

Linearly Convergent Algorithms for Nonsmooth Problems with Unknown Smooth Pieces

Zhe Zhang · Suvrit Sra

Received: date / Accepted: date

Abstract We develop efficient algorithms for optimizing piecewise smooth (PWS) functions where the underlying partition of the domain into smooth pieces is *unknown*. For PWS functions satisfying a quadratic growth (QG) condition, we propose a bundle-level (BL) type method [25] that achieves global linear convergence—to our knowledge, the first such result for any algorithm for this problem class. We extend this method to handle approximately PWS functions and to solve weakly-convex PWS problems, improving the state-of-the-art complexity to match the benchmark for smooth non-convex optimization. Furthermore, we introduce the first verifiable and accurate termination criterion for PWS optimization. Similar to the gradient norm in smooth optimization, this certificate tightly characterizes the optimality gap under the QG condition, and can moreover be evaluated without knowledge of any problem parameters. We develop a search subroutine for this certificate and embed it within a guess-and-check framework, resulting in an almost parameter-free algorithm for both the convex QG and weakly-convex settings.

1 Introduction

Non-smoothness is a major bottleneck in optimization. In its presence, theoretical convergence rates plummet from linear to sublinear for convex problems under the QG condition [21, 35, 34], and degrade by orders of magnitude for non-convex problems [43, 5]. Crucially, for general non-smooth functions, this performance gap is fundamentally unavoidable [34, 35].

In this paper, we study a specific, widely applicable class of non-smooth functions: piecewise smooth (PWS) functions. A function is PWS if its domain can be partitioned into a finite number of subsets (pieces) such that restricted to each piece (see Definition 1) the function is smooth. Clearly, PWS functions present a more structured form of non-smoothness compared to the general case, as their gradients are continuous within the interior of each piece, and non-differentiability occurs only on the lower-dimensional boundaries between pieces (that have measure zero).

The PWS structure appears in many important applications, including statistics (e.g., elastic net regularization), signal processing (e.g., compressive sensing and phase retrieval), economics (e.g., matrix games), control (e.g., multiparametric programming), and machine learning (e.g., the ReLU activation function). A key challenge arises in many practical scenarios where the underlying pieces comprising the PWS objective function are too complicated or unknown *a priori*. This lack of knowledge about the function’s specific structure raises the following central research question:

Can one optimize PWS functions with unknown pieces using almost the same oracle complexity as that of smooth optimization?

Z. Zhang
School of Industrial Engineering, Purdue University, USA
E-mail: zhan5111@purdue.edu

S. Sra
Department of Mathematics, CIT, TU Munich, Germany
E-mail: s.sra@tum.de

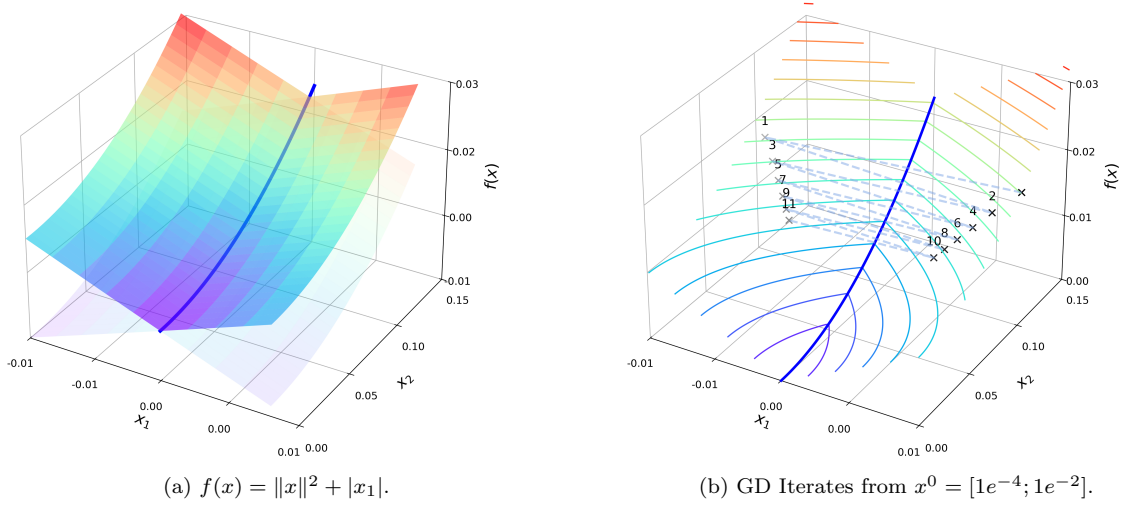


Fig. 1: Illustration of a Piecewise Smooth Function.

We answer this question by developing new algorithms for optimizing PWS problems of the form:

$$\min_{x \in X} f(x), \quad (1.1)$$

where the objective function f is PWS, and the feasible region X is closed, convex, and simple [3] (i.e., sets onto which projection is computationally efficient). An illustrative example of such a PWS function is given by

$$f(x) = \|x\|^2 + |x_1|, \quad x \in \mathbb{R}^2. \quad (1.2)$$

As plotted in Figure 1a, the non-smooth component $|x_1|$ divides the domain \mathbb{R}^2 into two regions, $X_1 := (-\infty, 0] \times \mathbb{R}$ and $X_2 := (0, \infty) \times \mathbb{R}$. Observe that f is differentiable everywhere except on the boundary between these regions (the x_2 -axis, where $x_1 = 0$), and the gradient $\nabla f(x)$ is continuous within the interior of each piece.

Despite the structure of PWS functions, there has been limited success in designing algorithms capable of achieving “almost smooth performance” for optimizing them, especially when the pieces are unknown. Specifically, for iterative first-order (FO) methods, fast convergence typically relies on computing the next iterate based on accurate FO information obtained at the current iterate. However, obtaining such information is challenging for PWS functions with unknown pieces: the FO information (e.g., the gradient) generated at a given iterate is accurate only within the smooth piece containing that iterate. Ensuring this information remains relevant for determining the next iterate is difficult when the iterates jump between different unknown smooth pieces.

More concretely, the standard gradient descent (GD) method can perform poorly even on the simple example in (1.2). Since the GD update relies solely on the gradient at the previous iterate, it offers little guarantee of progress when the current iterate lands in a different smooth piece than the previous one. This potential inefficiency is demonstrated by the characteristic zig-zagging pattern often observed in GD iterates for such problems, as illustrated in Figure 1b for iterates generated using the optimal Polyak stepsize. In contrast, methods that utilize information from multiple past iterates, such as the bundle-level (BL) method, can potentially overcome this limitation. The BL method uses first-order (FO) information from the m most recent iterates (via cuts, see (1.3) later). Thus, as long as the next iterate x^{t+1} lands on the same piece as any recent iterate x^{t-i} ($i \in \{0, \dots, m-1\}$), the cut generated at x^{t-i} provides relevant information, potentially guiding x^{t+1} toward making significant progress.

1.1 Overview of Existing approaches

Before detailing our approach and contributions, we first review existing approaches more broadly. In the literature, two main directions exist for tackling the PWS optimization problem (1.1), though neither approach fully achieves efficient global convergence when the function’s underlying PWS structure is unknown. First, when the PWS function admits a known max-of-smooth representation (conceptually illustrated by the shadow pieces in Figure 1a), the prox-linear update [33, 35, 20] can be employed. This method evaluates gradients on all constituent smooth functions at the current iterate, and uses all this information to generate the next iterate. It can achieve “almost smooth performance” because at least one of these gradients provides accurate information for the active (maximal) function near the next iterate. However, the prerequisite of requiring access to all component functions is far too strong in applications where pieces are unknown.

The second approach exploits the so-called $\mathcal{V}\mathcal{U}$ structure [31, 26, 14, 6] that characterizes the local geometry around an optimal solution. This structure decomposes into a smooth component (the \mathcal{U} -component) associated with an affine subspace, and a non-smooth component orthogonal to it (the \mathcal{V} -component). By analyzing projections of the iterates onto the \mathcal{U} -space, linear convergence can be achieved. However, the definition and identification of the \mathcal{U} -space rely on local information around the optimum. This limits these convergence guarantees to a local neighborhood, thus providing little insight into the global oracle complexity needed to answer our research question.

The approach taken in our paper proceeds by re-examining the classical bundle-level (BL) type methods, and affirmatively answers the posed research question. As motivated above and unlike the methods reviewed, BL methods do not require explicit knowledge of the pieces, and can be analyzed globally. Rather than measuring one-step progress, the key to our approach is a novel technique that analyzes multi-step progress. Under this new perspective, we prove that the BL method makes significant progress once two iterates (that need not be consecutive) land within the same piece. When the total number of pieces is finite, say k , (by the pigeonhole principle) this event is guaranteed to happen every $k + 1$ steps. Thus, the iterates make significant progress periodically, resulting in the desired linear convergence under the strongly convex setting.

1.2 Main Contributions

More concretely, for the PWS objective function in (1.1), the BL method generates the new iterate x^{t+1} by projecting the current iterate x^t onto a level set constructed from the preceding m cuts (linearizations based on past gradients):

$$\begin{aligned} x^{t+1} \leftarrow \arg \min_{x \in X} \frac{1}{2} \|x - x^t\|^2 \\ \text{s.t.} \quad f(x^{t-i}) + \langle \nabla f(x^{t-i}), x - x^{t-i} \rangle \leq l \quad \forall i \in \{0, 1, \dots, m-1\}. \end{aligned} \tag{1.3}$$

Crucially, this update (1.3) depends only on the algorithm parameters m (number of cuts) and l (the level value). It does not require explicit knowledge of the PWS function’s underlying pieces, making it directly applicable to our unknown PWS setting.

Based on this method and our novel multi-step analysis, this paper contributes to the complexity theory of nonsmooth PWS optimization along the following directions:

i). A New Perspective and Global Linear Convergence. When the objective function f is convex, satisfies the quadratic growth condition, and the optimal objective value f^* is known, we provide a simple proof showing that the single-loop BL method achieves a global linear convergence rate if the number of cuts, m , is greater than the total number of pieces k . Importantly, our analysis reveals that the complexity bound depends not on the global number of pieces k , but rather on how often nearby iterates land on the same piece. This occurs more frequently as the algorithm approaches a solution, suggesting faster convergence in practice.

ii). *Generalization and Local Linear Convergence.* To handle broader settings and relax assumptions on the oracle, we introduce a class of approximately PWS (apx-PWS) functions. In this setting, we show that the BL iterates achieve a rate of $O(1/t)$. More importantly, we prove that once the iterates enter a neighborhood of the solution, the algorithm accelerates to a *local linear convergence* rate if m exceeds the local number of pieces, k_{local} . This provides the first theoretical justification for the well-known superior empirical performance of bundle-level methods over gradient-descent type methods for this problem class.

iii). *Adaptive Methods and Application to Non-Convex Problems.* We address a critical drawback of the baseline BL method by removing the impractical requirement of knowing the optimal value f^* . We propose an adaptive method, μ -BL, that instead only requires the QG modulus μ —a much weaker assumption. This method achieves the same oracle complexity as the original BL method. We then apply this adaptive approach to the non-smooth, non-convex setting. By using μ -BL as a subroutine within an inexact proximal point framework, we improve the oracle complexity for finding an ϵ -Moreau stationary point from $O(1/\epsilon^4)$ down to $O(1/\epsilon^2)$, matching the optimal order for smooth non-convex optimization.

iv). *A Verifiable Stationarity Certificate and Parameter-Free Algorithms.* We introduce a novel \mathcal{W} -stationarity certificate that resolves the long-standing challenge of finding a verifiable and accurate termination criterion for PWS optimization. Similar to gradient norms in the smooth setting, our certificate provides an error bound, tightly characterizes the optimality gap under the QG condition, and can be computed without prior knowledge of problem parameters. We believe this is the first certificate for this problem class with these properties. Crucially, this certificate and its search subroutine provide the mechanism to exploit unknown growth conditions. By embedding our subroutine within a guess-and-check framework, we successfully design almost parameter-free BL algorithms for both convex QG and weakly-convex settings, a key practical advance.

1.3 Related Literature

Bundle Type Algorithms. The cutting plane method, introduced in the 1960s [17], constructs a piecewise linear model of the objective function using the maximum of linear approximations (cuts) derived from previous iterates, selecting the model’s minimizer as the next iterate. However, its practical performance can be unstable, as iterates may change drastically with each new cut incorporated into the model.

The bundle level (BL) method was introduced by Lemaréchal, Nesterov, and Nemirovski [25] primarily to address the stability issues inherent in the cutting plane method. To mitigate instability, the BL method ensures more controlled progress by selecting the next iterate as the projection of the current iterate onto a specific level set derived from the cutting plane model. Reference [25] established that this method achieves an $O(1/\epsilon^2)$ oracle complexity for optimizing general Lipschitz-continuous, non-smooth convex objective functions. Later, in his textbook [33], Nemirovski observed empirically that the BL method significantly outperformed subgradient descent on the MAXQUAD problem. Since MAXQUAD involves a piecewise smooth (PWS) objective, our development provides theoretical justification for Nemirovski’s observation.

The BL method has since been extended, for example, to incorporate non-Euclidean Bregman distances [4] and to achieve accelerated convergence rates [20]. More recently, [9] proposed a BL-type method for function-constrained optimization, demonstrating promising numerical results.

A closely related line of work is the proximal bundle method, introduced in the 1970s [24, 29, 42]. These methods enhance stability by adding a proximal regularization term (typically penalizing deviation from the current iterate) to the cutting plane model minimization subproblem. The proximal bundle method has been extended, for instance, to handle inexact oracles [8] and non-convex objectives [15, 7].

Complexity guarantees for the proximal bundle method have been analyzed in works such as [18, 12, 10, 27]. However, these complexity bounds, at best, match those of subgradient-type methods (e.g., $O(1/\epsilon^2)$ for the general non-smooth convex case) and typically fail to demonstrate theoretical benefits from leveraging multiple cuts simultaneously—a key factor often cited for the

method’s strong empirical performance. An intriguing research direction involves exploiting the so-called \mathcal{UV} structure around the optimal solution to achieve superlinear local convergence [31, 32]. However, as this structure is inherently local, the corresponding analysis is restricted to local convergence regimes.

Structured Non-smooth Optimization. To overcome the challenges of non-smoothness, significant effort has focused on exploiting specific problem structures when available. For instance, in sparse optimization, [2, 37] developed composite optimization methods adept at handling objectives combining a smooth term with a prox-friendly non-smooth regularizer. In areas like function-constrained optimization and risk minimization, methods based on the prox-linear operation [35, 20, 44, 23] (related to bundle methods) leverage structures involving the maximum of known smooth functions. Furthermore, if the Fenchel conjugate of the non-smooth function is accessible, Nesterov [36] proposed influential techniques for constructing a smooth approximation of the original function. This smoothing approach is closely related to concepts like the Moreau envelope [1] and randomized smoothing [13]. However, these approximation strategies typically cannot match the oracle complexity of direct smooth optimization, often due to the difficulty of the sub-problems involved or large Lipschitz constants associated with the smoothed surrogate functions.

More recently, interesting developments have focused on algorithms achieving rapid local convergence (e.g., linear or superlinear) by exploiting unknown local geometric structures near an optimal solution. These methods often rely on the existence of a manifold within a neighborhood of the solution on which the objective function behaves smoothly. Of particular importance is the \mathcal{VU} decomposition framework [31, 30], which covers the PWS functions considered in this paper as a special case.

Leveraging this structure, [31] combines the proximal bundle method with Newton-like updates on the smooth \mathcal{U} -subspace component to achieve local superlinear convergence. Recently, [14] proposed a simpler survey descent method specifically designed for PWS problems. In this method, each iteration generates a new batch of survey points by solving k quadratically constrained quadratic programs (QCQPs), ensuring that all local smooth pieces are represented in the “survey”. At a high level, these survey points track the local geometry of f , enabling the generation of informed descent directions to achieve local linear convergence. However, initializing such a method to guarantee all necessary pieces are surveyed from the beginning presents a key challenge.

Davis and Jiang [6] propose the normal tangent descent method. This approach aims to find the minimal-norm element of the Goldstein subgradient to serve as a reliable descent direction for achieving local linear convergence. The method handles a more general class of non-smooth problems satisfying a local \mathcal{VU} manifold decomposition condition. They design an efficient subroutine that exploits this local manifold structure to find the minimum-norm subgradient for descent.

Leveraging Growth Conditions. A growth condition of order p relates the optimality gap to the distance to the solution set X^* via the inequality $\mu \text{dist}^p(x, X^*) \leq f(x) - f^*$. This concept is closely related to error bounds [38] and the Polyak-Lojasiewicz (PL) and Kurdyka-Lojasiewicz (KL) conditions [39, 28, 19]. Two special cases are particularly important: the sharp growth condition ($p = 1$) and the quadratic growth (QG) condition ($p = 2$).

In his seminal work, Polyak showed that for smooth functions, the QG condition is sufficient for gradient descent to achieve a linear convergence rate, even without strong convexity [39]. For non-smooth functions, he later established that the sharp growth condition allows gradient descent with the Polyak stepsize to converge linearly [40]. More recent results have focused on exploiting these conditions in more adaptive ways or in more complex settings [41, 10, 11]. Our work focuses on the PWS non-smooth setting under the QG condition. We show that by re-examining the bundle-level method from a new perspective, we can improve the sublinear convergence rates implied by prior works to a linear one. Furthermore, our technique can be adapted to the sharp growth setting to achieve an even faster quadratic convergence rate.

1.4 Notation and Assumptions

Notation

- $[m] := \{0, 1, 2, \dots, m\}$ denotes the set of integers from 0 to m .
- $B(x; \delta) := \{y \in \mathbb{R}^n \mid \|y - x\| \leq \delta\}$ denotes the closed ball in \mathbb{R}^n centered at x with radius $\delta \geq 0$.
- $f'(x)$ denotes an arbitrary subgradient from the subdifferential $\partial f(x)$. If f is differentiable at x , $\nabla f(x)$ denotes its unique gradient.
- $l_f(x; y)$ denotes the linear model of f constructed at point y and evaluated at point x :

$$l_f(x; y) := f(y) + \langle f'(y), x - y \rangle.$$

Assumptions Throughout the paper, we make the following standard assumptions about the objective function f .

- We assume f is M -Lipschitz continuous for some $M \geq 0$, i.e., $|f(x) - f(y)| \leq M\|x - y\|$ for all $x, y \in X$.
- We assume f is bounded below, i.e., $f^* := \min_{x \in X} f(x) > -\infty$, and use Δ_f to denote the initial function value gap, i.e., $\Delta_f := f(x^0) - f^*$.
- We assume the set of minimizers, $\mathcal{X}^* := \arg \min_{x \in \mathcal{X}} f(x)$, is non-empty. We use x^* to denote an arbitrary optimal solution from this set.

2 The Warm-Up: When f^* Is Known

We illustrate the key idea for achieving global linear convergence under the simplest setting where the optimal function value f^* is available. Specifically, Subsection 2.1 provides the detailed setting for our discussion, Subsection 2.2 presents the bundle-level (BL) algorithm and discusses the key insight, and Subsection 2.4 furnishes the technical proofs.

2.1 The problem set-up

In this section, we assume the objective function f in (1.1) is convex, (k, L) -piecewise smooth, and satisfies the quadratic growth property with some modulus $\mu > 0$. The specific definitions for piecewise smoothness and the growth property are provided below.

Definition 1 (Piecewise Smoothness) We say a function $f : X \rightarrow \mathbb{R}$ is (k, L) -piecewise smooth if there exists a covering of its domain X by k sets (pieces) $\{X_i\}_{i=1}^k$ (i.e., $X \subseteq \bigcup_{i=1}^k X_i$) such that for each piece X_i , the restriction $f_i := f|_{X_i}$ is L -smooth for some $L > 0$. Specifically, we assume access to a first-order oracle $f'(x)$ (which could be the gradient $\nabla f(x)$ where f is differentiable, or a specific subgradient otherwise) such that the following inequality holds for all $i \in \{1, \dots, k\}$:

$$f(x) - f(\bar{x}) - \langle f'(\bar{x}), x - \bar{x} \rangle \leq \frac{L}{2} \|x - \bar{x}\|^2, \quad \forall x, \bar{x} \in X_i. \quad (2.1)$$

A few remarks are in order regarding Definition 1. This definition includes the k -max-of-smooth functions, $f(x) := \max_{i=1, \dots, k} \tilde{f}_i(x)$, as a special case, where the covering sets $X_i = \{x \in X \mid \tilde{f}_i(x) = f(x)\}$ can simply be chosen based on which component function is maximal. Furthermore, the first-order oracle $f'(x)$ required by (2.1) is slightly stronger than a standard subgradient oracle from convex analysis. While $f'(x)$ coincides with the standard gradient $\nabla f(x)$ at points where f is differentiable within a piece X_i , condition (2.1) imposes constraints on the choice of subgradient $f'(\bar{x}) \in \partial f(\bar{x})$ at boundary points \bar{x} . Specifically, the chosen subgradient must satisfy the inequality for all x in a given piece X_i containing \bar{x} . For instance, consider $f(x) = |x|$ on $X = \mathbb{R}$ with the partition $X_1 = (-\infty, 0]$ and $X_2 = (0, \infty)$. Condition (2.1), when applied with $\bar{x} = 0$ and considered relative to the piece X_1 (which contains the non-differentiable point 0), requires choosing the specific subgradient $f'(0) = -1$ from $\partial f(0) = [-1, 1]$. However, access to such specific boundary subgradients is not critical for the practical algorithm design presented later, as means exist to relax this requirement (e.g., by ensuring iterates land on differentiable points almost surely). For analytical simplicity throughout this section, we nevertheless assume access to an oracle $f'(x)$ satisfying (2.1) for some valid covering $\{X_i\}$.

