$f(x_T) - \min_{x\in\mathbb{R}} f(x) \geq \frac{1}{2}\eta_T$$

for all $T \geq 0$.

Proof The map $s \mapsto (s+1)^{-\alpha}$ is convex on $(-1, \infty)$. Therefore, the sequence $(\eta_t)_{t\geq 0}$ is convex, strictly decreasing, and $\eta_t \to 0$. In particular, for any fixed t, the differences $d_j := \eta_{t+j} - \eta_{t+j+1}$ are positive and non-increasing in j.

Since $\eta_{t+j} \to 0$, the alternating series

$$a_t := \sum_{j=0}^{\infty} (-1)^j \eta_{t+j}$$

converges for every $t \geq 0$. The sequence $(a_t)_{t\geq 0}$ satisfies

$$a_t + a_{t+1} = \eta_t \tag{37}$$

and

$$a_t \geq \frac{1}{2}\eta_t. \tag{38}$$

(37) follows by re-indexing $a_{t+1} = -\sum_{k\geq 1}(-1)^k \eta_{t+k}$. (38) follows as

$$\eta_t = \sum_{j\geq 0}\left(\eta_{t+j} - \eta_{t+j+1}\right) = \sum_{i\geq 0} d_{2i} + \sum_{i\geq 0} d_{2i+1} \leq 2a_t.$$

The inequality follows since $(d_j)_{j\geq 0}$ is non-increasing and $\sum_{i\geq 0} d_{2i+1} \leq \sum_{i\geq 0} d_{2i} = a_t$.

Set $x_t := (-1)^t a_t$. By (38) and $\eta_t > 0$, $x_t \neq 0$ and $\operatorname{sign}(x_t) = (-1)^t$. Moreover, by (37),

$$x_t - \eta_t f'(x_t) = x_t - \eta_t \operatorname{sign}(x_t) = (-1)^t(a_t - \eta_t) = (-1)^t(-a_{t+1}) = x_{t+1}.$$

Therefore, $(x_t)_{t\geq 0}$ is the sequence of the subgradient method with the initial point $x_0 = a_0$ and step size sequence $\eta_t = \frac{\eta}{(t+1)^\alpha}$.

Thus, we obtain

$$f(x_T) - \min_{x\in\mathbb{R}} f(x) = |x_T| = a_T \geq \frac{1}{2}\eta_T$$

for all $T \geq 0$. □

6 Numerical Experiments

In this section, we demonstrate that the theoretical convergence bounds established in Section 5 closely align with the experimental results, suggesting the existence of the smooth surrogate in the sense of Definition 2, even when its existence is difficult to verify explicitly.

6.1 Locally Smooth Case

For visualization, we consider the squared loss of a single-neuron neural network with Leaky ReLU activation parameterized by $w, b \in \mathbb{R}$.

$$f(w,b) = \sum_{i=1}^{m} \left(\max\left\{\alpha(w\cdot x_i + b),\ w\cdot x_i + b\right\} - y_i\right)^2. \tag{39}$$

$(x_i)_{i=1}^m \in \mathbb{R}^m$ are the inputs and $(y_i)_{i=1}^m \in \mathbb{R}^m$ are the corresponding labels, generated by a teacher model of the form $y_i = \max\left\{\alpha(w^*\cdot x_i + b^*),\ w^*\cdot x_i + b^*\right\}$. Here, (w^*, b^*) denotes a global minimizer of $f(w,b)$. If $w^*\cdot x_i + b^* \neq 0$ holds for all i, then $f(w,b)$ is locally quadratic around (w^*, b^*), and therefore satisfies the locally smooth condition given by Assumption 1. Figure 1 illustrates this scenario. The left panel shows a trajectory of the subgradient method with a constant step size. Since the trajectory passes through a set of non-differentiable points, the classical analyses for smooth optimization do not apply. The right panel demonstrates that the subgradient method with a constant step size achieves linear convergence, owing to the local smoothness of f, consistent with Theorem 4.