The next definition, for quadratic growth, is standard in the optimization literature. For a convex function f , this property is slightly weaker than the strong convexity condition (see [16]).

Definition 2 (Quadratic Growth) Consider a function $f : X \rightarrow \mathbb{R}$ with a non-empty minimizer set $X^* \subseteq X$. We say f satisfies the quadratic growth condition with modulus $\mu > 0$ if the following inequality holds for any optimal solution $x^* \in X^*$:

$$f(x) - f(x^*) \geq \frac{\mu}{2} \text{dist}^2(x, X^*), \quad \forall x \in X.$$

(Here, $\text{dist}(x, X^*) := \inf_{y \in X^*} \|x - y\|$ denotes the distance from point x to the set X^* .)

2.2 The Algorithm and the Key Idea

Algorithm 2.1 The Bundle Level Method with Known f^* , $\text{BL}(m, f^*, x^0)$

Input: Optimal value f^* ; initial point $x^0 \in X$; number of cuts parameter $m \geq 1$.

1: **for** $t = 0, 1, 2, \dots$ **do**

2: Define the level set using the m most recent cuts:

$$X(t) := \{x \in X \mid \langle f'(x^{t-i}), x - x^{t-i} \rangle + f(x^{t-i}) \leq f^*, \quad \forall i \in \{0, 1, \dots, m-1\}\}.$$

3: Compute the next iterate by projection: $x^{t+1} \leftarrow \arg \min_{x \in X(t)} \frac{1}{2} \|x - x^t\|^2$.

4: **end for**

The bundle level method for optimizing the convex piecewise smooth objective function f , assuming f^* is known, is provided in Algorithm 2.1. It is closely related to the method proposed in [25]. In each iteration t , the next iterate x^{t+1} is generated by projecting the current point x^t onto the level set $X(t)$. This set $X(t)$ is constructed from m cuts generated using first-order information $f'(x^{t-i})$ from the current and previous $m-1$ iterates (specifically, for $i \in \{0, 1, \dots, m-1\}$).

Towards establishing global linear convergence, we show that the iterates make significant progress towards the optimal set X^* every time a matching pair is encountered, defined as follows:

Definition 3 (l -Matching Pair) For a trajectory $\{x^t\}_{t \geq 0}$ and an integer $l \geq 1$, we call the pair (x^t, x^{t+j}) an l -matching pair with respect to the pieces $\{X_i\}_{i=1}^k$ if $j \in \{1, \dots, l\}$ and there exists some piece index $\bar{i} \in \{1, \dots, k\}$ such that both x^t and x^{t+j} belong to the same piece $X_{\bar{i}}$.

Since there are only k pieces, the pigeonhole principle guarantees that a k -matching pair occurs within any $k+1$ consecutive iterations (i.e., for any t , there exists $j \in \{1, \dots, k\}$ such that (x^t, x^{t+j}) is a k -matching pair). Let us focus on such a k -matching pair (x^t, x^{t+j}) with both points belonging to the same piece $X_{\bar{i}}$. Intuitively, the iterates x^t, \dots, x^{t+j-1} explore the function landscape, potentially landing on different pieces X_i . However, because x^{t+j} and the earlier iterate x^t lie in the same piece $X_{\bar{i}}$ (over which f behaves smoothly according to Definition 1), the older cut generated at x^t remains highly relevant when computing x^{t+j} . This cut is included in the bundle defining the level set $X(t+j-1)$ (since $j \leq k$ via pigeonhole and the convergence analysis requires $m > k$), providing sufficient information about f 's behavior on $X_{\bar{i}}$ to ensure the iterate x^{t+j} makes significant progress towards X^* .

To be more precise, let us consider the exploration iterates between the matching pair. From the convexity of f and the definition of $X(t)$, we have¹ that the optimal set X^* is contained within the level set $X(t)$ for all t . Therefore, the projection operation defining x^{t+1} in Line 3 of Algorithm 2.1 implies the standard separating hyperplane inequality associated with projections onto convex sets:

$$\langle x^{t+1} - x^t, x^{t+1} - x^* \rangle \leq 0, \quad \forall x^* \in X^*. \quad (2.2)$$

Rearranging this (or using the equivalent form $\langle x^t - x^{t+1}, x^* - x^{t+1} \rangle \leq 0$) yields the crucial three-point inequality [21]:

$$\|x^{t+1} - x^*\|^2 + \|x^t - x^{t+1}\|^2 \leq \|x^t - x^*\|^2, \quad \forall x^* \in X^*. \quad (2.3)$$

¹ Since $x^* \in X^*$, convexity implies $f(x^{t-i}) + \langle f'(x^{t-i}), x^* - x^{t-i} \rangle \leq f(x^*) \leq f^*$ for any $x^* \in X^*$ and any i . Thus, any $x^* \in X^*$ satisfies all conditions defining $X(t)$, implying $X^* \subseteq X(t)$.

This inequality (2.3) is key, implying that the distance from the iterates to any optimal point $x^* \in X^*$ is monotonically non-increasing, i.e., $\|x^{t+1} - x^*\| \leq \|x^t - x^*\|$, ensuring the exploration iterates remain bounded relative to the optimal set X^* . Importantly, these inequalities can be combined over consecutive steps to obtain the following "bridged" version.

Lemma 2.1 *Assume the iterates $\{x^t\}_{t \geq 0}$ and a point $x^* \in X^*$ satisfy $\langle x^{t+1} - x^t, x^{t+1} - x^* \rangle \leq 0$ for all $t \geq 0$. Then the following **bridged three-point inequality** holds:*

$$\|x^{t+j} - x^*\|^2 + \frac{1}{j} \|x^{t+j} - x^t\|^2 \leq \|x^t - x^*\|^2, \quad \forall t \geq 0, j \geq 1.$$

Proof Adding up $\|x^i - x^{i+1}\|^2 + \|x^{i+1} - x^*\|^2 \leq \|x^i - x^*\|^2$ from $i = t$ to $t + j - 1$, we get

$$\|x^{t+j} - x^*\|^2 + \sum_{i=1}^j \|x^{t+i} - x^{t+i-1}\|^2 \leq \|x^t - x^*\|^2.$$

The result then follows from the algebraic identity

$$(j) \sum_{i=1}^j \|x^{t+i} - x^{t+i-1}\|^2 \geq \|(x^{t+j} - x^{t+j-1}) + (x^{t+j-1} - x^{t+j-2}) + \dots (x^{t+1} - x^t)\|^2.$$

■

Now, to illustrate the significant improvement derived from the iterate x^{t+j} within a k -matching pair (x^t, x^{t+j}) , let us assume for simplicity that the optimal set X^* is a singleton, $X^* = \{x^*\}$. As discussed, (x^t, x^{t+j}) belong to the same piece $X_{\bar{i}}$, and the cut generated at x^t is included in the level set $X(t+j-1)$ used to compute x^{t+j} via projection (Line 3 of Algorithm 2.1). The feasibility of x^{t+j} with respect to this specific cut implies:

$$\begin{aligned} \langle f'(x^t), x^{t+j} - x^t \rangle + f(x^t) &\leq f^* \\ \stackrel{(a)}{\implies} f(x^{t+j}) - f^* &\leq \frac{L}{2} \|x^{t+j} - x^t\|^2 \\ \stackrel{(b)}{\implies} \frac{\mu}{2} \|x^{t+j} - x^*\|^2 &\leq \frac{L}{2} \|x^{t+j} - x^t\|^2, \end{aligned}$$

where implication (a) follows from the L -smoothness of f on the piece $X_{\bar{i}}$ (Eq. (2.1)) and (b) follows from the quadratic growth condition. This yields the inequality $\mu \|x^{t+j} - x^*\|^2 \leq L \|x^{t+j} - x^t\|^2$.

Combining this inequality with the bridged three-point inequality $kL \|x^{t+j} - x^*\|^2 + \frac{k}{j} \frac{L}{L} \|x^{t+j} - x^*\|^2 \leq kL \|x^t - x^*\|^2$ from Lemma 2.1, and using the fact that $j \leq k$ for the k -matching pair, we directly obtain the progress guarantee:

$$\|x^{t+j} - x^*\|^2 \leq \frac{kL}{kL + \mu} \|x^t - x^*\|^2. \quad (2.4)$$

Since the distance to the optimal solution x^* is monotonically non-increasing (from (2.3)), and the squared distance contracts by at least the factor $\frac{kL}{kL + \mu} < 1$ every time a k -matching pair occurs (which happens within every $k + 1$ iterations), this establishes the desired global linear convergence guarantee if the number of cuts m is bigger than k .

2.3 The Convergence Guarantee

To quantify the contraction resulting from multiple matching pairs in a more precise fashion, the following definitions associated with a sequence of matching pairs are useful.

Definition 4 (*l*-Matching Pair Sequence and Statistics) We call a sequence of index pairs $\{(l_i, r_i)\}_{i \geq 1}$ an *l-matching pair sequence* with respect to the pieces $\{X_i\}_{i=1}^k$ if every pair (x^{l_i}, x^{r_i}) in the corresponding trajectory $\{x^t\}$ is an *l*-matching pair (see Definition 3) and the pairs correspond to non-overlapping index intervals, i.e., $r_i \leq l_{i+1} < r_{i+1}$ for all $i \geq 1$. Given such a sequence and an iteration limit $N \geq l$, let $P(N) := \max\{i \in N_+ \mid r_i \leq N\}$ denote the number of matching pairs completed by iteration N . We define the *average inter-arrival time* $\bar{\kappa}(N)$ and the *average length* $\bar{\sigma}(N)$ associated with the sequence up to iteration N as:

$$\bar{\kappa}(N) := \frac{N}{p(N)}, \quad \bar{\sigma}(N) := p(N) \left/ \sum_{i=1}^{p(N)} \frac{1}{r_i - l_i} \right. . \quad (2.5)$$

For any (k, L) -piecewise smooth function, the pigeonhole principle implies that there always exists some k -matching pair sequence with $\bar{\kappa}(N) \leq 2k$ and $\bar{\sigma}(N) \leq k + 1$.

Now we are ready to state the formal convergence guarantee.

Theorem 2.1 *For problem (1.1), assume f is a convex and (k, L) -piecewise smooth function (Definition 1). Consider the iterates $\{x^t\}_{t \geq 0}$ generated by Algorithm 2.1 with inputs including the number of cuts m satisfying $m \geq k$, the optimal value f^* , and an initial point $x^0 \in X$. Let $\{(x^{l_i}, x^{r_i})\}$ be any k -matching pair sequence (see Definition 3) associated with this trajectory, and let $\bar{\kappa}(N)$ and $\bar{\sigma}(N)$ be the corresponding statistics defined in (2.5) for a given iteration count N . The following convergence guarantees hold:*

a) *If f is convex, then for $N \geq 1$:*

$$\min_{t \in \{1, \dots, N\}} f(x^t) - f^* \leq \frac{L \bar{\sigma}(N) \bar{\kappa}(N)}{2N} \text{dist}^2(x^0, X^*).$$

b) *Moreover, if f satisfies the quadratic growth condition (Definition 2) with modulus $\mu > 0$, then for $N \geq 1$, the sequence converges linearly to X^* :*

$$\begin{aligned} \text{dist}^2(x^N, X^*) &\leq \left(1 + \frac{1}{\bar{\kappa}(N)}\right)^{-N} C \cdot \text{dist}^2(x^0, X^*), \\ \min_{t \in \{1, \dots, N\}} f(x^t) - f^* &\leq \left(1 + \frac{1}{\bar{\kappa}(N)}\right)^{-N} C \cdot [f(x^0) - f^*], \end{aligned}$$

where $C > 0$ is an absolute constant and the effective condition number $\kappa(N)$ is given by

$$\kappa(N) = O(1) \left(\frac{\bar{\sigma}(N) \bar{\kappa}(N) L}{\mu} \right).$$

Proof We defer the detailed analysis to Subsection 2.4.

We make a few remarks regarding Theorem 2.1.

First, the complexity bound effectively depends on the matching pair statistics $\bar{\sigma}(N)$ and $\bar{\kappa}(N)$, which reflect the actual iterate behavior rather than relying solely on the total number of pieces k (worst case). This offers a potentially more precise characterization of the oracle complexity. Specifically, rearranging the linear convergence result from Theorem 2.1(b) yields an iteration complexity $N(\epsilon)$ to reach ϵ -accuracy of:

$$N(\epsilon) = O \left(\frac{L \bar{\sigma}(N(\epsilon)) \bar{\kappa}(N(\epsilon))}{\mu} \log \left(\frac{1}{\epsilon} \right) \right).$$

Here, the factor L/μ corresponds to the condition number common in non-accelerated analyses for strongly convex smooth optimization. The additional cost for handling the unknown PWS structure is captured by the multiplicative factors $\bar{\sigma}(N)$ and $\bar{\kappa}(N)$ (see Definition 4). Since the pigeonhole principle guarantees the existence of a k -matching pair sequence with $\bar{\kappa}(N) \leq O(k)$ and $\bar{\sigma}(N) \leq k$, a conservative upper bound on the oracle complexity is:

$$N(\epsilon) \leq O \left(\frac{Lk^2}{\mu} \log \left(\frac{1}{\epsilon} \right) \right).$$

For comparison, consider minimizing a k -max-of-smooth objective $f(x) = \max_{i=1,\dots,k} \tilde{f}_i(x)$. If oracle access to all individual smooth components \tilde{f}_i is available, the unaccelerated prox-linear method achieves $O\left(\frac{Lk}{\mu} \log\left(\frac{1}{\epsilon}\right)\right)$ complexity [35]. Comparing these bounds suggests our algorithm incurs an additional factor related to k due to the lack of knowledge of the individual pieces. However, the dependence on the potentially smaller $\bar{\sigma}(N)$ and $\bar{\kappa}(N)$ might yield better practical performance than the $O(k^2)$ worst-case bound suggests.

Second, Algorithm 2.1 itself does not require knowledge of the problem parameters μ (quadratic growth modulus) or L (smoothness constant) to execute. This advantage of being relatively parameter-free (regarding L, μ) is common to bundle-level type algorithms [20, 25]. The crucial inputs for Algorithm 2.1 as presented are the optimal value f^* and the number of cuts m . Note that the linear convergence analysis in Theorem 2.1 requires selecting $m \geq k$. As discussed in Contribution 3, Section 4 and Section 6 will propose methods to remove the dependence on f^* . The parameter m controls the memory of the method (number of cuts stored) and represents a trade-off between the iteration complexity and the computational cost per iteration (solving the quadratic program in Line 3 of Algorithm 2.1).

Third, the choice of m affects the convergence behavior. While $m \geq k$ is assumed for the global linear rate in Theorem 2.1, choosing a smaller m (e.g., m related to the number of pieces k_{local} in a neighborhood around x^*) is computationally cheaper per iteration. Subsequent analysis in Section 3 shows that with such smaller m , the algorithm typically first achieves an $O(1/N)$ sublinear convergence rate, followed by local linear convergence once the iterates enter the relevant neighborhood. Furthermore, if the objective satisfies a sharper growth condition than quadratic growth, such as $f(x) - f^* \geq \mu \text{dist}(x, X^*)$, Algorithm 2.1 might achieve faster local convergence, potentially quadratic, improving over the linear convergence guarantee for the proximal bundle method in [10].

2.4 The Detailed Convergence Proof

The following technical lemma regarding iterate properties and function value convergence is useful for analyzing both the general convex setting and the quadratic growth setting.

Lemma 2.2 *Assume f is a convex and (k, L) -piecewise smooth function (Definition 1) for problem (1.1). Let the iterates $\{x^t\}_{t \geq 0}$ be generated by Algorithm 2.1 with inputs including $m \geq k$, the optimal value f^* , and an initial point $x^0 \in X$. The following relations hold:*

a) *The iterates are non-expansive with respect to the optimal set X^* :*

$$\|x^t - x^*\|^2 \leq \|x^{t-1} - x^*\|^2, \quad \forall x^* \in X^*, \forall t \geq 1.$$

b) *For every k -matching pair (x^l, x^r) (see Definition 3), we have:*

$$\frac{1}{(r-l)L} (f(x^r) - f^*) \leq \frac{1}{2} (\|x^l - x^*\|^2 - \|x^r - x^*\|^2), \quad \forall x^* \in X^*.$$

Proof a) As argued previously (e.g., in the discussion leading to (2.3) or the footnote on Page 7), convexity implies $X^* \subseteq X(t)$ for all t . The projection in Line 3 of Algorithm 2.1 yields the condition $\langle x^t - x^{t-1}, x^t - x^* \rangle \leq 0$ for all $x^* \in X^*$ and $t \geq 1$. This directly implies the standard three-point inequality:

$$\|x^t - x^{t-1}\|^2 + \|x^t - x^*\|^2 \leq \|x^{t-1} - x^*\|^2, \quad \forall x^* \in X^*. \quad (2.6)$$

The non-expansiveness result $\|x^t - x^*\|^2 \leq \|x^{t-1} - x^*\|^2$ follows immediately by dropping the non-negative term $\|x^t - x^{t-1}\|^2$.

b) For the k -matching pair (x^l, x^r) , we have $1 \leq r - l \leq k$. Since the algorithm runs with $m \geq k$, the cut generated at x^l is included in the level set $X(r-1)$ used to compute x^r . Feasibility of $x^r \in X(r-1)$ implies:

$$\langle f'(x^l), x^r - x^l \rangle + f(x^l) \leq f^*. \quad (2.7)$$

Since x^l, x^r belong to the same piece $X_{\bar{i}}$, combining (2.7) with the L -smoothness property on $X_{\bar{i}}$ (Definition 1) yields:

$$f(x^r) - f^* \leq \frac{L}{2} \|x^r - x^l\|^2. \quad (2.8)$$

Applying Lemma 2.1 with $t = l$ and $j = r - l$ gives:

$$\|x^r - x^*\|^2 + \frac{1}{r-l} \|x^r - x^l\|^2 \leq \|x^l - x^*\|^2. \quad (2.9)$$

Combining the consequence of (2.8) (i.e., $\frac{1}{r-l} \|x^r - x^l\|^2 \geq \frac{2}{L(l-r)}(f(x^r) - f^*)$) with (2.9) leads directly to the result stated in part (b). ■

Next, we show part a) of Theorem 2.1.

Proof of Theorem 2.1(a) Let $x^* \in X^*$ be an arbitrary optimal solution. For every k -matching pair (x^{l_i}, x^{r_i}) in the given sequence, combining Lemma 2.2(b) (multiplied by L) and Lemma 2.2(a) (non-expansiveness, implying $\|x^{l_i} - x^*\|^2 \leq \|x^{r_{i-1}} - x^*\|^2$ for $i \geq 2$), we have:

$$\frac{2}{r_i - l_i} (f(x^{r_i}) - f^*) + L \|x^{r_i} - x^*\|^2 \leq L \|x^{l_i} - x^*\|^2.$$

Summing this relation from $i = 1$ to $P(N)$ (where $P(N)$ is the number of pairs completed by iteration N) yields a telescoping sum on the right-hand side (after applying non-expansiveness):

$$\sum_{i=1}^{P(N)} \frac{2}{r_i - l_i} (f(x^{r_i}) - f^*) \leq L \|x^{l_1} - x^*\|^2 \leq L \|x^0 - x^*\|^2.$$

Let $f_{\min, N} := \min_{t \in \{1, \dots, N\}} f(x^t)$. Lower bounding $f(x^{r_i})$ by $f_{\min, N}$:

$$\left(\sum_{i=1}^{P(N)} \frac{2}{r_i - l_i} \right) (f_{\min, N} - f^*) \leq \sum_{i=1}^{P(N)} \frac{2}{r_i - l_i} (f(x^{r_i}) - f^*) \leq L \|x^0 - x^*\|^2.$$

By definition (2.5) (see also Definition 4), we have $\sum_{i=1}^{P(N)} \frac{1}{r_i - l_i} = P(N)/\bar{\sigma}(N)$ and $P(N) = N/\bar{\kappa}(N)$. Therefore, the sum is $\frac{N}{\bar{\kappa}(N)\bar{\sigma}(N)}$. Substituting this gives:

$$\frac{2N}{\bar{\kappa}(N)\bar{\sigma}(N)} (f_{\min, N} - f^*) \leq L \|x^0 - x^*\|^2.$$

Rearranging and minimizing the right-hand side over $x^* \in X^*$ (yielding $\text{dist}^2(x^0, X^*)$) gives:

$$\min_{t \in \{1, \dots, N\}} [f(x^t) - f^*] \leq \frac{L\bar{\kappa}(N)\bar{\sigma}(N)}{2N} \text{dist}^2(x^0, X^*). \quad (2.10)$$
■

Now we bootstrap the result from part (a) to prove the linear convergence in part (b) of Theorem 2.1. The analysis is slightly more involved than the simple argument in (2.4) because the optimal solution set X^* is not necessarily a singleton.

Proof (Proof of Theorem 2.1(b)) We partition the matching pair sequence $\{(l_i, r_i)\}$ into phases. Define the index of the last pair in phase $j \geq 1$ as:

$$R(j) := \min \left\{ s \in \{1, 2, \dots\} \mid \sum_{i=1}^s \frac{1}{r_i - l_i} \geq \frac{eL}{\mu} j + j - 1 \right\}. \quad (2.11)$$

Let $r_{R(0)} := 0$, so $x^{r_{R(0)}} = x^0$. We will show by induction on $j \geq 0$ that:

$$\text{dist}^2(x^{r_{R(j)}}, X^*) \leq e^{-j} \text{dist}^2(x^0, X^*). \quad (2.12)$$

The base case $j = 0$ holds since $e^0 = 1$. Assume (2.12) holds for some $j - 1 \geq 0$. Consider phase j , containing pairs indexed from $i = R(j - 1) + 1$ to $R(j)$. By the definition (2.11) of $R(j)$ and $R(j - 1)$, the sum of inverse lengths within this phase satisfies:

$$\begin{aligned} \sum_{i=R(j-1)+1}^{R(j)} \frac{1}{r_i - l_i} &= \sum_{i=1}^{R(j)} \frac{1}{r_i - l_i} - \sum_{i=1}^{R(j-1)} \frac{1}{r_i - l_i} \\ &\geq \frac{e(j)L}{\mu} + j - \left[\frac{e(j-1)L}{\mu} + s \right] \geq \frac{eL}{\mu}. \end{aligned} \quad (2.13)$$

Now, let x^{m_j} be an iterate within the range $t \in \{r_{R(j-1)} + 1, \dots, r_{R(j)}\}$ such that $f(x^{m_j}) = \min\{f(x^t) \mid r_{R(j-1)} < t \leq r_{R(j)}\}$. Applying the result derived in the proof of Theorem 2.1(a) (specifically, the inequality relating the sum to the distance squared, adapted to start from $x^{r_{R(j-1)}}$ and summing over phase j), we have:

$$\left(\sum_{i=R(j-1)+1}^{R(j)} \frac{2}{r_i - l_i} \right) (f(x^{m_j}) - f^*) \leq L \text{dist}^2(x^{r_{R(j-1)}}, X^*).$$

Using the quadratic growth condition $f(x^{m_j}) - f^* \geq (\mu/2) \text{dist}^2(x^{m_j}, X^*)$ and the sum lower bound (2.13), this implies:

$$\left(\frac{2eL}{\mu} \right) \left(\frac{\mu}{2} \text{dist}^2(x^{m_j}, X^*) \right) \leq L \text{dist}^2(x^{r_{R(j-1)}}, X^*).$$

Simplifying gives $\text{dist}^2(x^{m_j}, X^*) \leq e^{-1} \text{dist}^2(x^{r_{R(j-1)}}, X^*)$. Finally, by Lemma 2.2(a), the distance $\text{dist}(x^t, X^*)$ is non-increasing. Since $r_{R(j)} \geq m_j$, we have $\text{dist}^2(x^{r_{R(j)}}, X^*) \leq \text{dist}^2(x^{m_j}, X^*)$. Combining these inequalities yields:

$$\text{dist}^2(x^{r_{R(j)}}, X^*) \leq \text{dist}^2(x^{m_j}, X^*) \leq e^{-1} \text{dist}^2(x^{r_{R(j-1)}}, X^*). \quad (2.14)$$

Applying the induction hypothesis for $j - 1$, $\text{dist}^2(x^{r_{R(j-1)}}, X^*) \leq e^{-(j-1)} \text{dist}^2(x^0, X^*)$, completes the inductive step:

$$\text{dist}^2(x^{r_{R(j)}}, X^*) \leq e^{-1} \left(e^{-(j-1)} \text{dist}^2(x^0, X^*) \right) = e^{-j} \text{dist}^2(x^0, X^*).$$

By the principle of mathematical induction, (2.12) holds for all $j \geq 0$.

Now we derive the convergence bound for the fixed iteration count $N \geq 1$ from the per-phase contraction (2.12). Let $\bar{j} := \max\{j \in \{1, 2, \dots\} \mid r_{R(j)} \leq N\}$ be the index of the last phase fully completed by iteration N . The definition of $R(j)$ in (2.11) along with the definitions of $P(N)$, $\bar{\sigma}(N)$, $\bar{\kappa}(N)$ in (2.5) establish a relationship between \bar{j} and N . Specifically, since $R(\bar{j} + 1) \geq P(N)$, the following inequality can be shown to hold:

$$(\bar{j} + 1) \left(\frac{eL}{\mu} + 1 \right) \geq \sum_{i=1}^{P(N)} \frac{1}{r_i - l_i} = \frac{N}{\bar{\sigma}(N)\bar{\kappa}(N)}. \quad (2.15)$$

Let $\kappa(N) := \bar{\sigma}(N)\bar{\kappa}(N)(\frac{eL}{\mu} + 1)$. Then (2.15) implies:

$$\bar{j} + 1 \geq \frac{N}{\kappa(N)}.$$

Therefore, $\bar{j} \geq N/\kappa(N) - 1$.

Now, using monotonicity (Lemma 2.2(a)) and the per-phase decay (2.12):

$$\begin{aligned} \text{dist}^2(x^N, X^*) &\leq \text{dist}^2(x^{r_{R(\bar{j})}}, X^*) \quad (\text{by non-expansiveness}) \\ &\leq e^{-\bar{j}} \text{dist}^2(x^0, X^*) \quad (\text{using (2.12)}) \\ &\leq e^{-(N/\kappa(N)-1)} \text{dist}^2(x^0, X^*) \quad (\text{using the lower bound on } \bar{j}) \\ &= e \cdot \exp\left(-\frac{N}{\kappa(N)}\right) \text{dist}^2(x^0, X^*). \end{aligned}$$

Using the algebraic inequality $e^{-\frac{1}{A}} \leq (1 + \frac{1}{A})^{-1}$ for $A > 0$, we get

$$\text{dist}^2(x^N, X^*) \leq e \cdot \left(1 + \frac{1}{\kappa(N)}\right)^{-N} \text{dist}^2(x^0, X^*).$$

This precisely matches the first result stated in part (b) with the constant $C = e$.

The function value convergence result in Theorem 2.1(b) follows via analogous arguments, yielding the same convergence factor $e \cdot (1 + 1/\kappa(N))^{-N}$. We list the central recursive relation derived during the induction:

$$\begin{aligned} \min_{t \in [r_{R(j)}]} f(x^t) - f^* &\leq \frac{L}{2} \frac{1}{(\sum_{i=R(j-1)+1}^{R(j)} \frac{1}{r_i - l_i})} \text{dist}^2(x^{r_{R(j-1)}}, X^*) \\ &\leq \frac{\mu}{2e} \text{dist}^2(x^{r_{R(j-1)}}, X^*) \leq \frac{1}{e} \min_{t \in [r_{R(j-1)}]} f(x^t) - f^*. \end{aligned}$$

■

3 Approximately Piecewise Smooth Function

In this section, we introduce a more general class of approximately piecewise smooth functions (see Definition 5). This framework allows us to relax the strict requirement from Definition 1 to select a specific ("right") subgradient at non-differentiable points and potentially broadens the applicability to more general non-smooth functions. We will show that the proposed bundle-level method, with minor modifications, can still achieve linear convergence for this function class.

3.1 The Problem Set-up

Definition 5 (Approximate Piecewise Smoothness) We call a convex function $f : X \rightarrow \mathbb{R}$ (k, L, δ) -**approximately piecewise smooth** (apx-PWS) if there exists some covering $\{X_i\}_{i=1}^k$ of X (i.e., $X \subseteq \cup_{i=1}^k X_i$) and an oracle which, when queried at a point $\bar{x} \in X$, returns a linear support function $\tilde{l}_f(\cdot; \bar{x})$ satisfying the following requirements for some $L \geq 0, \delta \geq 0$:

- a) Support function property: $f(y) \geq \tilde{l}_f(y; \bar{x})$ for all $y \in X$.
- b) Approximate smoothness on pieces: if $\bar{x} \in X_i$ for some $i \in \{1, \dots, k\}$, then

$$f(y) - \tilde{l}_f(y; \bar{x}) \leq \frac{L}{2} \|y - \bar{x}\|^2 + \delta, \quad \forall y \in X_i.$$

Clearly, a (k, L) -PWS function (see Definition 1) is also a $(k, L, 0)$ -apx-PWS function if the oracle $f'(x)$ is chosen appropriately. Let us motivate this definition further by discussing an important application.

- *Relaxing the requirement for specific subgradients:* Recall that Definition 1 requires the first-order oracle $f'(x)$ to return a potentially specific subgradient at non-differentiable points to satisfy the smoothness condition (2.1) within each piece X_i . This might not be feasible with only black-box subgradient access.

To address this, consider a convex (k, L) -PWS function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that is also M -Lipschitz continuous. We construct an approximate linear functional $\tilde{l}_f(\cdot; \bar{x})$ using gradient information at a perturbed point \tilde{x} . When the oracle is queried at $\bar{x} \in X$, we sample \tilde{x} uniformly from the Euclidean ball $B(\bar{x}; \bar{\delta})$ and define:

$$\tilde{l}_f(x; \bar{x}) := f(\tilde{x}) + \langle \nabla f(\tilde{x}), x - \tilde{x} \rangle. \quad (3.1)$$

(Here, \tilde{x} is differentiable w.p.1 by Rademacher's theorem). This \tilde{l}_f satisfies the support function property (a) by convexity. Choosing the radius $\bar{\delta} \leq \min\{\sqrt{\delta/(8L)}, \delta/(4L)\}$ for a target $\delta > 0$, we can verify the approximate smoothness property (b). Specifically, based on the original

covering $\{X_i\}_{i=1}^k$, for any finite set of evaluation points $\{x^t\}_{t=1}^N$ ², we assign x^t to piece \tilde{X}_i if its corresponding perturbed point $\tilde{x}^t \in X_i$. This collection $\{\tilde{X}_i\}_{i=1}^k$ provides a covering for the evaluated points $\{x^t\}_{t=1}^N$. For any two points x^i, x^j belonging to the same derived piece \tilde{X}_i , Lemma 7.1 implies that:

$$f(x^i) - \tilde{l}_f(x^i; x^j) = f(x^i) - l_f(x^i; \tilde{x}^j) \leq L\|x^i - x^j\|^2 + \delta, \quad \forall x^i, x^j \in \tilde{X}_i.$$

Thus, the proposed perturbation-based support function oracle in (3.1) allows us to formulate any convex, M -Lipschitz, (k, L) -PWS function as a $(k, 2L, \delta)$ -apx-PWS function satisfying Definition 5, without the subgradient selection requirement.

- *General Lipschitz continuous (non-smooth) convex functions:* Any M -Lipschitz continuous convex function f can be viewed as approximately piecewise smooth with one piece ($k = 1$). Specifically, f is $(1, M^2/\delta, \delta)$ -apx-PWS for any $\delta > 0$. This follows from subgradient properties and Young’s inequality:

$$f(x) - f(\bar{x}) - \langle f'(\bar{x}), x - \bar{x} \rangle \leq M\|x - \bar{x}\| \leq \frac{M^2}{2\delta}\|x - \bar{x}\|^2 + \frac{\delta}{2}, \quad \forall f'(\bar{x}) \in \partial f(\bar{x}), \forall x, \bar{x}.$$

From this perspective, the iterates generated by Algorithm 3.1 converge to the optimal solution for any choice of the number of cuts $m \geq 1$. Indeed, since Algorithm 3.1 does not require L nor δ as input parameters (cf. Definition 5), we can make appropriate choices in the analysis for Theorem 3.1 to show that the optimal oracle complexity for optimizing general M -Lipschitz non-smooth convex functions [35] can be achieved by setting $m = 1$.

3.2 The Algorithm and the Convergence Guarantee

Algorithm 3.1 The Approximate Bundle Level Method with Known f^* , apx-BL(m, f^*, x^0)

Input: Optimal value f^* ; initial point $x^0 \in X$; number of cuts parameter $m \geq 1$.

1: **for** $t = 0, 1, 2, \dots$ **do**

2: Define the level set using the m most recent approximate linear models:

$$X(t) := \{x \in X \mid \tilde{l}_f(x; x^{t-i}) \leq f^*, \quad \forall i \in \{0, 1, \dots, m-1\}\}.$$

3: Compute the next iterate by projection: $x^{t+1} \leftarrow \arg \min_{x \in X(t)} \frac{1}{2}\|x - x^t\|^2$.

4: **end for**

As shown in Algorithm 3.1, the proposed approximate bundle level method (apx-BL) for handling (k, L, δ) -apx-PWS functions (see Definition 5) is a slight generalization of Algorithm 2.1 from the previous section. In each iteration, the new iterate x^{t+1} is generated by projecting the current iterate x^t onto the bundle level set $X(t)$. This set $X(t)$ is now constructed using the approximate linear functionals $\tilde{l}_f(\cdot; x^{t-i})$ provided by the apx-PWS oracle instead of the exact linearizations based on $f'(x^{t-i})$. If the approximate linear functional $\tilde{l}_f(\cdot; x^i)$ happens to be the standard linearization $l_f(x; x^i) = f(x^i) + \langle f'(x^i), x - x^i \rangle$ (corresponding to the case where $\delta = 0$ and the oracle f' satisfies Definition 1), then Algorithm 3.1 is exactly the same as Algorithm 2.1. Due to this similarity, the concepts introduced in the last section, such as matching pairs, can be readily adapted, and the convergence results are analogous. Specifically, the formal convergence guarantee for apx-BL is presented in the next theorem.

Theorem 3.1 *For problem (1.1), assume f is a convex and (k, L, δ) -apx-PWS (Definition 5) with associated pieces $\{X_i\}_{i=1}^k$. Consider the iterates $\{x^t\}_{t \geq 0}$ generated by Algorithm 3.1 with inputs including the number of cuts m satisfying $m \geq k$, the optimal value f^* , and an initial point $x^0 \in X$.*

² Considering only a finite set of points is formally weaker than the requirement in Definition 5; however, it is sufficient for algorithm analysis since only the finite sequence of iterates generated is relevant. Another subtlety is that the piece membership \tilde{X}_i associated with an iterate x^t might change if x^t is queried multiple times (due to resampling \tilde{x}^t). However, within a single bundle-level step, all cuts use distinct evaluation centers. Furthermore, as our algorithm handles unknown coverings, the analysis is robust to such potentially changing set assignments.

Let $\{(x^{l_i}, x^{r_i})\}$ be any k -matching pair sequence (Definition 3) with respect to $\{X_i\}$, and let $\bar{\kappa}(N)$ and $\bar{\sigma}(N)$ be the corresponding statistics from (2.5). The following convergence guarantees hold:

a) If f is convex, then for $N \geq 1$:

$$\min_{t \in \{1, \dots, N\}} f(x^t) - f^* \leq \frac{L\bar{\sigma}(N)\bar{\kappa}(N)}{2N} \text{dist}^2(x^0, X^*) + \delta.$$

b) Moreover, if f satisfies the quadratic growth condition (Definition 2) with modulus $\mu > 0$, then for $N \geq 1$:

$$\begin{aligned} \text{dist}^2(x^N, X^*) &\leq \left(1 + \frac{1}{\kappa(N)}\right)^{-N} C \cdot \text{dist}^2(x^0, X^*) + \frac{4\delta}{\mu}, \\ \min_{t \in \{1, \dots, N\}} f(x^t) - f^* &\leq \left(1 + \frac{1}{\kappa(N)}\right)^{-N} C \cdot [f(x^0) - f^*] + 2e\delta, \end{aligned}$$

where $C > 0$ is an absolute constant and the effective condition number $\kappa(N)$ is given by

$$\kappa(N) = O(1) \left(\frac{\bar{\sigma}(N)\bar{\kappa}(N)L}{\mu} \right).$$

Proof We defer the detailed analysis to Subsection 3.3.

A couple of remarks are in order regarding Theorem 3.1.

First, the complexity bound effectively depends on the matching pair statistics $\bar{\sigma}(N)$ and $\bar{\kappa}(N)$, reflecting the generated iterates' properties rather than just the global worst-case piece count k . Specifically, to reach an ϵ -accuracy where the approximation error δ is small (e.g., $\epsilon \geq 4e\delta$), the iteration complexity $N(\epsilon)$ is determined by:

$$N(\epsilon) = O \left(\kappa(N(\epsilon)) \log \left(\frac{f(x^0) - f^*}{\epsilon} \right) \right) = O \left(\frac{L\bar{\sigma}\bar{\kappa}}{\mu} \log \left(\frac{1}{\epsilon} \right) \right).$$

Since the statistics $\bar{\sigma}$ and $\bar{\kappa}$ ($\bar{\sigma}(N(\epsilon))$ and $\bar{\kappa}(N(\epsilon))$ from Definition 4) effectively average behavior over the entire trajectory, their values depend mostly on what happens in a neighborhood of the optimal set X^* , especially for large N . Thus, the oracle complexity essentially depends mostly on the number of smooth pieces encountered near X^* rather than over the entire domain.

Second, Algorithm 3.1 requires only m and f^* as inputs, not the problem parameters L, μ , or δ . This relative parameter-independence (regarding L, μ, δ) is common to bundle methods [20, 25]. Third, the framework is robust to misspecification of the number of cuts parameter m . Since any M -Lipschitz convex function can be regarded as $(1, M^2/\epsilon, \epsilon)$ -apx-PWS, Algorithm 3.1 can be applied with $m = 1$ and is guaranteed to converge (albeit potentially sublinearly). Furthermore, using a more refined analysis involving an adaptive choice for the inexactness parameter δ (making it vanish relative to progress, $\delta^t = o(1) \min_{j \in [t]} f(x^t) - f^*$), the method can achieve the optimal $O(M^2/(\mu\epsilon))$ oracle complexity for general non-smooth strongly convex optimization [35], matching the theoretical lower bound.

3.3 The Detailed Convergence Analysis

Due to the similarity between Algorithm 3.1 (apx-BL) and Algorithm 2.1 (BL), the convergence proofs presented in Section 2 for the exact setting can be adapted here with minor modifications to take into account the approximation error δ . To avoid repetition, we highlight the key changes required for the analysis of Algorithm 3.1. In particular, the next lemma serves as the counterpart to Lemma 2.2 in the apx-PWS setting.

Lemma 3.1 *Assume f is a convex and (k, L, δ) -approximately piecewise smooth function (Definition 5) for problem (1.1). Let the iterates $\{x^t\}_{t \geq 0}$ be generated by Algorithm 3.1 with inputs including the number of cuts m satisfying $m \geq k$, the optimal value f^* , and an initial point $x^0 \in X$. The following relations hold:*

a) The iterates are non-expansive with respect to the optimal set X^* :

$$\|x^t - x^*\|^2 \leq \|x^{t-1} - x^*\|^2, \quad \forall x^* \in X^*, \forall t \geq 1.$$

b) For every k -matching pair (x^l, x^r) (see Definition 3), we have:

$$\frac{1}{(r-l)L} (f(x^r) - f^* - \delta) \leq \frac{1}{2} (\|x^l - x^*\|^2 - \|x^r - x^*\|^2), \quad \forall x^* \in X^*.$$

Proof Let $x^* \in X^*$ be given. Since the approximate linear support function $\tilde{l}_f(\cdot; x^i)$ provides a lower approximation to f (see part (a) of Definition 5), we always have $\tilde{l}_f(x^*; x^i) \leq f(x^*) = f^*$ for any query point x^i . Thus, x^* belongs to the level set $X(t)$ defined in Algorithm 3.1, and the non-expansiveness relation stated in part a) follows immediately.

Part b) follows from Definition 5 b). For the k -matching pair (x^l, x^r) in the same piece $X_{\bar{t}}$, applying the definition gives:

$$f(x^r) - \tilde{l}_f(x^r; x^l) - \delta \leq \frac{L}{2} \|x^r - x^l\|^2.$$

The counterpart to (2.8) then follows from the feasibility of x^r with respect to the cut from x^l (i.e., $\tilde{l}_f(x^r; x^l) \leq f^*$, as $x^r \in X(r-1)$ and $r-l \leq k \leq m$):

$$\begin{aligned} \tilde{l}_f(x^r; x^l) &\leq f^*, \\ \Rightarrow \tilde{l}_f(x^r; x^l) + \delta + \frac{L}{2} \|x^r - x^l\|^2 - f^* - \delta &\leq \frac{L}{2} \|x^r - x^l\|^2, \\ \Rightarrow \frac{2}{r-l} [f(x^r) - f^* - \delta] &\leq \frac{L}{r-l} \|x^r - x^l\|^2. \end{aligned}$$

■

Now we highlight the necessary modifications to the proofs from Section 2 to obtain Theorem 3.1.

Proof to Theorem 3.1. Part a) follows directly from Lemma 3.1 b). The argument is identical to that for Theorem 2.1 a), replacing the term $f(x^{r_i}) - f^*$ within the summation with its counterpart $[f(x^{r_i}) - f^* - \delta]$, which leads directly to the additional $+\delta$ term in the final bound.