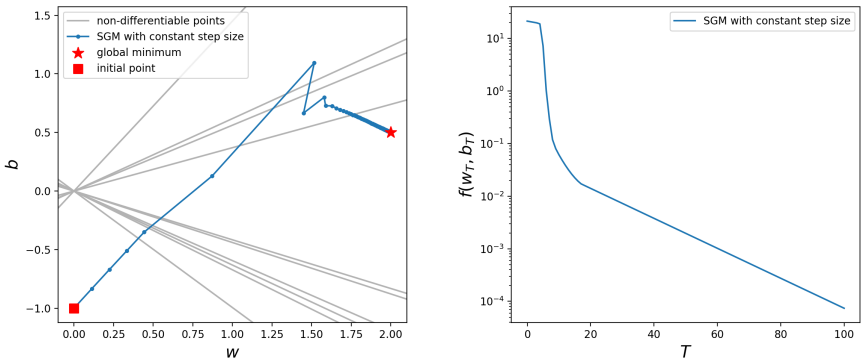


Fig. 1 [Left] Trajectory of the subgradient method (SGM) in the paramter space, along with the initial point, global minimum, and set of non-differentiable points. The parameters of $f(w, b)$ are set as $m = 10$, $\alpha = 0.1$, and $(w^*, b^*) = (2, 0.5)$. $\{x_i\}_{i=1}^m$ are independently sampled from the standard normal distribution. The hyperparameters of the subgradient method with constant step size are configured as $\eta = 0.075$ and $(w_0, b_0) = (0, -1)$. **[Right] Transition of the loss for the subgradient method (SGM) with constant step size.** The parameter settings are identical to those used in the left figure.

6.2 Locally Sharp Case

As an example of a loss function with a non-differentiable local minimum, we consider the l_1-regularized squared loss of a single-neuron neural network with Leaky ReLU activation, parameterized by $w, b \in \mathbb{R}$.

$$f(w, b; \lambda) = \sum_{i=1}^{m} \left(\max\left\{\alpha(w \cdot x_i + b),\ w \cdot x_i + b\right\} - y_i\right)^2 + \lambda(|w| + |b|). \tag{40}$$

We numerically demonstrate the advantage of geometrically decaying step sizes over polynomially decaying step sizes, as suggested by Theorem 6 and Proposition 3. The left panel of Figure 2 illustrates that both step sizes approach the same local minimizer, which is a non-differentiable point. The right panel of Figure 2 demonstrates that a geometrically decaying step size achieves linear convergence in line with Theorem 6, whereas a polynomially decaying step size exhibits slower convergence, as predicted by Proposition 3.

We finally numerically demonstrate the advantage of geometrically decaying step sizes with a longer first stage over the standard geometrically decaying step sizes, as suggested by Remark 2. The left panel of Figure 3 represents the schedules of three different geometrically decaying step sizes of the form

$$\eta_t = \begin{cases} \eta & \text{if } t < T_0 \\ \eta \cdot \gamma^{\left\lceil \frac{t+1-T_0}{N} \right\rceil} & \text{if } t \geq T_0 \end{cases},$$

which share the same η_0 and γ but use different values of T_0 and N. The right panel of Figure 3 represents the transition of the loss under these step sizes. The blue line,

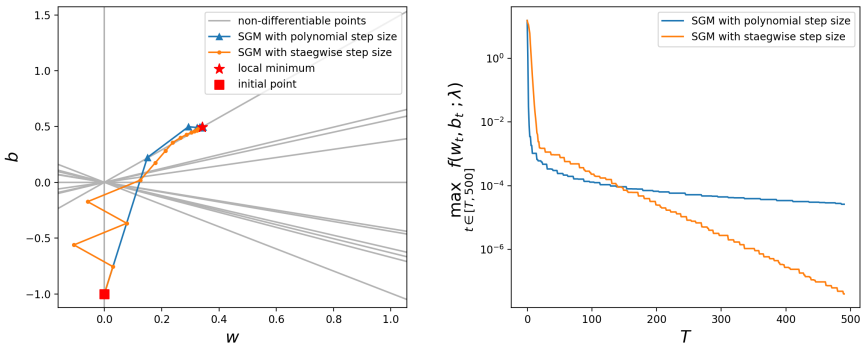


Fig. 2 [Left] Trajectories of the subgradient method (SGM) with polynomially and geometrically decaying step sizes in the parameter space, along with the initial point, global minimum, and set of non-differentiable points. The parameters of $f(w, b; \lambda)$ are set as $m = 10$, $\alpha = 0.1$, $\lambda = 10$ and $(w^*, b^*) = (2, 0.5)$. $\{x_i\}_{i=1}^m$ are independently sampled from the standard normal distribution. For the subgradient method with polynomially decaying step size, the hyperparameter is set as $\eta_0 = 0.1$. For the subgradient method with geometrically decaying step size, the hyperparameters are set as $\eta_0 = 0.02$ and $T_0 = N = 10$. Both two algorithms share the same initialization $(w_0, b_0) = (-2, -2)$. **[Right] Transition of the loss for the subgradient method (SGM) with geometrically decaying step size and polynomially decaying step size.** The parameter settings are identical to those used in the left figure.

corresponding to a geometrically decaying step-size schedule with a small stage length N, fails to converge to a local minimum because the decay factor γ is too small relative to N. The orange line, corresponding to a geometrically decaying step-size schedule with a longer stage length, converges slowly to a local minimum because the decay factor γ is too small relative N. These observations align with Theorem 6. The green line, employing a longer initial stage T_0 and a shorter subsequent stage length N, achieves faster convergence than the orange line, as suggested by Remark 2.