For part b), we focus on the case where $\text{dist}^2(x^N, X^*) \geq \frac{4\delta}{\mu}$, as the desired bound holds trivially otherwise due to the additive error term. The analysis closely follows the proof of Theorem 2.1b). We define the phases using a slightly modified threshold in the definition of $R(j)$ (cf. (2.11)):

$$R(j) := \min \left\{ s \in \{1, 2, \dots\} \mid \sum_{i=1}^s \frac{1}{r_i - l_i} \geq \frac{2eL}{\mu} j + j - 1 \right\}.$$

Let m_j be an index in the range $\{r_{R(j-1)} + 1, \dots, r_{R(j)}\}$ such that $f(x^{m_j}) = \min\{f(x^t) \mid r_{R(j-1)} < t \leq r_{R(j)}\}$. With this definition, the counterpart to the recursive inequality (2.14) used in the previous proof becomes:

$$\frac{\mu}{4} \text{dist}^2(x^{m_j}, X^*) \leq \frac{\mu}{2} \text{dist}^2(x^{m_j}, X^*) - \delta \leq f(x^{m_j}) - f^* - \delta \leq \frac{\mu}{4e} \text{dist}^2(x^{r_{R(j-1)}}, X^*). \quad (3.2)$$

Here, the first inequality follows from the assumption $\text{dist}^2(x^{m_j}, X^*) \geq \text{dist}^2(x^N, X^*) \geq \frac{4\delta}{\mu}$ (noting distance is non-increasing), the second inequality is the quadratic growth condition (2), and the third follows from adapting the derivation in the proof of Theorem 2.1(a) using the modified phase definition which yields $\sum_{i=R(j-1)+1}^{R(j)} (r_i - l_i)^{-1} \geq 2eL/\mu$.

The convergence guarantee follows by applying the same induction argument as before, using (3.2) to establish the per-phase decay, and translating back to iteration N :

$$\begin{aligned} \text{dist}^2(x^N, X^*) &\leq \max \left\{ \left(1 + \frac{1}{\kappa(N)} \right)^{-N} e \text{dist}^2(x^0, X^*), \frac{4\delta}{\mu} \right\} \\ &\quad \text{with } \kappa(N) = \bar{\sigma}(N) \bar{\kappa}(N) \left(\frac{2eL}{\mu} + 1 \right). \end{aligned}$$

For the function value gap convergence, if $f(x^t) - f^* \geq 2e\delta$ for all relevant t up to some $\tilde{N} \geq m_j$, we have the following recursion:

$$\begin{aligned} f(x^{m_j}) - f^* - \delta &\leq \frac{\mu}{4e} \text{dist}^2(x^{r_{R(j-1)}}, X^*) \leq \frac{1}{2e} \left(\min_{t \in \{1, \dots, r_{R(j-1)}\}} f(x^t) - f^* \right), \\ \implies f(x^{m_j}) - f^* &\leq \max \left\{ 2e\delta, \frac{1}{e} \left(\min_{t \in \{1, \dots, r_{R(j-1)}\}} f(x^t) - f^* \right) \right\}. \end{aligned}$$

■

4 Assuming the Knowledge of the Quadratic Growth Parameter

In this section, we relax the requirement that the optimal objective value f^* is known, assuming instead knowledge of the quadratic growth parameter μ . This assumption is more practical, as any value in $[0, \mu]$ is a valid modulus. We propose a bundle-level method that dynamically searches for upper and lower bounds on f^* while achieving the same order of oracle complexity as Algorithm 3.1. The techniques developed here for the strongly convex setting will also be instrumental for handling weakly convex problems in later sections.

4.1 The Bundle Level Method with a Known μ .

Algorithm 4.1 The Gap Reduction subroutine (\mathcal{GR})

Input: The quadratic growth parameter $\mu > 0$; initial point x^0 ; initial upper bound $\bar{f} = f(x^0)$; initial lower bound \underline{f} ; number of cuts m .

Output: Updated iterate x^+ ; updated upper bound $\bar{f}^+ = f(x^+)$; updated lower bound \underline{f}^+ .

- 1: **Initialize:** $\bar{f}^0 \leftarrow \bar{f}$, $\underline{f}^0 \leftarrow \underline{f}$, $\bar{x}^0 \leftarrow x^0$, $\Delta_0 \leftarrow \bar{f}^0 - \underline{f}^0$, $\tilde{S}_r(0) \leftarrow 0$, $\tilde{S}_l(0) \leftarrow 0$.
- 2: **for** $t = 0, 1, 2, \dots$ **do**
- 3: Define level set: $f^{t+1} \leftarrow \frac{2}{3}\underline{f}^t + \frac{1}{3}\bar{f}^t$.
- 4: Compute $x^{t+1} \leftarrow \arg \min_{x \in X(t+1)} \|x - x^t\|^2$ where the level is given by

$$X(t+1) := \{x \in \mathcal{X} \mid \tilde{l}_f(x; x^{t-j}) \leq f^{t+1} \text{ for } j \in \{0, \dots, m-1\}\}$$

- 5: Update upper bound: $\bar{f}^{t+1} \leftarrow \min\{\bar{f}^t, f(x^{t+1})\}$ and $\bar{x}^{t+1} \leftarrow \arg \min_{x \in \{\bar{x}^t, x^{t+1}\}} f(x)$.
- 6: Update lower bound: $\underline{f}^{t+1} \leftarrow \max\left\{\min_{x \in \mathcal{X}} \max_{j \in \{0, \dots, m-1\}} \tilde{l}_f(x; x^{t+1-j}), \underline{f}^t\right\}$.
- 7: Update the gap: $\Delta_{t+1} \leftarrow \bar{f}^{t+1} - \underline{f}^{t+1}$.
- 8: With $\tau \leftarrow \max\{0, t+1-m\}$, update the empirical progress parameter and the smoothness constant:

$$\tilde{S}_r(t+1) \leftarrow \max_{\tau \leq q \leq t} \left\{ \tilde{S}_l(q) + \frac{1}{(t+1-q)\tilde{L}(t+1, q; \frac{\Delta_t}{6})} \right\}, \tilde{S}_l(t+1) \leftarrow \max_{\tau \leq q \leq t+1} \tilde{S}_r(q).$$

- 9: **if** $\Delta_{t+1} \leq \frac{2}{3}\Delta_0$ **then return** $(\bar{x}^{t+1}, \bar{f}^{t+1}, \underline{f}^{t+1})$.
 - 10: **else if** $\tilde{S}_r(t+1) \geq \frac{6}{\mu}$ **then**
 - 11: Set $\underline{f}^{t+1} \leftarrow \min_{i \in \{0, \dots, t+1\}} f^i$, and **return** $(\bar{x}^{t+1}, \bar{f}^{t+1}, \underline{f}^{t+1})$.
 - 12: **end if**
 - 13: **end for**
-

Algorithm 4.2 presents an iterative method to find an ϵ -optimal solution. The algorithm's outer loop narrows the gap between an upper bound \bar{f}^s and a lower bound \underline{f}^s for the optimal value f^* , terminating once the gap is certifiably less than the target accuracy ϵ :

$$f(x^S) - f^* \leq \bar{f}^S - \underline{f}^S \leq \epsilon.$$

Progress in each outer iteration is driven by the Gap Reduction (\mathcal{GR}) subroutine (Algorithm 4.1). Provided with valid initial upper and lower bounds \bar{f} and \underline{f} , this subroutine is guaranteed

Algorithm 4.2 The Bundle Level method with a Known μ (BL- μ)**Input:** Initial point x^0 , quadratic growth parameter $\mu > 0$, number of cuts m , and target accuracy $\epsilon > 0$.**Output:** An ϵ -optimal solution \hat{x} .

```

1: Initialize:  $\bar{f}^0 \leftarrow f(x^0)$  and  $\underline{f}^0 \leftarrow f(x^0) - 2\|f'(x^0)\|^2/\mu$ .
2: for  $s = 0, 1, 2, \dots$  do
3:   Set  $(x^{s+1}, \bar{f}^{s+1}, \underline{f}^{s+1}) \leftarrow \mathcal{GR}(\mu, x^s, \bar{f}^s, \underline{f}^s)$  with objective function  $f$ .
4:   if  $\bar{f}^{s+1} - \underline{f}^{s+1} \leq \epsilon$  return  $x^{s+1}$ .
5: end for

```

to shrink the current gap by at least a one-third factor, returning an improved iterate x^+ and updated bounds \bar{f}^+ and \underline{f}^+ such that $\bar{f}^+ - \underline{f}^+ \leq \frac{2}{3}(\bar{f} - \underline{f})$.

The core challenge for the \mathcal{GR} subroutine is the one-sided performance of the bundle-level method. In each iteration, a level set is constructed using a parameter f^t . If this parameter is always an overestimate of the true optimum ($f^t \geq f^*$ for all t), the iterates converge quickly. However, if f^t becomes an underestimate ($f^t < f^*$), the level set may be empty, causing the algorithm to fail. We resolve this with a decision rule to infer when f^t is an underestimate. The rule uses a contrapositive argument: assuming all $f^t \geq f^*$, we can calculate the number of iterations T required to guarantee a significant reduction in the optimality gap. Therefore, if the algorithm runs for T iterations *without* achieving this gap reduction, our assumption must be false. This implies that some $f^t < f^*$, which allows us to improve the lower bound estimate and ensure progress.

This argument for handling the level-set parameter is implemented in Algorithm 4.1. The algorithm has two termination triggers to ensure it always makes provable progress. In each iteration, it performs a bundle-level update and refines its upper and lower bounds on the optimal value in Line 3-7. It then terminates if either of two conditions is met: (1) the update successfully reduces the gap between the bounds (Line 9), or (2) a contrapositive trigger fires which allows the algorithm to update the lower bound (Line 10-11).

The effectiveness of this contrapositive trigger hinges on tracking theoretical progress (assuming all the level set parameters satisfy $f^t \geq f^*$), even though key problem parameters like the true smooth pieces $\{X_i\}$ and the smoothness constant L are unknown. We resolve this issue by introducing an empirical smoothness constant, which is calculated directly from the iterates.

Definition 6 (Empirical Smoothness Constant) For an approximately piecewise smooth (apx-PWS) objective f , the *empirical smoothness constant* between any two points $x, \bar{x} \in \mathcal{X}$ with an inexactness level $\bar{\delta} \geq 0$ is defined as:

$$\tilde{L}(x, \bar{x}; \bar{\delta}) := \max \left\{ \frac{2 \left(f(x) - \tilde{l}_f(x; \bar{x}) - \bar{\delta} \right)}{\|x - \bar{x}\|^2}, 0 \right\}. \quad (4.1)$$

This measure is well-behaved. In general, for a (k, L, δ) -apx-PWS function, if iterates x and y land on the same piece, $\tilde{L}(x, y; \bar{\delta}) \leq L$ as long as $\bar{\delta} \geq \delta$. This computable measure allows us to quantify the total progress of the algorithm. For a given sequence of iterate pairs $\{(l_i, r_i)\}$, we define the cumulative progress $S(t)$ as

$$S(t; \{(l_i, r_i)\}) := \sum_{l_i, r_i \leq t} \frac{1}{\tilde{L}(x^{r_i}, x^{l_i}; \Delta_0/6)(r_i - l_i)}.$$

This measure relates to the gap between the generated upper and lower bounds via $\bar{f}^t - \underline{f}^t \leq O(1/S(t))$. Therefore, to obtain the tightest bound, Algorithm 4.1 uses dynamic programming (Line 8) to find the sequence of iterate pairs that maximizes $S(t)$. Specifically, $\tilde{S}_l(t+1)$ denotes the maximal progress made if iterate $t+1$ is chosen as the beginning of a matching pair, and $\tilde{S}_r(t+1)$ represents the maximal progress made if iterate $t+1$ is chosen as the end of a matching pair.

For the complexity analysis of this scheme, the following empirical smoothness statistics associated with the generated iterates is useful.

Definition 7 (m -Pair Sequence and Empirical Smoothness Statistics) An m -*matching pair sequence* is a set of index pairs $\{(l_i, r_i)\}_{i=1}^p$ from an iterate sequence $\{x^i\}_{i=1}^N$ with non-overlapping intervals $(r_i \leq l_{i+1})$ where pairs are separated by at most m ($r_i \leq l_i + m + 1$). For such a sequence, we define:

- **Average inter-arrival time:** $\bar{\kappa} := \frac{N}{p}$.
- **Average length:** $\bar{\sigma} := p / \sum_{i=1}^p \frac{1}{r_i - l_i}$.
- **Average empirical smoothness:** $\bar{L} := \sum_{i=1}^p \frac{1}{r_i - l_i} / \sum_{i=1}^p \frac{1}{(r_i - l_i) \bar{L}(x^{r_i}, x^{l_i}, \Delta/6)}$.

These statistics are bounded for appropriately chosen m -pair sequences. If the objective function f is (k, L, δ) -apx PWS with $\delta < \Delta/6$ and the chosen sequence $\{(l_i, r_i)\}$ corresponds to a matching-pair sequence associated with the underlying pieces (see Definition 4), we have $\bar{L} \leq L$. Moreover, if $m \geq k$, there always exists some m -pair sequence such that $\bar{L} \leq L$, $\bar{\kappa} \leq O(1)k$ and $\bar{\sigma} \leq O(1)k$.

Now we are ready to state the convergence guarantee associated with Algorithm 4.2 in the following theorem.

Theorem 4.1 *Consider a convex objective function f that satisfies the Quadratic Growth (QG) condition with modulus $\mu > 0$. When Algorithm 4.2 is run with inputs (x^0, μ, m, ϵ) , it returns an ϵ -optimal solution in S outer iterations, where*

$$S \leq O(1) \cdot \left\lceil \log \left(\frac{\|f'(x^0)\|^2}{\mu\epsilon} \right) \right\rceil.$$

Moreover, the algorithm's oracle complexity is characterized as follows:

- a) Let $(\bar{L}_s, \bar{\kappa}_s, \bar{\sigma}_s)$ be the empirical statistics associated with some m -pair sequence (Definition 7) in the s -th call to the \mathcal{GR} subroutine. Define the worst-case statistics over all S calls as $\bar{L} := \max_{s \in [S]} \bar{L}_s$, $\bar{\kappa} := \max_{s \in [S]} \bar{\kappa}_s$, and $\bar{\sigma} := \max_{s \in [S]} \bar{\sigma}_s$. The total number of oracle evaluations is bounded by:

$$O(1) \cdot \left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\mu} \right\rceil \cdot \left\lceil \log \left(\frac{\|f'(x^0)\|^2}{\mu\epsilon} \right) \right\rceil.$$

By definition of the m -pair sequence, the statistics satisfy $\bar{\kappa} \leq m$ and $\bar{\sigma} \leq 2m$.

- b) If f is also a (k, L, δ) -apx-PWS function with $k \leq m$ and $6\delta \leq \epsilon$, then a matching pair sequence can be chosen such that the worst-case empirical smoothness is bounded by the true smoothness constant, i.e., $\bar{L} \leq L$.

The convergence result in Theorem 4.1 has important practical implications. First, the oracle complexity of $O\left(\frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\mu} \log\left(\frac{1}{\epsilon}\right)\right)$ depends on the statistics $\bar{\kappa}$ and $\bar{\sigma}$, which are determined by the local piecewise geometry near the solution set X^* . This is a key advantage, as the performance is not dictated by the function's global structure. This complexity is as efficient as methods that require the optimal value f^* to be known. Additionally, Algorithm 4.2 produces a gap sequence $\{\Delta_\ell\}$ that provides a verifiable certificate of suboptimality at termination. It is important to note, however, that the guaranteed geometric convergence rate relies on the target accuracy ϵ being sufficiently larger than the function's inexactness (e.g., $\epsilon \geq 9\delta$). While the gap certificate remains valid even in the high-accuracy regime where $\epsilon < \delta$, the empirical Lipschitz constant \bar{L} might increase and the convergence speed may degrade.

Second, the algorithm is notably practical as it only requires the quadratic growth parameter μ to run. Although our analysis assumes a (k, L, δ) -apx-PWS structure, the algorithm itself is independent of L , δ , and k . It is also robust to the choice of the number of cuts, m . Indeed, for any general M -Lipschitz continuous convex function, the method can be applied with $m \geq 1$ to achieve an oracle complexity of $O\left(\frac{M^2}{\mu\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)$. A more refined analysis, similar to that following Theorem 3.1, would show that the optimal oracle complexity of $O\left(\frac{M^2}{\mu\epsilon}\right)$ is attainable.

4.2 The Convergence Analysis of the BL- μ Method

The following lemma establishes the correctness of the GR subroutine (Algorithm 4.1). It guarantees that the subroutine successfully reduces the optimality gap and provides an upper bound on its iteration count.

Lemma 4.1 (Properties of the GR Subroutine) *Let f be a convex objective function satisfying the Quadratic Growth (QG) condition with modulus $\mu > 0$. Consider the iterates generated by the GR subroutine with inputs $(\mu, x^0, \bar{f}, \underline{f}, m)$, where $\bar{f} = f(x^0)$, $\underline{f} \leq f^*$, and the initial gap is $\Delta = \bar{f} - \underline{f}$. The following properties hold:*

- a) *The subroutine returns an updated iterate x^+ , an upper bound $\bar{f}^+ = f(x^+)$, and a valid lower bound \underline{f}^+ such that the new gap is reduced by a constant factor:*

$$\bar{f}^+ - \underline{f}^+ \leq \frac{2}{3}(\bar{f} - \underline{f}).$$

- b) *The subroutine terminates in at most $\bar{\tau} = \left\lceil \frac{3\bar{\kappa}\bar{\sigma}\bar{L}}{\mu} \right\rceil$ iterations, where $(\bar{\kappa}, \bar{\sigma}, \bar{L})$ are the empirical statistics from Definition 7 for the chosen m -pair sequence.*

- c) *If f is also a (k, L, δ) -apx-PWS function and the inputs satisfy $m \geq k$ and $\Delta \geq 6\delta$, then an m -pair sequence can be chosen such that its worst-case empirical smoothness is bounded by the true smoothness, i.e., $\bar{L} \leq L$.*

Proof Part a). The proof is by contradiction. Assume the subroutine terminates via the lower-bound update rule but the returned lower bound is invalid, i.e., $\underline{f}^+ > f^*$. This implies $f^t \geq f^*$ for all t up to the termination iteration τ . We show this leads to a contradiction.

If $f^t \geq f^*$, the optimal set X^* is contained in every level set $X(t)$. This ensures the standard non-expansive property of the iterates with respect to any $x^* \in X^*$:

$$\begin{aligned} \|x^t - x^*\|^2 + \|x^t - x^{t-1}\|^2 &\leq \|x^{t-1} - x^*\|^2, \quad \forall x^* \in X^*, t \in [\tau], \\ \Rightarrow \|x^t - x^*\|^2 + \frac{1}{t-\bar{t}}\|x^t - x^{\bar{t}}\|^2 &\leq \|x^{\bar{t}} - x^*\|^2, \quad \forall x^* \in X^*, \forall \bar{t} < t \leq \tau. \end{aligned}$$

Furthermore, for any $\bar{t} \in [0, t]$, since $\tilde{l}_f(x^t; x^{\bar{t}}) \leq f^t \leq \underline{f}^t + \frac{1}{3}\Delta_t \leq \underline{f}^t + \frac{1}{3}\Delta$, the definition of \tilde{L} in Definition 6 implies:

$$\begin{aligned} f(x^t) - \tilde{l}_f(x^t; x^{\bar{t}}) - \frac{\Delta}{6} &\leq \frac{\tilde{L}(t, \bar{t}; \frac{\Delta}{6})}{2} \|x^t - x^{\bar{t}}\|^2, \\ \Rightarrow \frac{2}{\tilde{L}(t, \bar{t}; \frac{\Delta}{6})(t - \bar{t})} \left[f(x^t) - \underline{f}^t - \frac{\Delta}{2} \right] &\leq \|x^t - x^{\bar{t}}\|^2. \end{aligned}$$

Combining the preceding relations yields a per-step progress inequality:

$$\frac{2}{\tilde{L}(t, \bar{t}; \frac{\Delta}{6})(t - \bar{t})} \left[f(x^t) - \underline{f}^t - \frac{\Delta}{2} \right] + \|x^t - x^*\|^2 \leq \|x^{\bar{t}} - x^*\|^2.$$

Summing this guarantee over the optimal matching pair sequence $\{(l_i, r_i)\}$ and using the monotonicity of $\{\|x^t - x^*\|\}$ gives:

$$\sum_{r_i \leq \tau} \frac{2}{\tilde{L}(r_i, l_i; \frac{\Delta}{6})(r_i - l_i)} \left[f(x^{r_i}) - \underline{f}^{r_i} - \frac{\Delta}{2} \right] \leq \|x^0 - x^*\|^2.$$

From the definition of $\tilde{S}_r(\tau)$ and the QG condition, it follows that:

$$\min_{r_i \leq \tau} \left[f(x^{r_i}) - \underline{f}^{r_i} - \frac{\Delta}{2} \right] \leq \frac{f(x^0) - f^*}{\mu \tilde{S}_r(\tau)} \leq \frac{\bar{f} - \underline{f}}{\mu \tilde{S}_r(\tau)} = \frac{\Delta}{\mu \tilde{S}_r(\tau)} \leq \frac{\Delta}{6}.$$

Let $r_{\bar{i}}$ be the index achieving this minimum. From the monotonicity of $\{\underline{f}^t\}$ and $\{\Delta_t\}$, we have:

$$\Delta_{r_{\bar{i}}} \leq f(x^{r_{\bar{i}}}) - \underline{f}^{l_{\bar{i}}} \leq \frac{\Delta}{2} + \frac{\Delta}{6} = \frac{2}{3}\Delta.$$

This shows the gap-reduction condition (Line 9) must have been met, contradicting the assumption that the algorithm terminated by the lower-bound update rule.

Part b). Assume for contradiction that the method runs for more than $\bar{\tau} + 1$ iterations. The optimality of the dynamic programming step ensures the progress measure satisfies:

$$\tilde{S}_r(\bar{\tau}) = \sum_{r_i \leq \bar{\tau}} \frac{1}{\tilde{L}(r_i, l_i; \frac{\Delta}{6})(r_i - l_i)} \geq \sum_{\hat{r}_i \leq \bar{\tau}} \frac{1}{\tilde{L}(\hat{r}_i, \hat{l}_i; \frac{\Delta}{6})(\hat{r}_i - \hat{l}_i)} \geq \frac{3}{\mu}.$$

This implies the termination condition in Lines 10-11 would have been triggered in iteration τ , a contradiction.

Part c). Since $m \geq k$, the pigeonhole principle guarantees the existence of an m -matching pair sequence $\{(x^{\bar{l}_i}, x^{\bar{r}_i})\}$ associated with the underlying pieces $\{X_i\}$. For each such pair, since $\Delta \geq 6\delta$, the definition of \tilde{L} gives:

$$\tilde{L}(x^{\bar{r}_i}, x^{\bar{l}_i}; \Delta/6) \leq \left[\frac{2(f(x^{\bar{r}_i}) - \tilde{l}_f(x^{\bar{r}_i}); x^{\bar{l}_i}) - \Delta/6}{\|x^{\bar{r}_i} - x^{\bar{l}_i}\|^2} \right]_+ \leq \left[\frac{2(f(x^{\bar{r}_i}) - \tilde{l}_f(x^{\bar{r}_i}); x^{\bar{l}_i}) - \delta}{\|x^{\bar{r}_i} - x^{\bar{l}_i}\|^2} \right]_+ \leq L.$$

By choosing a sequence of such pairs, we ensure $\bar{L} \leq L$. ■

We now use the convergence result for each GR subroutine to prove the main convergence result for the entire algorithm in Theorem 4.1.

Proof (Proof of Theorem 4.1) First, we establish the correctness of the algorithm. By Lemma 7.2, the initial value \underline{f}^0 is a valid lower bound on f^* . Lemma 4.1a) guarantees that each subsequent lower bound \underline{f}^s generated by the \mathcal{GR} subroutine also remains valid, i.e., $\underline{f}^s \leq f^*$ for all $s \geq 1$. Thus, when the algorithm terminates at an iteration S with the condition $f(x^S) - \underline{f}^S \leq \epsilon$, the returned solution x^S is guaranteed to be ϵ -optimal.

Next, we analyze the algorithm's efficiency to prove part a). Lemma 4.1a) ensures that each call to the GR subroutine reduces the gap $\Delta_s = \bar{f}^s - \underline{f}^s$ by a factor of at least $2/3$. The number of outer iterations S required to reduce the initial gap $\Delta_0 := 2\|f'(x^0)\|^2/\mu$ to ϵ is therefore bounded by $\log_{3/2}(\Delta_0/\epsilon)$, which gives the stated bound $S \leq O(1) \left\lceil \log\left(\frac{\|f'(x^0)\|^2}{\mu\epsilon}\right) \right\rceil$. The total oracle complexity is found by multiplying this number of outer iterations by the maximum number of iterations required for each \mathcal{GR} call, which is given in Lemma 4.1b).

Finally, part b) of the theorem is a direct consequence of Lemma 4.1c). The conditions assumed in Theorem 4.1b), namely that f is a (k, L, δ) -apx-PWS function with $k \leq m$ and $6\delta \leq \epsilon$, ensure that the prerequisite of Lemma 4.1c) holds for each subroutine call before the gap is reduced to ϵ . This guarantees the existence of an m -pair sequence such that the worst-case empirical smoothness \bar{L} is bounded by the true smoothness L . ■

4.3 The Weakly Convex Problem

We now consider finding approximate stationarity points for a weakly-convex apx-PWS function, for which the weak convexity constant ρ is known. We begin by reviewing the standard terminology for weakly-convex optimization [5].

Definition 8 (Weak Convexity and Moreau Stationarity) A function f is ρ -weakly convex if for any $\bar{x} \in X$, the surrogate function $F_{2\rho}(x; \bar{x})$ is ρ -strongly convex with respect to x , where

$$\begin{aligned} F_{2\rho}(x; \bar{x}) &:= f(x) + \rho\|x - \bar{x}\|^2, \\ f_{2\rho}(\bar{x}) &:= \min_{x \in X} F_{2\rho}(x; \bar{x}). \end{aligned} \tag{4.2}$$

Furthermore, a point \bar{x} is (ρ, ϵ) -Moreau stationary for some $\epsilon \geq 0$ if $\|\rho(\bar{x} - \hat{x})\| \leq \epsilon$, where $\hat{x} \leftarrow \arg \min_{x \in X} F_{2\rho}(x; \bar{x})$.

Algorithm 4.3 The Inexact Proximal Point Method for ρ -Weakly Convex Problems

Input: Starting point \bar{x}^0 , weakly convex modulus $\rho > 0$, number of cuts m .

```

1: for  $s = 0, 1, 2, \dots$  do
2:   Set up proximal subproblem:  $P_s(x) := f(x) + \rho\|x - \bar{x}^s\|^2$ ;  $\bar{P}^0 \leftarrow f(\bar{x}^s)$ ;  $\hat{x}^0 \leftarrow \bar{x}^s$ ;  $\underline{P}^0 \leftarrow \bar{P}^0 - \|f'(\bar{x}^s)\|^2/(2\rho)$ .
3:   for  $i = 1, 2, 3, \dots$  do
4:      $(\hat{x}^i, \bar{P}^i, \underline{P}^i) \leftarrow \mathcal{GR}(\rho, \hat{x}^{i-1}, \bar{P}^{i-1}, \underline{P}^{i-1}, m)$  with objective  $P_s(x)$ .
5:     if  $\bar{P}^0 - P_s(\hat{x}^i) \geq P_s(\hat{x}^i) - \underline{P}^i$  then
6:       Set  $\bar{x}^{s+1} \leftarrow \hat{x}^i$ , store the gap  $\Delta_s \leftarrow \bar{P}^0 - \underline{P}^i$ , and break.
7:     end if
8:   end for
9: end for

```

As shown in [5], the Moreau stationarity condition is equivalent to the gradient of the Moreau envelope $f_{2\rho}$ being small, i.e., $\|\nabla f_{2\rho}(\bar{x})\| \leq \epsilon$. The following lemma connects this stationarity measure to the function value gap of the surrogate problem. Its proof is deferred to the appendix.

Lemma 4.2 *For a ρ -weakly convex function f , let $F_{2\rho}$ and $f_{2\rho}$ be the perturbed function and its Moreau envelope defined in (4.2). The gradient of the Moreau envelope is bounded by the function value gap as follows:*

$$\|\nabla f_{2\rho}(\bar{x})\|^2 \leq 8\rho \left(F_{2\rho}(\bar{x}; \bar{x}) - \min_{x \in X} F_{2\rho}(x; \bar{x}) \right) = 8\rho (f(\bar{x}) - f_{2\rho}(\bar{x})).$$

The result from Lemma 4.2 shows that Moreau stationarity at a point \bar{x} is bounded by the potential descent in the proximal objective function $F_{2\rho}(x; \bar{x})$. This motivates using the inexact proximal point method (IPPM) to find an approximately stationary point. The method, shown in Algorithm 4.3, generates a sequence of prox-centers $\{\bar{x}^s\}$, where each new center \bar{x}^{s+1} is found by approximately solving the surrogate problem $F_{2\rho}(x; \bar{x}^s)$.

More specifically, in each outer iteration s , Algorithm 4.3 constructs the ρ -strongly convex proximal subproblem $P_s(x)$. It then uses a procedure based on the \mathcal{GR} subroutine (Algorithm 4.1) to find a new iterate \hat{x}^i that achieves at least half of the maximal possible descent on $P_s(x)$ from the prox-center \bar{x}^s . Once this condition is met, the inner loop terminates and the algorithm proceeds to the next outer iteration.

The convergence of this scheme can be analyzed using the original objective f as a potential function. In each outer iteration s , the update to the new prox-center \bar{x}^{s+1} decreases the objective value f by at least $\Delta_s/2$, where $\Delta_s := P_s(\bar{x}^s) - \min_{x \in X} P_s(x)$. Since the total possible descent, $f(\bar{x}^0) - \min_{x \in X} f(x)$, is finite, the sequence of potential descents $\{\Delta_s\}$ must converge to zero. As Lemma 4.2 ties the stationarity measure to Δ_s , we have $\lim_{s \rightarrow \infty} \|\nabla f_{2\rho}(\bar{x}^s)\| = 0$. A more concrete, finite-time convergence guarantee is provided in the next theorem.

Theorem 4.2 *Consider an M -Lipschitz continuous and ρ -weakly convex objective function f that is bounded from below, i.e., $\Delta_f := f(\bar{x}^0) - \min_{x \in X} f(x) < \infty$. When run with inputs (\bar{x}^0, ρ, m) , Algorithm 4.3 generates an iterate \bar{x}^S that is (ρ, ϵ) -Moreau stationary, satisfying $\Delta_S \leq \epsilon^2/(8\rho)$, in a number of outer loops S bounded by:*

$$S \leq \left\lceil 16\rho\Delta_f/\epsilon^2 \right\rceil.$$

Moreover, the algorithm's oracle complexity is characterized as follows:

a) Let $(\bar{L}_{s,i}, \bar{\kappa}_{s,i}, \bar{\sigma}_{s,i})$ be the empirical statistics associated with some m -pair sequence (see Definition 7) from the i -th \mathcal{GR} call for the s -th subproblem. Let $\bar{L}, \bar{\kappa}, \bar{\sigma}$ be the maximum of these statistics over the entire run. The total number of oracle evaluations is bounded by

$$O(1) \frac{\Delta_f}{\epsilon^2} \left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\rho} \right\rceil \left\lceil \log \frac{M}{\epsilon} \right\rceil = O(1) \frac{\Delta_f \bar{\kappa} \bar{\sigma} \max\{\bar{L}, \rho\}}{\epsilon^2} \log \left(\frac{M}{\epsilon} \right),$$

where the statistics satisfy $\bar{\kappa} \leq 2m$ and $\bar{\sigma} \leq m$.

b) If f is also a (k, L, δ) -apx-PWS function with $k \leq m$ and $\delta \leq \epsilon^2/(48\rho)$, then an m -pair sequence can be chosen such that the empirical smoothness satisfies $\bar{L} \leq L + 2\rho$.

We make two comments regarding the convergence result. First, it is useful to situate our complexity result in the context of more familiar settings.

- For smooth, non-convex objectives (i.e., a $(1, L)$ -PWS function), the oracle complexity bound simplifies to $O\left(\frac{\Delta_f L}{\epsilon^2} \log \frac{M}{\epsilon}\right)$, which nearly matches the well-known complexity of gradient descent [21, 22].
- For general non-smooth, non-convex objectives (i.e., a $(1, M^2\rho/\epsilon^2, \epsilon^2/(16\rho))$ -apx-PWS function), the complexity becomes $O\left(\frac{\Delta_f \rho M^2}{\epsilon^4}\right)$, which is consistent with the current state-of-the-art result in [5].

The main advantage of Algorithm 4.3 is for objectives that are PWS but not smooth. In this case, our method improves upon the general non-smooth rate of $O(1/\epsilon^4)$ to a much faster "smooth-like" rate of $\tilde{O}(1/\epsilon^2)$, with the additional factors depending on the local PWS geometry.

Second, the method is practical to implement. It automatically adapts to the local smoothness and requires only two parameters: the number of cuts m and the weak convexity constant ρ . The algorithm is quite robust to the choice of m (e.g., $m = 10$ is often sufficient). It is, however, sensitive to the misspecification of ρ . We address this challenge in the subsequent sections by developing a verifiable criterion to detect if ρ is misspecified.

4.4 The Convergence Analysis

Proof (Proof of Theorem 4.2) Let \bar{x}^S denote the first prox-center to satisfy the termination condition $\bar{\Delta}_S \leq \epsilon^2/(8\rho)$. We will calculate the number of oracle evaluations required to generate this solution.

Bounding the number of outer iterations. We first calculate an upper bound on S using a descent argument on f . For any outer iteration $s < S$, the algorithm does not terminate, which implies $\bar{\Delta}_s > \epsilon^2/(8\rho)$. The termination condition for the inner loop in Algorithm 4.3 is

$$\bar{P}_s(\bar{x}^s) - P_s(\bar{x}^{s+1}) \geq P_s(\bar{x}^{s+1}) - \underline{P}_s^i,$$

which implies $2[P_s(\bar{x}^s) - P_s(\bar{x}^{s+1})] \geq P_s(\bar{x}^s) - \underline{P}_s^i \geq \bar{\Delta}_s$. This gives the following descent guarantee for each outer iteration:

$$\begin{aligned} f(\bar{x}^s) - f(\bar{x}^{s+1}) &\geq f(\bar{x}^s) - [f(\bar{x}^{s+1}) + \rho\|\bar{x}^{s+1} - \bar{x}^s\|^2] \\ &= P_s(\bar{x}^s) - P_s(\bar{x}^{s+1}) \\ &\geq \frac{1}{2}\bar{\Delta}_s. \end{aligned}$$

Summing this over all iterations from $s = 0$ to $S - 1$ yields:

$$\frac{S\epsilon^2}{16\rho} \leq \frac{1}{2} \sum_{s=0}^{S-1} \bar{\Delta}_s \leq \sum_{s=0}^{S-1} (f(\bar{x}^s) - f(\bar{x}^{s+1})) = f(\bar{x}^0) - f(\bar{x}^S) \leq f(\bar{x}^0) - \min_{x \in X} f(x) = \Delta_f.$$

Rearranging gives the bound on the number of outer loops, $S \leq \lceil 16\rho\Delta_f/\epsilon^2 \rceil$.

Bounding the total complexity. Next, we bound the number of oracle evaluations within each outer loop s . Since f is M -Lipschitz, the initial gap of the subproblem is bounded: $\Delta_s^0 \leq M^2/(2\rho)$. The subproblem $P_s(x)$ satisfies the QG condition with modulus ρ . Since we would have triggered the break condition in Line 6 when $\Delta_s^i \leq \epsilon^2/8\rho$, an argument similar to that for Theorem 4.1 shows that the number of gradient evaluations required to solve this subproblem is bounded by:

$$O(1) \left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\rho} \right\rceil \left\lceil \log \frac{M^2/(2\rho)}{\epsilon^2/(8\rho)} \right\rceil = O(1) \left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\rho} \right\rceil \left\lceil \log \frac{M}{\epsilon} \right\rceil.$$

Combining this with the bound on the number of outer loops S gives the desired total oracle complexity.

The apx-PWS case. Finally, if f is a (k, L, δ) -apx-PWS function, then the proximal function $P_s(x)$ is a $(k, L + 2\rho, \delta)$ -apx-PWS function. Given that the gap remains above $\epsilon^2/(8\rho)$ before termination, an argument similar to that in the proof of Lemma 4.1.c) shows that the empirical smoothness \bar{L} is appropriately bounded.



5 Verifiable Termination Condition

The soundness of both the μ -BL method (Algorithm 4.2) and the IPPM method (Algorithm 4.3) relies on knowing a lower-curvature parameter. In practice, however, this parameter is often unavailable. This section addresses this issue by developing a novel \mathcal{W} -stationarity certificate for PWS functions, which can be used to verify algorithmic progress without this prior knowledge.

This section is organized as follows. First, we propose the normalized Wolfe-gap stationarity (\mathcal{W} -stationarity) certificate and compare it to more familiar termination conditions to illustrate its unique advantages. Second, we present an algorithm to compute a certificate for \mathcal{W} -stationarity. In the next section, we will apply this certificate to design almost parameter-free versions of our bundle-level method for both the convex quadratic growth and the weakly convex settings.

5.1 Requirements for a Termination Criterion

Before presenting our \mathcal{W} -stationarity certificate, it is useful to recall why the gradient norm, $\|\nabla f(\bar{x})\|$, is the widely accepted termination certificate for the simpler setting of smooth and strongly convex optimization [22]. Two properties are essential. First, it is easily *verifiable*; given a point \bar{x} , one can compute $\|\nabla f(\bar{x})\|$ without knowledge of any other problem parameters. Second, the gradient norm provides an accurate characterization of the optimality gap:

$$\frac{1}{2L}\|\nabla f(\bar{x})\|^2 \leq f(\bar{x}) - f^* \leq \frac{1}{2\mu}\|\nabla f(\bar{x})\|^2.$$

The right-hand inequality shows that the gradient norm provides a computable *upper bound* on the optimality gap, while the left-hand inequality shows that it is *proportional* to the gap. These two properties are critical for developing optimal parameter-free algorithms for smooth optimization, as in [22].

In the PWS setting, the gradient norm remains verifiable. However, it fails to be proportional to the function value gap due to non-smoothness. This shortcoming prevents its use in developing parameter-free algorithms and prompts the search for a more suitable termination certificate. To that end, we formalize the essential requirements for such a certificate.

Definition 9 (Requirements for a Reasonable Certificate) We call a termination certificate $V(\bar{x})$ for an apx-PWS function *reasonable* if it satisfies the following conditions:

- Handles Constraints:** The certificate can be applied to problems with a simple, closed, convex feasible region X .
- Verifiable:** The certificate can be verified by calling the black-box first-order oracle without requiring any problem parameters (e.g., L or μ).
- Provides an Upper Bound:** Under the QG setting, the certificate provides a computable upper bound on the optimality gap, $f(\bar{x}) - f^*$.
- Computable & Proportional:** For any feasible point \bar{x} , there exists a finite-time, first-order method to generate a certificate whose value is proportional to the optimality gap.

It is illuminating to see how frequently used stationarity certificates fare on these requirements. As summarized in Table 1, these common criteria are lacking in one way or another, which motivates our development of the \mathcal{W} -stationarity certificate.

- The *optimality gap*, $f(x^t) - f^*$, provides an ideal characterization of progress. However, it is not a practical certificate because f^* cannot be computed or verified by any first-order method.

Certificate	Verifiable	Optimality Gap			Handles Constraints
		Upper Bound	Proportional	Computable	
Optimality Gap	×	✓	✓	×	✓
Gradient Mapping	✓	✓	×	✓	✓
Moreau Stationarity	✓	✓	✓	×	✓
Approx. Moreau	×	✓	✓	✓	✓
Goldstein Stationarity	✓	✓	?	✓	×
\mathcal{W} -Stationarity (This work)	✓	✓	✓	✓	✓

Table 1: Comparison of termination certificates for PWS problems against the criteria from Definition 9.

- The *gradient norm* (or gradient mapping for constrained problems) is computationally attractive, requiring only a single gradient evaluation. As discussed, however, it fails to be proportional to the optimality gap in the PWS setting. For example, with $f(x) := |x|$, we have $\|\nabla f(x)\| = 1$ for all $x > 0$, while the gap $f(x) - f^* = x$ approaches zero.
- The *Moreau stationarity* certificate, $M_\eta(\bar{x}) := \eta(\bar{x} - \hat{x}_\eta)$ where $\hat{x}_\eta \leftarrow \arg \min_{x \in X} f(x) + \frac{\eta}{2}\|x - \bar{x}\|^2$, is proportional to the optimality gap but is not computable in finite time because it requires finding the exact minimizer \hat{x}_η [5]. Approximating \hat{x}_η requires yet another termination certificate for the proximal subproblem.
- The *Goldstein stationarity* certificate is another popular choice for non-smooth problems [43]. Its primary drawbacks are its inability to handle constrained optimization and its reliance on a carefully chosen radius parameter $\delta(\bar{x})$ to maintain proportionality with the function value gap.

5.2 The \mathcal{W} -Stationarity Certificate

Inspired by the Wolfe-gap termination condition and the Goldstein stationarity certificate [43], we propose the following \mathcal{W} -stationarity certificate to meet the requirements in Definition 9.

Definition 10 (\mathcal{W} -Stationarity Certificate) Given an apx-PWS function f , an evaluation center $\bar{x} \in X$, a radius $\iota > 0$, and a set of search points $\{x^i\}_{i \in [m]} \subset B(\bar{x}; \iota)$, we first define the model function $\psi(x)$ as the maximum of several linear approximations:

$$\psi(x) := \max \left\{ \max_{i \in [m]} \{\tilde{l}_f(x; x^i)\}, l_f(x; \bar{x}) \right\},$$

where $l_f(x; \bar{x}) := f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle$ is the linear support function at the evaluation center. The normalized Wolfe-gap (or \mathcal{W} -gap) is then defined as the maximal rate of descent on this model within a local ball:

$$\mathcal{V}_\iota(\bar{x}; \{x^i\}_{i \in [m]}) := \frac{1}{\iota} \left\{ \max_{x \in X} (\psi(\bar{x}) - \psi(x)) \quad \text{s.t.} \quad \|x - \bar{x}\| \leq \iota \right\}. \quad (5.1)$$

We say the set $\{x^i\}_{i \in [m]}$ constitutes a valid (ι, ν) - \mathcal{W} -stationarity certificate for \bar{x} if $\|x^i - \bar{x}\| \leq \iota$ for all $i \in [m]$, and $\mathcal{V}_\iota(\bar{x}) \leq \nu$. Moreover, if the search points are clear from the context, we would drop $\{x^i\}_{i \in [m]}$ to write $\mathcal{V}_\iota(\bar{x})$ for simplicity.