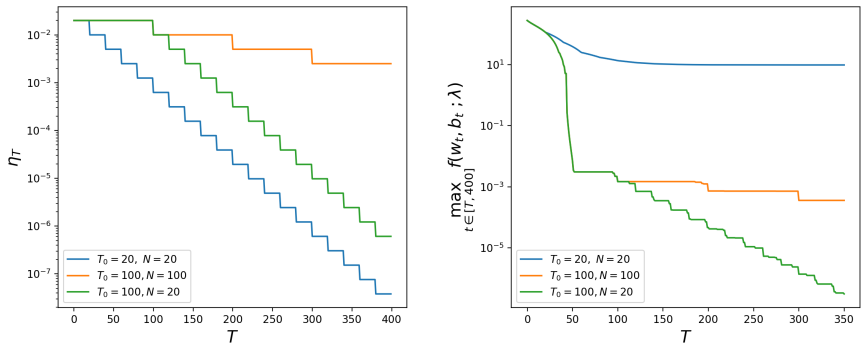


Fig. 3 [Left] Schedules of different geometrically decaying step sizes. Three geometrically decaying step sizes share the same initial step size $\eta_0 = 0.02$ and decaying factor $\gamma = 0.5$, but differ in the initial stage length T_0 and subsequent stage length N. **[Right]: Transition of the loss under different geometrically decaying step sizes for the subgradient method (SGM).** The parameters of $f(w, b; \lambda)$ are set as $m = 10$, $\alpha = 0.1$, $\lambda = 10$, and $(w^*, b^*) = (2, 0.5)$. $\{x_i\}_{i=1}^m$ are independently sampled from the standard normal distribution. The subgradient method is applied separately with each step size from the left panel, sharing the same initialization $(w_0, b_0) = (-2, -2)$.

7 Conclusion

In this work, we introduced the class of functions that strictly contains the class of weakly convex functions. For this broader class, we established convergence bounds across non-structured, locally smooth, and locally sharp settings. Furthermore, we provided explicit specializations to weakly convex functions, recovering results similar to those in [9, 12]. These convergence guarantees yield key insights into step-size selection: (i) a polynomially decaying step size is one of the most robust choices for guaranteeing convergence in the non-structured setting; (ii) a polynomially decaying step size can be suboptimal in both locally smooth and locally sharp settings; and (iii) in the locally sharp setting, geometrically decaying step sizes with an extended first stage can achieve faster convergence than standard geometrically decaying schedules. These observations align with existing empirical findings in [15, 31–33] and are further validated by our numerical experiments.

A limitation of our current framework is that verifying the existence of a smooth surrogate remains challenging in general. Developing a systematic methodology for constructing smooth surrogates for non-weakly convex functions constitutes a promising direction to broaden the practical applicability of our framework.

8 Declarations

Funding

The authors did not receive support from any organization for the submitted work.

Competing Interests:

The authors have no relevant financial or non-financial interests to disclose.

References

[1] Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M.: Automatic differentiation in machine learning: A survey. J. Mach. Learn. Res. **18**(153), 1–43 (2018)

[2] Park, S., Chun, S., Lee, W.: What does automatic differentiation compute for neural networks? In: The Twelfth International Conference on Learning Representations (2024)

[3] Clarke, F.H.: Generalized gradients and applications. Trans. Amer. Math. Soc. **205**, 247–262 (1975) https://doi.org/10.1090/S0002-9947-1975-0367131-6

[4] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS 2017 Workshop on Autodiff (2017)

[5] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, pp. 265–283. USENIX Association, USA (2016)

[6] Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, 1st edn. Springer, New York, NY (2014). https://doi.org/10.1007/978-1-4419-8853-9

[7] Bubeck, S.: Convex optimization: Algorithms and complexity. FFound. Trends Mach. Learn. **8**(3–4), 231–357 (2015) https://doi.org/10.1561/2200000050