The \mathcal{W} -gap measures the steepness of the model ψ in the neighborhood $B(\bar{x}; \iota)$ of the point of interest \bar{x} . As $\iota \rightarrow 0_+$, the \mathcal{W} -gap reduces to the gradient norm $\|\nabla f(\bar{x})\|$. Our proposed certificate differs from the classic Wolfe-gap in three important ways:

- It utilizes a model function $\psi(x)$ constructed from a *bundle* of linear supports from multiple points, which is essential for handling the non-smoothness of PWS functions.
- It focuses on descent within a local ι -ball, which is a less conservative stationarity measure than considering descent over the entire feasible region X .

- It restricts the evaluation points $\{x^i\}$ to lie within the same ι -ball, a feature that is important for handling non-convexity.

We now discuss how the proposed \mathcal{W} -certificate meets several of the requirements outlined in Definition 9. The analysis of its computability and proportionality is deferred to the following subsections.

First, for any fixed set of evaluation points, the \mathcal{W} -gap is computable by solving a quadratically constrained quadratic program (QCQP). This makes the certificate *verifiable* without knowledge of any hidden problem parameters. Second, since the feasible region \mathcal{X} is incorporated into its definition, the \mathcal{W} -gap naturally *handles constraints*.

Showing that the \mathcal{W} -certificate provides a meaningful *upper bound* on the optimality gap requires a preliminary monotonicity result. The following lemma shows that for a fixed set of search points, the \mathcal{W} -gap is a monotonically non-increasing function of its radius ι .

Lemma 5.1 *Fixing the search points $\{x^i\}_{i \in [m]}$, the \mathcal{W} -gap $\mathcal{V}_\iota(\bar{x})$ is a monotonically non-increasing function of its radius ι for any $\iota \geq \max_{i \in [m]} \|x^i - \bar{x}\|$.*

Proof Let $\iota_1 \geq \iota_2$ be given, and let x_{ι_1} and x_{ι_2} be the respective optimal solutions in the definition of the \mathcal{W} -gap (5.1). By the convexity of \mathcal{X} , the point $\tilde{x} := \bar{x} + \frac{\iota_2}{\iota_1}(x_{\iota_1} - \bar{x})$ is also in \mathcal{X} . Since the model $\psi(x)$ is convex and the points \bar{x} , \tilde{x} , and x_{ι_1} are collinear, the property of secant lines for convex functions implies:

$$\frac{\psi(\tilde{x}) - \psi(\bar{x})}{\|\tilde{x} - \bar{x}\|} \leq \frac{\psi(x_{\iota_1}) - \psi(\bar{x})}{\|x_{\iota_1} - \bar{x}\|}.$$

Since $\|\tilde{x} - \bar{x}\| = \iota_2$ and $\|x_{\iota_1} - \bar{x}\| = \iota_1$, this is equivalent to:

$$\frac{\psi(\bar{x}) - \psi(\tilde{x})}{\iota_2} \geq \frac{\psi(\bar{x}) - \psi(x_{\iota_1})}{\iota_1}.$$

By definition of the \mathcal{W} -gap, $\mathcal{V}_{\iota_2}(\bar{x}) = \frac{\psi(\bar{x}) - \psi(x_{\iota_2})}{\iota_2} \geq \frac{\psi(\bar{x}) - \psi(\tilde{x})}{\iota_2}$. Combining these inequalities gives the desired result:

$$\mathcal{V}_{\iota_2}(\bar{x}) \geq \mathcal{V}_{\iota_1}(\bar{x}).$$

■

The next lemma shows that the \mathcal{W} -certificate provides an upper bound on the optimality gap for a convex function satisfying the QG condition. Importantly, the lemma only requires the QG condition to hold at the evaluation center \bar{x} , not globally. This relaxation is instrumental for providing bounds for general convex functions.

Lemma 5.2 *Given a convex apx-PWS objective function f , suppose there exists an (ι, ν) - \mathcal{W} -stationarity certificate for a feasible point \bar{x} . If \bar{x} satisfies the QG condition $f(\bar{x}) - f^* \geq \frac{\mu}{2} \text{dist}^2(\bar{x}, X^*)$ for some $\mu > 0$, then the \mathcal{W} -stationarity certificate implies the following error bound condition:*

$$\text{dist}(\bar{x}, X^*) \leq \max\left\{\tau, \frac{\nu}{\mu}\right\}.$$

Moreover its optimality gap is bounded by:

$$f(\bar{x}) - f^* \leq \max\left\{\iota\nu, \frac{2\nu^2}{\mu}\right\}. \quad (5.2)$$

Proof By definition, $f(\bar{x}) \leq \psi(\bar{x})$. We derive a lower bound on f^* by considering two cases based on the location of the projection $x_p^* := \text{proj}_{X^*}(\bar{x})$.

Case 1: $\|\bar{x} - x_p^*\| \leq \iota$. In this case, x_p^* is inside the ball used to define the \mathcal{W} -gap. Since f is convex, $f(x_p^*) \geq \psi(x_p^*)$. We have:

$$f(\bar{x}) - f^* \leq \psi(\bar{x}) - \psi(x_p^*) \leq \max_{x \in B(\bar{x}; \iota) \cap X} (\psi(\bar{x}) - \psi(x)) = \iota \mathcal{V}_\iota(\bar{x}) \leq \iota\nu.$$

Case 2: $\|\bar{x} - x_p^*\| > \iota$. We first bound the optimality gap by the distance. Using the monotonicity of the \mathcal{W} -gap from Lemma 5.1, we get:

$$f(\bar{x}) - f^* \leq \psi(\bar{x}) - \psi(x_p^*) \leq \|\bar{x} - x_p^*\| \mathcal{V}_{\|\bar{x} - x_p^*\|}(\bar{x}) \leq \|\bar{x} - x_p^*\| \mathcal{V}_\iota(\bar{x}) \leq \nu \|\bar{x} - x_p^*\|.$$

Now, we incorporate the QG condition with the inequality we just derived:

$$\frac{\mu}{2} \|\bar{x} - x_p^*\|^2 \leq f(\bar{x}) - f^* \leq \nu \|\bar{x} - x_p^*\|.$$

Solving this for the distance gives $\|\bar{x} - x_p^*\| \leq \frac{2\nu}{\mu}$. Substituting this back into our gap inequality yields:

$$f(\bar{x}) - f^* \leq \nu \|\bar{x} - x_p^*\| \leq \nu \left(\frac{2\nu}{\mu} \right) = \frac{2\nu^2}{\mu}.$$

Combining the two cases gives the desired result. ■

The bound from (5.2) has two implications. First, in the QG setting where the objective satisfies the condition with a modulus $\mu > 0$, we can choose the radius $\iota \leq 2\nu/\mu$, which simplifies the bound to $f(\bar{x}) - f^* \leq O(\nu^2/\mu)$ and $\text{dist}(\bar{x}, X^*) \leq \nu/\mu$. This guarantee is analogous to the one provided by the standard gradient norm for smooth optimization. Second, for a general convex function where a global QG condition does not hold, we can still derive a meaningful bound. By defining a local QG modulus, $\mu_{\text{loc}} := 2(f(\bar{x}) - f^*)/D_X^2$, where D_X is the diameter of the feasible set \mathcal{X} , and substituting it into the bound from Lemma 5.2, we get:

$$f(\bar{x}) - f^* \leq \frac{2\nu^2 D_X^2}{f(\bar{x}) - f^*} \implies (f(\bar{x}) - f^*)^2 \leq 2\nu^2 D_X^2.$$

This implies an optimality gap of $f(\bar{x}) - f^* \leq \sqrt{2\nu} D_X$, which is also similar to the upper bound provided by the gradient norm in the general convex setting.

Non-convex Setting

We now show that in the non-convex setting, our \mathcal{W} -stationarity certificate implies other well-known stationarity conditions.

Lemma 5.3 *For the unconstrained setting ($X = \mathbb{R}^n$), suppose the linear support function $\tilde{l}_f(x; x^i)$ is computed using a gradient at a nearby point \tilde{x}^i , i.e., $\nabla \tilde{l}_f(x; x^i) = \nabla f(\tilde{x}^i)$ with $\|x^i - \tilde{x}^i\| \leq \bar{\delta}$. Then a (ι, ν) - \mathcal{W} -stationarity certificate for a point \bar{x} is also a $(\iota + \bar{\delta}, \nu)$ -Goldstein stationarity certificate [43].*

Proof For simplicity, let $\tilde{l}_f(x; x^0) := f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle$ denote the linear approximation constructed from the evaluation center \bar{x} . Let $\bar{\lambda}$ be the optimal dual variables from the definition of the \mathcal{W} -gap, i.e., $\bar{\lambda} := \arg \max_{\lambda \in \Delta_{m+1}^+} \min_{x \in B(\bar{x}, \iota)} \sum_{i=0}^m \lambda_i \tilde{l}_f(x, x^i)$. Let $g = \sum_{i=0}^m \bar{\lambda}_i \nabla f(\tilde{x}^i)$. It suffices to show that $\|g\| \leq \nu$.

Since $\mathcal{V}_\iota(\bar{x}) \leq \nu$, from the definition of the \mathcal{W} -gap we have:

$$\psi(\bar{x}) - \sum_{i=0}^m \bar{\lambda}_i \tilde{l}_f(x; x^i) \leq \nu \iota, \quad \forall x \in B(\bar{x}; \iota).$$

Using the definition $\psi(\bar{x}) = \max_{i \in [m]} \tilde{l}_f(\bar{x}; x^i)$ and the fact that $\sum \bar{\lambda}_i = 1$, we can deduce:

$$\begin{aligned} \sum_{i=0}^m \bar{\lambda}_i \left(\tilde{l}_f(\bar{x}; x^i) - \tilde{l}_f(x; x^i) \right) &\leq \nu \iota \\ \implies \sum_{i=0}^m \bar{\lambda}_i \langle \nabla f(\tilde{x}^i), \bar{x} - x \rangle &\leq \nu \iota \\ \implies \langle g, \bar{x} - x \rangle &\leq \nu \iota, \quad \forall x \in B(\bar{x}; \iota). \end{aligned}$$

By choosing the specific point $x = \bar{x} - \iota \frac{g}{\|g\|}$ from the ball, we get $\iota \|g\| \leq \nu \iota$, which simplifies to $\|g\| \leq \nu$.

■

Furthermore, if the objective function is weakly convex, the \mathcal{W} -certificate also implies Moreau stationarity, which is the standard benchmark in this setting. The next lemma shows that an (ι, ν) - \mathcal{W} -certificate implies a point is $O(\nu)$ -Moreau stationary, provided the radius ι is chosen appropriately relative to ν and the weak convexity constant ρ . This result provides practical guidance on how to select the radius parameter in our algorithms.

Lemma 5.4 *Given a ρ -weakly convex function f , if there exists an (ι, ν) - \mathcal{W} -stationarity certificate for a point \bar{x} , then \bar{x} is a $(\rho, 2\nu + 4\iota\rho)$ -Moreau stationary point.*

Proof The proof centers on the surrogate function $F_{2\rho}(x; \bar{x}) := f(x) + \rho\|x - \bar{x}\|^2$ and involves two steps.

Step 1: Transfer the \mathcal{W} -certificate to the surrogate function. We first show that an (ι, ν) - \mathcal{W} -certificate for f implies an $(\iota, \nu + 2\rho\iota)$ - \mathcal{W} -certificate for $F_{2\rho}(x; \bar{x})$. Let $\psi_f(x)$ be the model for f from the certificate's search points $\{x^i\}$, and let $\psi_{F_{2\rho}}(x)$ be the corresponding model for $F_{2\rho}$. For any $x \in B(\bar{x}; \iota)$, the difference between the two models is bounded by:

$$|\psi_f(x) - \psi_{F_{2\rho}}(x)| \leq \rho \max_{i \in [m]} |\langle x^i - \bar{x}, x - x^i \rangle| \leq \rho\iota^2.$$

The \mathcal{W} -gap for the surrogate function is therefore bounded. Specifically, we have:

$$\max_{x \in B(\bar{x}; \iota)} (\psi_{F_{2\rho}}(\bar{x}) - \psi_{F_{2\rho}}(x)) \leq \max_{x \in B(\bar{x}; \iota)} (\psi_f(\bar{x}) - \psi_f(x)) + 2\rho\iota^2 \leq \nu + 2\rho\iota^2.$$

This confirms that we have an $(\iota, \nu + 2\rho\iota)$ - \mathcal{W} -certificate for the surrogate function $F_{2\rho}$.

Step 2: Derive Moreau stationarity from the surrogate's certificate. Since $F_{2\rho}$ is ρ -strongly convex, we can apply Lemma 5.2 to it. This gives an upper bound on the optimality gap of the surrogate problem:

$$\begin{aligned} \frac{\rho}{2}\|\bar{x} - \hat{x}\|^2 &\leq F_{2\rho}(\bar{x}; \bar{x}) - \min_{x \in X} F_{2\rho}(x; \bar{x}) \leq \max \left\{ \iota(\nu + 2\rho\iota), \frac{2(\nu + 2\rho\iota)^2}{\rho} \right\} \\ &\leq \frac{2}{\rho}(\nu + 2\rho\iota)^2. \end{aligned}$$

Rearranging this inequality gives $\|\rho(\bar{x} - \hat{x})\| \leq \sqrt{4(\nu + 2\rho\iota)^2} = 2\nu + 4\rho\iota$, which is the desired Moreau stationarity bound.

■

Thus, we have illustrated that the proposed \mathcal{W} -stationarity certificate satisfies the *verifiability*, *providing upper bound*, and *handling constraints* requirements specified in Definition 9. The next subsection proposes an algorithm to efficiently compute a proportional \mathcal{W} -stationarity certificate.

5.3 Computing the \mathcal{W} -stationarity Certificate

This subsection introduces a method for computing a \mathcal{W} -stationarity certificate for convex functions using a first-order oracle. We establish the utility of this certificate in two key settings. First, under the Quadratic Growth (QG) condition, we show the certificate yields an upper bound proportional to the true optimality gap. Second, for weakly convex problems, we apply the method to the proximal surrogate function to obtain an useful non-convex stationarity certificate associated with the original objective function. The ability to compute these certificates is instrumental for the almost parameter-free algorithms we develop later.

As detailed in **Algorithm 5.1**, our proposed method computes a \mathcal{W} -stationarity certificate for a convex function f . The subroutine takes as input the evaluation center \bar{x} , an estimated optimality gap Δ , the number of search points m , and a maximal search radius ι_{\max} . The radius parameter ι_{\max} is particularly useful for the non-convex setting, as discussed preceding Lemma 5.4.

The core of the algorithm is an iterative process. In each iteration t , it generates a new search point x^t by projecting the center point \bar{x} onto a level set (Line 3). Crucially, the level-set parameter

Algorithm 5.1 The \mathcal{W} -Certificate Search Subroutine

Input: Convex objective f ; candidate solution \bar{x} ; estimated optimality gap Δ ; number of evaluation points m ; maximal radius ι_{\max} .

Output: A (ι, ν) \mathcal{W} -stationarity certificate if $f(\bar{x}) - f^* \leq \Delta$; otherwise, False.

1: **Initialize:** $x^0 \leftarrow \bar{x}$, $t \leftarrow 0$, $l \leftarrow f(\bar{x}) - 2\Delta$, and $l_f(x; x^0) \leftarrow f(x^0) + \langle f'(x^0), x - x^0 \rangle$.

2: **for** $t = 0, 1, \dots, m$ **do**

3: Compute the projection:

$$x^{t+1} \leftarrow \arg \min_{x \in X(t)} \|x - \bar{x}\|^2 \text{ where } X(t) := \{x \in \mathcal{X} \mid l_f(x; x^0) \leq l \text{ and } \tilde{l}_f(x; x^i) \leq l, \forall i \in \{1, \dots, t\}\}$$

4: **if** the computation is infeasible, set $\|x^{t+1} - \bar{x}\| = +\infty$.

5: **if** $\|x^{t+1} - \bar{x}\| > \iota_{\max}$, **return** $\{x^i\}_{i \in [t]}$ as the $(\iota_{\max}, \frac{2\Delta}{\iota_{\max}})$ \mathcal{W} -stationarity certificate.

6: **if** $\|x^{t+1} - \bar{x}\| = +\infty$, **return** $\{x^i\}_{i \in [t]}$ as the $(\mathbb{M}, \frac{2\Delta}{\mathbb{M}})$ \mathcal{W} -stationarity certificate.

7: **end for**

8: Compute empirical smoothness constant: $\tilde{L} \leftarrow \min_{1 \leq l < r \leq m+1} \frac{2(f(x^r) - \tilde{l}_f(x^r; x^l) - \Delta/6)}{\|x^l - x^r\|^2}$.

9: **if** $\|\bar{x} - x^{m+1}\| < \frac{1}{2} \sqrt{\frac{\Delta}{\tilde{L}}}$ **return** False.

10: Set radius: $\iota \leftarrow \|\bar{x} - x^{m+1}\|$, **return** search points $\{x^i\}_{i \in [m]}$ as an $(\iota, 2\Delta/\iota)$ \mathcal{W} -stationarity certificate.

is set to $l = f(\bar{x}) - 2\Delta$, a value intentionally chosen to be below the estimated optimal value f^* . This encourages the search to explore different smooth pieces of the function. By construction, the distance $\|x^t - \bar{x}\|$ increases monotonically with each iteration. The subroutine terminates under one of several conditions:

- If the search distance $\|x^{t+1} - \bar{x}\|$ exceeds the maximal radius ι_{\max} , the algorithm returns a certificate based on this radius (Line 5).
- If the projection becomes infeasible, it implies that the objective function is lower-bounded by l . The algorithm returns a certificate (Line 6). For analytical purposes, we use a large constant \mathbb{M} as a proxy for an infinite radius, and the returned certificate remains valid as $\mathbb{M} \rightarrow \infty$.
- If the subroutine completes $m+1$ iterations without early termination, it calculates the empirical smoothness \tilde{L} among all generated points. It then either returns ‘False’ if the final search point is too close to the center (Line 9) or returns a valid \mathcal{W} -certificate based on the final distance (Line 10).

The validity of this procedure is formally established in the subsequent lemma.

Lemma 5.5 *If the objective function f is convex, the iterates $\{x^i\}$ generated by **Algorithm 5.1** with inputs \bar{x}, Δ, m , and ι_{\max} satisfy the following properties:*

- a) *The sequence of distances to the center, $\{\|x^i - \bar{x}\|\}$, is monotonically non-decreasing.*
- b) *Any \mathcal{W} -stationarity certificate returned by the algorithm (from Lines 5, 6, or 10) is valid (see Definition 10), with its value ν bounded by $\nu \leq 4\sqrt{\Delta\tilde{L}}$.*
- c) *If f is a (k, L, δ) -apx-PWS function with $k \leq m$ and the inexactness $\delta \leq \frac{1}{6}\Delta$, then the empirical smoothness is bounded by the true smoothness, i.e., $\tilde{L} \leq L$.*
- d) *If the estimated optimality gap is well-specified (i.e., $f(\bar{x}) - f^* \leq \Delta$), the algorithm will not return ‘False’.*

Proof Let $\psi^t(x) := \max\{\max_{i \leq t} \tilde{l}_f(x; x^i), l_f(x; x^0)\}$.

Part a): As long as the projection in Line 3 is feasible, the iterate x^t is defined as the closest point to \bar{x} in the set $X(t-1) := \{x \in \mathcal{X} : \psi^{t-1}(x) \leq l\}$. By definition, the next iterate x^{t+1} must satisfy $\psi^{t-1}(x^{t+1}) \leq \psi^t(x^{t+1}) \leq l$, which implies that $x^{t+1} \in X(t-1)$. Therefore, $\|x^{t+1} - \bar{x}\| \geq \|x^t - \bar{x}\|$. If the projection to compute x^j becomes infeasible, we adopt the convention that $\|x^j - \bar{x}\| = +\infty$, and the monotonicity of the sequence $\{\|x^i - \bar{x}\|\}$ holds.

Part b): Let \bar{t} be the iteration in which the algorithm returns a \mathcal{W} -stationarity certificate. Since f is convex, $\psi^{\bar{t}}(\bar{x}) = f(\bar{x})$. We need only show that the maximal descent of the model, $\psi^{\bar{t}}(\bar{x}) - \psi^{\bar{t}}(x)$, is bounded by 2Δ within the ball $B(\bar{x}; \iota)$ for the corresponding radius $\iota \in \{\iota_{\max}, M, \|x^{m+1} - \bar{x}\|\}$.

- If the return is triggered by infeasibility (Line 6), then $\psi^{\bar{t}}(x) \geq l$ for all $x \in X$. Thus, for a sufficiently large \mathbb{M} , we have $\max_{x \in \mathcal{X} \cap B(\bar{x}; \mathbb{M})} (\psi^{\bar{t}}(\bar{x}) - \psi^{\bar{t}}(x)) \leq f(\bar{x}) - l = 2\Delta$.

- If the return is triggered by exceeding the maximal radius (Line 5), then for any point $x \in B(\bar{x}; \iota_{\max})$, we must have $\psi^{\bar{t}}(x) > l$, otherwise the projection $x^{\bar{t}+1}$ would have $\|x^{\bar{t}+1} - \bar{x}\| \leq \iota_{\max}$, a contradiction. Therefore, $\psi^{\bar{t}}(\bar{x}) - \psi^{\bar{t}}(x) \leq f(\bar{x}) - l \leq 2\Delta$ for all $x \in B(\bar{x}; \iota_{\max}) \cap X$. The argument for Line 10 is analogous.

Part c): Since $k \leq m$, the pigeonhole principle guarantees that among the $m+1$ points $\{x^i\}_{i=1}^{m+1}$, there exists a pair (x^i, x^j) that lie on the same smooth piece. For this pair, the (k, L, δ) -apx-PWS property implies

$$f(x^j) \leq \tilde{l}_f(x^j; x^i) + \frac{L}{2} \|x^i - x^j\|^2 + \delta.$$

Given that $\delta \leq \frac{\Delta}{6}$, the definition of the empirical smoothness constant \tilde{L} yields

$$\begin{aligned} \tilde{L} &:= \min_{0 \leq l < r \leq m+1} \frac{2(f(x^r) - \tilde{l}_f(x^r; x^l) - \frac{\Delta}{6})}{\|x^l - x^r\|^2} \leq \frac{2(f(x^j) - \tilde{l}_f(x^j; x^i) - \frac{\Delta}{6})}{\|x^j - x^i\|^2} \\ &\leq \frac{2(f(x^j) - \tilde{l}_f(x^j; x^i) - \delta)}{\|x^j - x^i\|^2} \leq L. \end{aligned}$$

Part d): Assume for the sake of contradiction that the algorithm returns ‘False’, which means $\|\bar{x} - x^{m+1}\| < \frac{1}{2} \sqrt{\frac{\Delta}{\tilde{L}}}$. Let (x^i, x^j) with $i < j$ be the pair of iterates that defines the empirical smoothness constant \tilde{L} . By the monotonicity established in part a), we have $\|x^i - \bar{x}\| \leq \|x^j - \bar{x}\| \leq \|x^{m+1} - \bar{x}\|$, which implies $\|x^i - x^j\| \leq \|x^i - \bar{x}\| + \|x^j - \bar{x}\| \leq \sqrt{\frac{\Delta}{\tilde{L}}}$.

The iterate x^j is feasible for the projection subproblem defining x^i (for $i < j$), so $\tilde{l}_f(x^j; x^i) \leq l$. This leads to the following chain of inequalities:

$$\begin{aligned} f(x^j) &\leq \tilde{l}_f(x^j; x^i) + \frac{\tilde{L}}{2} \|x^i - x^j\|^2 + \frac{\Delta}{6} \\ &\leq l + \frac{\tilde{L}}{2} \left(\sqrt{\frac{\Delta}{\tilde{L}}} \right)^2 + \frac{\Delta}{6} \\ &= (f(\bar{x}) - 2\Delta) + \frac{\Delta}{2} + \frac{\Delta}{6} = f(\bar{x}) - \frac{4}{3}\Delta. \end{aligned}$$

Since $f(\bar{x}) - \frac{4}{3}\Delta < f(\bar{x}) - \Delta$, and we assumed $f(\bar{x}) - \Delta \leq f^*$, this implies $f(x^j) < f^*$. This contradicts the definition of f^* as the minimum objective value. ■

5.4 Proportional \mathcal{W} -Stationarity Certificate under the QG Setting

The preceding lemma verifies that our proposed search method returns a valid \mathcal{W} -stationarity certificate when the estimated function value gap is well-specified. We now show that this certificate is also proportional to the gap estimate in the QG setting.

Proposition 5.1 *Given a convex and (k, L, δ) -apx-PWS objective function f , consider running Algorithm 5.1 with inputs \bar{x} , Δ , m , and $\iota_{\max} = +\infty$. The output satisfies the following properties.*

- If Δ is a valid upper bound on the optimality gap (i.e., $f(\bar{x}) - f^* \leq \Delta$), then the algorithm is guaranteed to return a \mathcal{W} -certificate from either Line 6 or Line 10.*
- If the algorithm terminates at Line 6, the returned \mathcal{W} -certificate implies an optimality gap bound of 2Δ .*
- If the algorithm terminates at Line 10 and the objective satisfies the μ -QG condition, the returned certificate implies an optimality gap of at most $\max\{2, \frac{32\tilde{L}}{\mu}\}\Delta$, where \tilde{L} is the empirical smoothness constant from Line 8. If, in addition, $\delta \leq \frac{1}{6}\Delta$ and $m \geq k$, this bound tightens to $\max\{2, \frac{32L}{\mu}\}\Delta$.*

Proof Part a): Since $\iota_{\max} = +\infty$, the termination condition in Line 5 is never met. As the input Δ is a valid upper bound on the optimality gap, Lemma 5.5.d) ensures that the algorithm will not return ‘False’ from Line 9. Therefore, the algorithm must terminate by returning a certificate from either Line 6 or Line 10.

Part b): If the certificate is returned from Line 6, it is an (ι, ν) - \mathcal{W} -certificate with $\iota = \mathbb{M}$ (for some large M) and $\nu = 2\Delta/\mathbb{M}$. Plugging these into the bound from Lemma 5.2 gives:

$$f(\bar{x}) - f^* \leq \lim_{\mathbb{M} \rightarrow +\infty} \max \left\{ \frac{2\Delta}{\mathbb{M}}, \frac{2(2\Delta/\mathbb{M})^2}{\mu} \right\} = 2\Delta.$$

Part c): If the certificate is returned from Line 10, its parameters are $\iota = \|x^{m+1} - \bar{x}\|$ and $\nu = 2\Delta/\iota$. The condition for termination in this line is $\iota \geq \frac{1}{2}\sqrt{\Delta/\tilde{L}}$, which implies $\nu \leq 4\sqrt{\Delta\tilde{L}}$. Applying the bound from Lemma 5.2 yields:

$$f(\bar{x}) - f^* \leq \max \left\{ \iota\nu, \frac{2\nu^2}{\mu} \right\} \leq \max \left\{ 2\Delta, \frac{2(4\sqrt{\Delta\tilde{L}})^2}{\mu} \right\} = \max \left\{ 2, \frac{32\tilde{L}}{\mu} \right\} \Delta.$$

Furthermore, if $\Delta \geq 6\delta$ and $m \geq k$, Lemma 5.5.c) guarantees that $\tilde{L} \leq L$, which gives the implied upper bound. ■

A few remarks are in order regarding this result. First, Proposition 5.1 demonstrates that the optimality gap guarantee implied by the \mathcal{W} -stationarity certificate is proportional to the gap estimate Δ , up to the condition number $\sqrt{L/\mu}$. This satisfies the final requirement for a “reasonable” certificate from Definition 9. Second, a key practical advantage is that to obtain this proportional optimality bound, the number of cuts m only needs to be larger than the number of smooth pieces in the local neighborhood $B(\bar{x}; \iota)$, rather than the total number of pieces globally. Finally, the result highlights why a bundle of cuts is necessary. If one sets $m = 1$ and treats a general M -Lipschitz continuous function as a $(1, M^2/\Delta, \Delta)$ -apx-PWS function, the empirical condition number \tilde{L}/μ can become arbitrarily large. In this case, the certificate value $\nu = O(\sqrt{\Delta\tilde{L}})$ can degrade to $O(M)$ even when the point \bar{x} is near optimal. This is expected, as a single-cut model is equivalent to the subgradient norm, which cannot provide a proportional optimality guarantee for general non-smooth functions.

5.5 Proportional \mathcal{W} -Stationarity Certificate under the Weakly Convex Setting

We now apply Algorithm 5.1 to the proximal surrogate function to generate a \mathcal{W} -stationarity certificate for the non-convex setting that is proportional to the Moreau stationarity criterion.

Proposition 5.2 *Consider a ρ -weakly convex function f and a point \bar{x} that is approximately Moreau stationary, such that $F_{2\rho}(\bar{x}; \bar{x}) - \min_{x \in X} F_{2\rho}(x; \bar{x}) \leq \Delta_{\bar{x}} = \epsilon^2/\rho$. When Algorithm 5.1 is applied to the surrogate function $F_{2\rho}(\cdot; \bar{x})$ with inputs $(\bar{x}, \Delta_{\bar{x}}, m, \iota_{\max} = \sqrt{\Delta_{\bar{x}}/\rho})$, the returned (ι, ν) - \mathcal{W} -certificate has the following properties.*

- a) *The certificate parameters satisfy $\nu \geq 2\iota\rho$ and $\nu \leq \max\{2\epsilon, 4\sqrt{\tilde{L}/\rho\epsilon}\}$, where \tilde{L} is the empirical smoothness constant from Line 8.*
- b) *The same search points also constitute a $(\iota, 2\nu)$ - \mathcal{W} -stationarity certificate for the original objective function f .*
- c) *If the surrogate function $F_{2\rho}(\cdot; \bar{x})$ is a (k, L, δ) -apx-PWS function with $m \geq k$ and $\delta \leq \Delta_{\bar{x}}/6$, then the empirical smoothness is bounded by $\tilde{L} \leq L$.*

Proof For simplicity, let $P_{\bar{x}}(x) := F_{2\rho}(x; \bar{x})$.

Part a): Since $\Delta_{\bar{x}}$ is an upper bound on the optimality gap of the convex function $P_{\bar{x}}$, Lemma 5.5 ensures that Algorithm 5.1 successfully returns a \mathcal{W} -certificate for $P_{\bar{x}}$. The condition $\iota \leq \iota_{\max}$ implies $\nu/\iota = 2\Delta_{\bar{x}}/\iota^2 \geq 2\Delta_{\bar{x}}/\iota_{\max}^2 = 2\rho$, which gives the lower bound $\nu \geq 2\iota\rho$.

For the upper bound on ν , we consider two cases for how the algorithm terminates. If termination occurs at $\iota = \iota_{\max}$, then $\nu = 2\Delta_{\bar{x}}/\iota_{\max} = 2\sqrt{\rho\Delta_{\bar{x}}} = 2\epsilon$. Otherwise, termination implies $\iota \geq \frac{1}{2}\sqrt{\Delta_{\bar{x}}/\tilde{L}}$, which gives the bound $\nu = 2\Delta_{\bar{x}}/\iota \leq 4\sqrt{\tilde{L}\Delta_{\bar{x}}} \leq 4\sqrt{\tilde{L}/\rho\epsilon}$. Combining these cases gives the result.

Part b): We now translate the certificate for $P_{\bar{x}}$ to one for f . Let $\psi_{P_{\bar{x}}}$ and ψ_f be the respective model functions. From the proof of Lemma 5.4, we know that $\max_{x \in B(\bar{x}; \iota)} (\psi_f(\bar{x}) - \psi_f(x)) \leq \iota\nu + 2\iota^2\rho$. Using the lower bound from part (a), $\nu \geq 2\iota\rho$, we have:

$$\max_{x \in B(\bar{x}; \iota)} (\psi_f(\bar{x}) - \psi_f(x)) \leq \iota\nu + \iota(2\iota\rho) \leq \iota\nu + \iota\nu = \iota(2\nu).$$

Thus, the search points form a valid $(\iota, 2\nu)$ - \mathcal{W} -stationarity certificate for the original function f .

Proof of (c). This follows directly from Lemma 5.5(c). Since the surrogate function $F_{2\rho}$ is (k, L, δ) -apx-PWS and the condition $\delta \leq \Delta_{\bar{x}}/6$ holds, the lemma guarantees that the empirical smoothness constant \tilde{L} computed by the algorithm is bounded by the true constant L .

Two remarks are in order. First, this result, combined with Lemma 5.4, shows that our generated \mathcal{W} -certificate is proportional to the Moreau stationarity measure, up to a factor related to the condition number $\sqrt{L/\rho}$. Second, the guarantee is reliable up to the inexactness parameter δ ; if the target gap $\Delta_{\bar{x}}$ is smaller than 6δ , the empirical smoothness constant \tilde{L} may no longer be bounded by L .

6 Almost Parameter-Free Bundle Level Method

In this section, we use the proposed \mathcal{W} -stationary certificate to design almost parameter-free algorithms for convex QG problems and weakly convex problems. Since the \mathcal{W} -certificate is both verifiable and proportional (see Definition 9), we use it to track if an algorithm is making the desired progress and adjust the algorithm accordingly in an adaptive fashion. This allows our methods to take advantage of unknown growth conditions in the apx-PWS non-smooth setting.

6.1 Convex Problems with an Unknown QG Constant

Algorithm 6.1 Almost Parameter-Free BL method (pfBL- μ) for Convex QG Objectives

Input: An initial point $x^0 \in X$; an upper bound on the QG modulus, $\tilde{\mu}$; the number of cuts m .

- 1: Initialize: $\bar{f}^0 \leftarrow f(x^0)$, $\underline{f}^0 \leftarrow f(x^0) - \frac{2\|f'(x^0)\|^2}{\tilde{\mu}}$, $\tau \leftarrow 0$, and $\Delta_0 \leftarrow \bar{f}^0 - \underline{f}^0$.
 - 2: **for** $s = 0, 1, 2, \dots$ **do**
 - 3: Run $(x^{s+1}, \bar{f}^{s+1}, \underline{f}^{s+1}) \leftarrow \mathcal{GR}(\tilde{\mu}, x^s, \bar{f}^s, \underline{f}^s)$ with objective f and set $\Delta_{s+1} \leftarrow \bar{f}^{s+1} - \underline{f}^{s+1}$.
 - 4: **if** $\Delta_{s+1} \leq \frac{1}{2}\Delta_\tau$ **then**
 - 5: Set $\tau \leftarrow s + 1$.
 - 6: Run Algorithm 5.1 on f with inputs $(\bar{x}^\tau, \Delta_\tau, m, +\infty)$ to generate a \mathcal{W} -stationarity certificate.
 - 7: **if** Algorithm 5.1 returns *False*, restart Algorithm 6.1 with inputs $(x^{s+1}, \tilde{\mu}/2, m)$.
 - 8: **end if**
 - 9: **end for**
-

We first consider the convex setting where the QG constant μ is unknown and the goal is to minimize the function value gap. The proposed method, Algorithm 6.1, produces a sequence of iterates $\{x^s\}$ whose objective values converge to f^* . Its implementation requires only an initial point x^0 , the number of cuts m , and an initial upper bound on the QG constant, $\tilde{\mu}$. As discussed in Lemma 7.2(b), this upper bound can be computed efficiently, making the method very practical.

More concretely, Algorithm 6.1 employs a "guess-and-check" strategy. It utilizes the μ -BL method (Algorithm 4.2) with a guessed QG constant $\tilde{\mu}$ to shrink the optimality gap Δ_s . Once the gap halves (Line 4), it runs the \mathcal{W} -certificate search algorithm (Line 6) to verify this progress. A successful verification implies that progress has been made. If the search algorithm returns *False*, it signals that the guess $\tilde{\mu}$ was too large. In this latter case (Line 7), the algorithm restarts the

entire scheme with a halved guess, $\tilde{\mu}/2$. The following theorem establishes the validity of this adaptive procedure by bounding the number of iterations required to find an ϵ -optimal solution.

Theorem 6.1 *Consider an M -Lipschitz continuous convex objective function f that satisfies the QG condition with modulus $\mu > 0$. When Algorithm 6.1 is run with inputs $(x^0, \tilde{\mu}, m)$, where $\tilde{\mu}$ is an initial upper bound on μ , it generates an ϵ -optimal solution. The number of restarts, Q , is bounded by $Q \leq O(1) \lceil \log(\tilde{\mu}/\mu) \rceil$. Moreover, the algorithm's efficiency is characterized as follows:*

- a) *Let $(\bar{L}_{q,s}, \bar{\kappa}_{q,s}, \bar{\sigma}_{q,s})$ be the empirical statistics from the s -th GR call within the q -th restart (see Definition 7). Let $\bar{L}, \bar{\kappa}, \bar{\sigma}$ be the maximum of these statistics, and let \hat{L} be the maximal empirical smoothness constant observed during calls to Algorithm 5.1³ before an ϵ -optimal solution is generated. The total number of required oracle evaluations is bounded by:*

$$O(1) \left(\left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\mu} \right\rceil + m \left\lceil \log \frac{\tilde{\mu}}{\mu} \right\rceil \right) \left\lceil \log \frac{M^2 \max\{\hat{L}, \tilde{\mu}\}}{\mu\epsilon} \right\rceil.$$

- b) *If f is also a (k, L, δ) -apx-PWS function with $k \leq m$ and $\delta \leq \frac{\epsilon\mu}{200L}$, there exists a choice of m -pair sequence such that the empirical constants are bounded by the true constants, i.e., $\bar{L} \leq L$ and $\hat{L} \leq L$.*

Proof For clarity in the analysis, we index quantities by the restart counter q . Thus, $\tilde{\mu}_q$ is the guessed QG modulus and $x^{q,s}$ is the s -th iterate during the q -th restart phase.

First, we establish a sufficient condition for ϵ -optimality. Let $\bar{\Delta} := \mu\epsilon/(32 \max\{\hat{L}, \tilde{\mu}\})$ denote the target accuracy. If Algorithm 5.1 is called with a gap $\Delta_{q,s} \leq \bar{\Delta}$, then Proposition 5.1 guarantees that the returned \mathcal{W} -certificate implies the solution is ϵ -optimal, regardless of whether the search terminates due to infeasibility (Line 6) or after completing all iterations (Line 10).

Next, we calculate the number of iterations required to reach this gap $\bar{\Delta}$ within a fixed restart phase q . The initial gap is bounded by $\Delta_{q,0} = 2\|f'(x^{q,0})\|^2/\tilde{\mu}_q \leq 2M^2/\tilde{\mu}_q$. The number of \mathcal{GR} evaluations needed to shrink this gap to $\bar{\Delta}$ is at most $\left\lceil \log(\frac{2M^2}{\tilde{\mu}_q\bar{\Delta}})/\log(3/2) \right\rceil$. From Lemma 4.1(b), this corresponds to a total of

$$\left\lceil \log \frac{2M^2}{\tilde{\mu}_q\bar{\Delta}} / \log(3/2) \right\rceil \left(\left\lceil \frac{3\bar{\kappa}\bar{\sigma}\bar{L}}{\tilde{\mu}_q} \right\rceil + m + 1 \right)$$

gradient evaluations for this phase, if it is not interrupted by a restart.

The algorithm stops restarting once $\tilde{\mu}_q \leq \mu$. The number of gradient evaluations in this final, successful phase is therefore upper bounded by $\left\lceil \log(\frac{4M^2}{\mu\bar{\Delta}})/\log(3/2) \right\rceil (\left\lceil \frac{4\bar{\kappa}\bar{\sigma}\bar{L}}{\mu} \right\rceil + m + 1)$. Summing the work from this final phase with the work from all prior, aborted phases yields the total oracle complexity bound stated in the theorem.

Finally, for part (b), the condition $\delta \leq \frac{\epsilon\mu}{200L}$ ensures that for any relevant gap $\Delta \geq \bar{\Delta}$, we have $6\delta \leq \Delta$. This allows the application of Lemma 4.1.c) and Lemma 5.5.c), which together provide the claimed bound on \bar{L} . ■

A few remarks regarding this complexity result are in order. First, for a (k, L, δ) -PWS function, the oracle complexity is dominated by the term $O(\frac{Lk^2}{\mu} \log(\frac{1}{\epsilon}))$. This matches the complexity of Algorithms 4.2 and 3.1, but removes the need to know the QG parameter μ or the optimal value f^* beforehand. This adaptivity to the true curvature is a significant practical advantage. In particular, if the feasible region \mathcal{X} is bounded by a diameter D_X , the algorithm achieves an oracle complexity of $O(\frac{Lk^2 D_X^2}{\epsilon} \log(\frac{1}{\epsilon}))$ even if the QG condition does not hold globally. Second, the proposed method is an anytime algorithm, as it does not require the target accuracy ϵ as an input. It is guaranteed to converge quickly until the optimality gap is on the order of the inexactness parameter δ . Beyond this point, the empirical Lipschitz constant \bar{L} may grow, leading to slower convergence, as established in Theorem 6.1.

³ If Line 8 is never invoked, \hat{L} is regarded as zero.

6.2 Nonconvex Problems with an Unknown Weak Convexity Constant

Algorithm 6.2 An Almost Parameter-Free IPPM for ρ -Weakly Convex Problems

Input: Initial point $\bar{x}^0 \in X$; initial guess for the weak convexity modulus, $\tilde{\rho} > 0$; number of cuts m ; stationarity requirement $\epsilon > 0$.