[8] Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019) https://doi.org/10.1137/18M1178244

[9] Davis, D., Drusvyatskiy, D., Macphee, K.J., Paquette, C.: Subgradient methods for sharp weakly convex functions. J. Optim. Theory Appl. **179**(3), 962–982 (2018) https://doi.org/10.1007/s10957-018-1372-8

[10] Duchi, J.C., Ruan, F.: Stochastic methods for composite and weakly convex optimization problems. SIAM J. Optim. **28**(4), 3229–3259 (2018) https://doi.org/10.1137/18M1172472

[11] Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J.D.: Stochastic subgradient method converges on tame functions. Found. Comput. Math. **20**(1), 119–154 (2020) https://doi.org/10.1007/s10208-018-09409-5

[12] Li, X., Zhao, L., Zhu, D., So, A.M.-C.: Revisiting subgradient method: Complexity

and convergence beyond lipschitz continuity. Vietnam J. Math. **53**(4), 735–755 (2025) https://doi.org/10.1007/s10013-024-00715-w

[13] Bolte, J., Pauwels, E.: A mathematical model for automatic differentiation in machine learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Curran Associates Inc., Red Hook, NY, USA (2020)

[14] Moreau, J.-J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. Fr. **93**, 273–299 (1965) https://doi.org/10.24033/bsmf.1625

[15] Wu, J., Zou, D., Braverman, V., Gu, Q., Kakade, S.M.: Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 24280–24314 (2022)

[16] Burke, J.V., Ferris, M.C.: Weak sharp minima in math. program. SIAM J. Control Optim. **31**(5), 1340–1359 (1993) https://doi.org/10.1137/0331063

[17] Charisopoulos, V., Davis, D.: A superlinearly convergent subgradient method for sharp semismooth problems. Math. Oper. Res. **49**(3), 1737–1770 (2024) https://doi.org/10.1287/moor.2023.1390

[18] Bolte, J., Pauwels, E.: Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. Math. Program. **188**, 19–51 (2021)

[19] Zhang, J., Lin, H., Sra, S., Jadbabaie, A.: First-order algorithms for non-smooth non-convex optimization in Goldstein stationarity. SIAM J. Optim. **32**(3), 1900–1928 (2022)

[20] Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J.D.: Subgradient methods for generic Lipschitz optimization. Math. Oper. Res. **47**(2), 1000–1025 (2022)

[21] Grimmer, B., Jia, Z.: Goldstein stationarity in Lipschitz constrained optimization. arXiv preprint arXiv:2310.03690 (2023)

[22] Kong, W., Lewis, A.S.: The cost of nonconvexity in deterministic nonsmooth optimization. Math. Oper. Res. **48**(4), 2051–2072 (2023)

[23] Tian, C., Zhou, Y., So, A.M.-C.: Finding near-approximate stationary points in non-smooth non-convex optimization. In: Proceedings of the 39th International Conference on Machine Learning (ICML), pp. 21340–21361 (2022)

[24] Zhang, P., Cutkosky, A.: Distributional stationarity and randomized algorithms for nonsmooth nonconvex optimization. In: Advances in Neural Information

Processing Systems (NeurIPS), vol. 37 (2024)

[25] Kornowski, G.: On the complexity of finding stationary points in nonsmooth nonconvex optimization. arXiv preprint arXiv:2609.17780 (2026)

[26] Beck, A.: First-Order Methods in Optimization. MOS-SIAM Series on Optimization, vol. 25. Society for Industrial and Applied Mathematics, Philadelphia, PA (2017). https://doi.org/10.1137/1.9781611974997

[27] Daniilidis, A., Burachik, R.S., Lima, P.: Filling the gap between lower-C^1 and lower-C^2 functions. J. Convex Anal. **12**(2), 485–508 (2005)

[28] Renaud, M., Leclaire, A., Papadakis, N.: On the moreau envelope properties of weakly convex functions. arXiv preprint arXiv:2509.13960 (2025)

[29] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, NJ (1970)

[30] Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 9851, pp. 795–811. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46128-1_50

[31] Yuan, Z., Yan, Y., Jin, R., Yang, T.: Stagewise training accelerates convergence of testing error over SGD. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 2604–2614 (2019)

[32] Ge, R., Kakade, S.M., Kidambi, R., Netrapalli, P.: The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 14951–14962 (2019)

[33] Wang, X., Magnússon, S., Johansson, M.: On the convergence of step decay step-size for stochastic optimization. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 14226–14238 (2021)