Output: A solution \hat{x} and an associated ϵ - \mathcal{W} -stationarity certificate.

```

1: for  $s = 0, 1, 2, \dots$  do
2:   Set up the subproblem and initialize its state:  $P_s(x) := f(x) + \tilde{\rho}\|x - \bar{x}^s\|^2$ ;  $\bar{P}^0 \leftarrow f(\bar{x}^s)$ ;  $\hat{x}^0 \leftarrow \bar{x}^s$ ; and  $\underline{P}^0 \leftarrow \bar{P}^0 - \|f'(\bar{x}^s)\|^2/(2\tilde{\rho})$ .
3:   for  $i = 1, 2, 3, \dots$  do
4:      $(\hat{x}^i, \bar{P}^i, \underline{P}^i) \leftarrow \mathcal{GR}(\tilde{\rho}, \hat{x}^{i-1}, \bar{P}^{i-1}, \underline{P}^{i-1}, m)$  with objective  $P_s(x)$ .
5:     if  $\bar{P}^0 - P_s(\hat{x}^i) \geq P_s(\hat{x}^i) - \underline{P}^i$  then
6:       Set  $\bar{x}^{s+1} \leftarrow \hat{x}^i$  and  $\Delta_s \leftarrow \bar{P}^0 - \underline{P}^i$ .
7:       Run Algorithm 5.1 on  $P_s$  with inputs  $(\bar{x}^s, \Delta_s, m, \sqrt{\Delta_s/\tilde{\rho}})$  to generate an  $(\iota_s, \nu_s)$ - $\mathcal{W}$ -certificate.
8:       Evaluate the corresponding  $(\iota_s, \nu_s^+)$ - $\mathcal{W}$ -certificate associated with the same search points for the original function  $f$  using Definition 10.
9:       if Algorithm 5.1 returns False or  $\nu_s^+ \geq 2\nu_s$  then
10:        Restart with a new guess for  $\rho$ : Call Algorithm 6.2 with inputs  $(\bar{x}^{s+1}, 2\tilde{\rho}, m, \epsilon)$ .
11:       else if  $\nu_s^+ \leq \epsilon$  then
12:         return  $\bar{x}^s$  and its  $(\iota_s, \nu_s^+)$ - $\mathcal{W}$ -stationarity certificate.
13:       else
14:         break (Continue to the next outer iteration  $s+1$ .)
15:       end if
16:     end if
17:   end for
18: end for

```

We now consider the non-convex apx-PWS setting where the weak convexity modulus ρ is unknown. Using the IPPM method (Algorithm 4.3) with a misspecified constant $\tilde{\rho}$ poses two significant challenges. First, if $\tilde{\rho} < \rho$, the proximal subproblems may not be convex, rendering the lower bounds from the \mathcal{GR} subroutine unreliable. Second, and more critically, there is no guarantee that descent on the surrogate problem is proportional to some stationarity measure for the true objective. This invalidates the descent argument used in Theorem 4.2. This issue has been particularly challenging in the literature due to the lack of verifiable and proportional stationarity measures for non-smooth, non-convex problems.

We address this issue by using our proposed \mathcal{W} -stationarity certificate as the measure of progress. The key idea is that the \mathcal{W} -certificate can be computed without knowing ρ . We can therefore check if the observed descent in the proximal subproblem is proportional to the generated certificate. If it is not, we can infer that our guess $\tilde{\rho}$ is misspecified and adapt it accordingly.

This “guess-and-check” strategy is implemented in Algorithm 6.2. The method takes an initial point x^0 , an initial guess for the weak convexity constant $\tilde{\rho}$, the number of cuts m , and a stationarity requirement ϵ . It proceeds by running the IPPM method with the current guess $\tilde{\rho}$. After sufficient progress is made on a subproblem (Line 5), it calls the \mathcal{W} -certificate search subroutine. It then checks if the computed certificate is consistent with the observed progress (Line 7). If not, this indicates that $\tilde{\rho}$ was too small, so the algorithm restarts with a doubled guess, $2\tilde{\rho}$. The efficiency of this adaptive method is established in the next theorem.

Theorem 6.2 Consider an M -Lipschitz continuous, ρ -weakly convex objective function f that is bounded from below with $f(x^0) - \min_{x \in X} f(x) \leq \Delta_f < +\infty$. Let Algorithm 6.2 be run with inputs $(x^0, \tilde{\rho}_0, m, \epsilon)$, where $\tilde{\rho}_0 > 1$ is an initial guess for the weak convexity modulus. The algorithm generates an (ι, ν) - \mathcal{W} -stationary point \bar{x} with $\nu \leq \epsilon$ and $\iota \leq \epsilon$.

Moreover, let $\rho_{\max} := \max\{\tilde{\rho}_0, \rho\}$. The efficiency of the algorithm is characterized as follows:

- a) Let $(\bar{L}_{q,s,i}, \bar{\kappa}_{q,s,i}, \bar{\sigma}_{q,s,i})$ be the empirical statistics from the i -th \mathcal{GR} call for the s -th subproblem within the q -th restart. Let $\bar{L}, \bar{\kappa}, \bar{\sigma}$ be the maximum of these statistics over the entire run, and let \hat{L} be the maximal empirical smoothness from calls to Algorithm 5.1.⁴ The total number of

⁴ If Line 8 is never invoked, \hat{L} is taken to be zero.

oracle evaluations is bounded by:

$$O(1) \frac{\Delta_f(\bar{L} + \rho_{\max})}{\epsilon^2} \left(\left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\tilde{\rho}_0} \right\rceil \left\lceil \log \frac{M(\hat{L} + \rho_{\max})}{\epsilon} \right\rceil + m \log \frac{\rho_{\max}}{\tilde{\rho}_0} \right),$$

where the m -pair statistics satisfy $\bar{\kappa} \leq 2m$ and $\bar{\sigma} \leq m$.

- b) If f is also a (k, L, δ) -apx-PWS function with $k \leq m$ and $\delta \leq \frac{\epsilon^2}{2500(\rho_{\max} + L)}$, then a m -pair sequence can be chosen such that the complexity bound in a) holds with the empirical constants bounded by the true problem parameters, i.e., $\bar{L} \leq L + 2\rho_{\max}$ and $\hat{L} \leq L + 2\rho_{\max}$.

Proof For the analysis, we label quantities by the restart counter q . For instance, $\tilde{\rho}_q$ is the guessed weak convexity modulus during the q -th restart phase.

First, we observe that the restart condition in Line 9 is not triggered if the guess is accurate, i.e., if $\tilde{\rho}_q \geq \rho$. Under this condition, the \mathcal{GR} subroutine's lower bounds are well-specified (Lemma 4.1), so the \mathcal{W} -certificate search will not return *False*. Furthermore, Proposition 5.2 ensures that the generated certificate is consistent with the observed descent (i.e., $\nu_s^+ < 2\nu_s$).

Next, for any fixed q and s , we show that achieving a gap $\Delta_{q,s} \leq \Delta_\epsilon := \epsilon^2/(144(\rho_{\max} + \hat{L}))$ is a sufficient condition for finding an (ϵ, ϵ) - \mathcal{W} -stationary point. Assuming the algorithm does not restart, Proposition 5.2 shows that the generated (ι, ν) - \mathcal{W} -certificate for f satisfies:

$$\iota \leq \sqrt{\Delta_\epsilon / \tilde{\rho}_q} \quad \text{and} \quad \nu \leq 12\sqrt{\max\{\hat{L}, \rho_{\max}\} \Delta_\epsilon},$$

which implies $\iota \leq \epsilon$ and $\nu \leq \epsilon$.

Now we bound the oracle complexity. For any fixed restart phase q , the algorithm must either terminate or restart before the subproblem gap $\Delta_{q,s}$ falls below Δ_ϵ . By the descent argument from Theorem 4.2, the number of calls to the \mathcal{W} -certificate search subroutine within this phase is at most $\lceil \Delta_f / \Delta_\epsilon \rceil$. The number of oracle evaluations associated with the \mathcal{GR} calls is bounded by an argument similar to that in Theorem 4.2:

$$O(1) \frac{\Delta_f}{\Delta_\epsilon} \left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\tilde{\rho}_q} \right\rceil \left\lceil \log \frac{M}{\Delta_\epsilon} \right\rceil \leq O(1) \frac{\Delta_f(\rho_{\max} + \hat{L})}{\epsilon^2} \left\lceil \frac{\bar{L}\bar{\kappa}\bar{\sigma}}{\tilde{\rho}_q} \right\rceil \log \left(\frac{M}{\epsilon} \right).$$

The total complexity for phase q is the sum of the cost of the certificate searches ($O(\lceil \frac{\Delta_f(\rho_{\max} + \hat{L})}{\epsilon^2} \rceil m)$) and the cost of the \mathcal{GR} calls. Since the algorithm must terminate in the first phase q where $\tilde{\rho}_q \geq \rho$, we can sum the complexity over all restart phases to recover the overall bound from the theorem statement.

Finally, part b) follows immediately from Lemma 4.1.c) and Lemma 5.5.c). ■

A few remarks regarding this result are in order. First, this is, to our knowledge, the first almost parameter-free algorithm for apx-PWS weakly convex optimization and the first to provide a verifiable termination criterion in the constrained setting. Second, for a (k, L, δ) -apx-PWS function, the oracle complexity is worse than that of Algorithm 4.3 (with known ρ) by a factor of $\lceil L/\tilde{\rho}_0 \rceil$. This dependence arises because the \mathcal{W} -certificate's proportionality to the true error depends on this condition number (see Propositions 5.1 and 5.2), which requires solving the subproblems to a higher accuracy to verify progress. This suggests a practical guideline: it is likely better to choose an initial guess $\tilde{\rho}_0$ that is closer to the Lipschitz smoothness constant L , rather than a smaller value. Finally, the method can be viewed as an anytime algorithm. If the termination condition in Line 12 of Algorithm 6.2 is removed, the method will continuously run to find a (ρ, ϵ) -Moreau stationary point for any $\epsilon > 0$, with the total number of oracle evaluations in this case being bounded by

$$O \left(\frac{\Delta_f \bar{\kappa} \bar{\sigma}}{\epsilon^2} \max \left\{ \frac{L^2}{(\tilde{\rho}_0)^2}, \left(\frac{\rho}{\tilde{\rho}_0} \right)^2 \right\} \log \frac{M(L + \rho_{\max})}{\epsilon} \right).$$

References

1. A. BECK, First-order methods in optimization, vol. 25, SIAM, 2017.
2. A. BECK AND M. TEOULLE, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM journal on imaging sciences, 2 (2009), pp. 183–202.
3. A. BEN-TAL AND A. NEMIROVSKI, Lectures on modern convex optimization: analysis, algorithms, and engineering applications, SIAM, 2001.
4. ———, Non-euclidean restricted memory level method for large-scale convex optimization, Mathematical Programming, 102 (2005), pp. 407–456.
5. D. DAVIS AND D. DRUSVYATSKIY, Stochastic model-based minimization of weakly convex functions, SIAM Journal on Optimization, 29 (2019), pp. 207–239.
6. D. DAVIS AND L. JIANG, A local nearly linearly convergent first-order method for nonsmooth functions with quadratic growth, Foundations of Computational Mathematics, (2024), pp. 1–82.
7. W. DE OLIVEIRA, Proximal bundle methods for nonsmooth dc programming, Journal of Global Optimization, 75 (2019), pp. 523–563.
8. W. DE OLIVEIRA, C. SAGASTIZÁBAL, AND C. LEMARÉCHAL, Convex proximal bundle methods in depth: a unified analysis for inexact oracles, Mathematical Programming, 148 (2014), pp. 241–277.
9. Q. DENG, G. LAN, AND Z. LIN, Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization, arXiv preprint arXiv:2412.06319, (2024).
10. M. DÍAZ AND B. GRIMMER, Optimal convergence rates for the proximal bundle method, SIAM Journal on Optimization, 33 (2023), pp. 424–454.
11. D. DRUSVYATSKIY AND A. S. LEWIS, Error bounds, quadratic growth, and linear convergence of proximal methods, Mathematics of operations research, 43 (2018), pp. 919–948.
12. Y. DU AND A. RUSZCZYŃSKI, Rate of convergence of the bundle method, Journal of Optimization Theory and Applications, 173 (2017), pp. 908–922.
13. J. C. DUCHI, P. L. BARTLETT, AND M. J. WAINWRIGHT, Randomized smoothing for stochastic optimization, SIAM Journal on Optimization, 22 (2012), pp. 674–701.
14. X. HAN AND A. S. LEWIS, Survey descent: A multipoint generalization of gradient descent for nonsmooth optimization, SIAM Journal on Optimization, 33 (2023), pp. 36–62.
15. W. HARE AND C. SAGASTIZÁBAL, A redistributed proximal bundle method for nonconvex optimization, SIAM Journal on Optimization, 20 (2010), pp. 2442–2473.
16. H. KARIMI, J. NUTINI, AND M. 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, Springer, 2016, pp. 795–811.
17. J. E. KELLEY, JR, The cutting-plane method for solving convex programs, Journal of the society for Industrial and Applied Mathematics, 8 (1960), pp. 703–712.
18. K. C. KIWIEL, Efficiency of proximal bundle methods, Journal of Optimization Theory and Applications, 104 (2000), pp. 589–603.
19. K. KURDYKA, On gradients of functions definable in o-minimal structures, in Annales de l’institut Fourier, vol. 48, 1998, pp. 769–783.
20. G. LAN, Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization, Mathematical Programming, 149 (2015), pp. 1–45.
21. ———, First-order and stochastic Optimization Methods for Machine Learning, Springer-Nature, 2020.
22. G. LAN, Y. OUYANG, AND Z. ZHANG, Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization, arXiv e-prints, (2023), pp. arXiv–2310.
23. G. LAN AND Z. ZHANG, Optimal methods for convex risk-averse distributed optimization, SIAM Journal on Optimization, 33 (2023), pp. 1518–1557.
24. C. LEMARÉCHAL, An extension of davidon methods to non differentiable problems, in Nondifferentiable optimization, Springer, 2009, pp. 95–109.
25. C. LEMARÉCHAL, A. NEMIROVSKII, AND Y. NESTEROV, New variants of bundle methods, Mathematical programming, 69 (1995), pp. 111–147.
26. C. LEMARÉCHAL, F. OUSTRY, AND C. SAGASTIZÁBAL, The u-lagrangian of a convex function, Transactions of the American mathematical Society, 352 (2000), pp. 711–729.
27. J. LIANG AND R. D. MONTEIRO, A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes, SIAM Journal on Optimization, 31 (2021), pp. 2955–2986.
28. S. ŁOJASIEWICZ, Une propriété topologique des sous-ensembles analytiques réels, Les équations aux dérivées partielles, 117 (1963).
29. R. MIFFLIN, An algorithm for constrained optimization with semismooth functions, Mathematics of Operations Research, 2 (1977), pp. 191–207.
30. R. MIFFLIN AND C. SAGASTIZÁBAL, Functions with primal-dual gradient structure and u-hessians, Nonlinear Optimization and Related Topics, (2000), pp. 219–233.
31. ———, A-vu algorithm for convex minimization, Mathematical programming, 104 (2005), pp. 583–608.
32. ———, A science fiction story in nonsmooth optimization originating at iiasa, this volume, (2012).
33. A. NEMIROVSKI, Information-based complexity of convex programming, Lecture notes, 834 (1995).
34. A. S. NEMIROVSKY AND D. B. YUDIN, Problem complexity and method efficiency in optimization., John Wiley UK/USA, 1983.
35. Y. NESTEROV, Introductory lectures on convex optimization: A basic course, vol. 87, Springer Science & Business Media, 2003.
36. ———, Smooth minimization of non-smooth functions, Mathematical programming, 103 (2005), pp. 127–152.

37. Y. NESTEROV, Gradient methods for minimizing composite functions, Mathematical programming, 140 (2013), pp. 125–161.
38. J.-S. PANG, Error bounds in mathematical programming, Mathematical Programming, 79 (1997), pp. 299–332.
39. B. T. POLYAK, Gradient methods for the minimisation of functionals, USSR Computational Mathematics and Mathematical Physics, 3 (1963), pp. 864–878.
40. B. T. POLYAK, Minimization of unsmooth functionals, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.
41. J. RENEGAR AND B. GRIMMER, A simple nearly optimal restart scheme for speeding up first-order methods, foundations of computational mathematics, 22 (2022), pp. 211–256.
42. P. WOLFE, A method of conjugate subgradients for minimizing nondifferentiable functions, in Nondifferentiable optimization, Springer, 2009, pp. 145–173.
43. J. ZHANG, H. LIN, S. JEGELKA, S. SRA, AND A. JADBABAIE, Complexity of finding stationary points of nonconvex nonsmooth functions, in International Conference on Machine Learning, PMLR, 2020, pp. 11173–11182.
44. Z. ZHANG AND G. LAN, Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization, arXiv preprint arXiv:2210.05807, (2022).

7 Appendix

Lemma 7.1 *Let the function $f : \mathcal{X} \rightarrow \mathbb{R}$ be M -Lipschitz continuous. Suppose that for two points $\tilde{x}, \tilde{y} \in \mathcal{X}$, the smoothness condition $f(\tilde{x}) - l_f(\tilde{x}; \tilde{y}) \leq \frac{L}{2} \|\tilde{x} - \tilde{y}\|^2$ holds for some $L \geq 0$. Then for any points $x \in B(\tilde{x}; \bar{\delta})$ and $y \in B(\tilde{y}; \bar{\delta})$ in the $\bar{\delta}$ -neighborhoods of \tilde{x} and \tilde{y} respectively, we have:*

$$f(x) - l_f(x; \tilde{y}) \leq L\|x - y\|^2 + 2M\bar{\delta} + 4L\bar{\delta}^2.$$

Proof We decompose the left-hand side into three parts:

$$f(x) - l_f(x; \tilde{y}) = \underbrace{(f(x) - f(\tilde{x}))}_{\text{Term 1}} + \underbrace{(f(\tilde{x}) - l_f(\tilde{x}; \tilde{y}))}_{\text{Term 2}} + \underbrace{(l_f(\tilde{x}; \tilde{y}) - l_f(x; \tilde{y}))}_{\text{Term 3}}.$$

We bound each of these terms separately.

- *Term 1:* By the M -Lipschitz continuity of f and since $x \in B(\tilde{x}; \bar{\delta})$, we have:

$$f(x) - f(\tilde{x}) \leq M\|x - \tilde{x}\| \leq M\bar{\delta}.$$

- *Term 2:* Using the lemma’s assumption and standard norm inequalities, we bound this term:

$$\begin{aligned} f(\tilde{x}) - l_f(\tilde{x}; \tilde{y}) &\leq \frac{L}{2} \|\tilde{x} - \tilde{y}\|^2 \\ &\leq \frac{L}{2} (2\|x - y\|^2 + 2\|(\tilde{x} - x) + (y - \tilde{y})\|^2) \\ &\leq L\|x - y\|^2 + L(\|\tilde{x} - x\| + \|y - \tilde{y}\|)^2 \quad (\text{by triangle inequality}) \\ &\leq L\|x - y\|^2 + L(2\bar{\delta})^2 = L\|x - y\|^2 + 4L\bar{\delta}^2. \end{aligned}$$

- *Term 3:* By the definition of $l_f(\cdot; \tilde{y})$ and the fact that $\|f'(\tilde{y})\|_* \leq M$:

$$l_f(\tilde{x}; \tilde{y}) - l_f(x; \tilde{y}) = \langle f'(\tilde{y}), \tilde{x} - x \rangle \leq \|f'(\tilde{y})\|_* \|\tilde{x} - x\| \leq M\bar{\delta}.$$

Summing the bounds for the three terms gives the desired result. ■

Proof (Proof to Lemma 4.2)

Let $\hat{x} \leftarrow \arg \min_{x \in X} F_{2v}(x; \bar{x})$. It follows from Lemma 2.2 in [5] that $\nabla f_{2v}(\bar{x}) = 2v(\hat{x} - \bar{x})$. Since F_{2v} is strongly convex with modulus v , we have

$$\begin{aligned} \|\nabla f_{2v}(\bar{x})\|^2 &= (8v) \frac{v}{2} \|\bar{x} - \hat{x}\|^2 \\ &\leq 8v[F_{2v}(\bar{x}; \bar{x}) - F_{2v}(\hat{x}; \bar{x})] \\ &= 8v[f(\bar{x}) - f_{2v}(\bar{x})]. \end{aligned}$$
■

The following lemma provides a lower bound on the optimal objective value f^* and a computable upper bound on the QG modulus μ .

Lemma 7.2 *Let f be a convex objective function. The following statements are valid.*

a) *If f satisfies the QG condition with modulus $\mu > 0$, then for any point $x^0 \in X$, we have the lower bound:*

$$f(x^0) - \frac{2}{\mu} \|f'(x^0)\|^2 \leq f^*.$$

b) *For any two points $x, y \in X$ with $f(x) > f(y)$, an upper bound on a valid QG modulus μ for the point x is given by:*

$$\mu \leq \frac{2\|f'(x)\|^2}{f(x) - f(y)}.$$

c) *If f is $\bar{\mu}$ -strongly convex, then for any point $x^0 \in \mathcal{X}$, we have the tighter lower bound:*

$$f(x^0) - \frac{1}{2\bar{\mu}} \|f'(x^0)\|^2 \leq f^*.$$

Proof Part a): Let x_p^* be the projection of x^0 onto the optimal set X^* . By convexity and the QG property, we have:

$$\frac{\mu}{2} \|x^0 - x_p^*\|^2 \leq f(x^0) - f^* = f(x^0) - f(x_p^*) \leq \langle f'(x^0), x^0 - x_p^* \rangle \leq \|f'(x^0)\| \|x^0 - x_p^*\|.$$

Rearranging the first and last terms gives $\|x^0 - x_p^*\| \leq \frac{2\|f'(x^0)\|}{\mu}$. Substituting this back into the inequality $f(x^0) - f^* \leq \|f'(x^0)\| \|x^0 - x_p^*\|$ yields the result.

Part b): The result follows directly from the Polyak-Lojasiewicz (PL) type inequality derived in the Part a), $f(x) - f^* \leq \frac{2\|f'(x)\|^2}{\mu}$, combined with the fact that $f^* \leq f(y)$.

Part c): The definition of $\bar{\mu}$ -strong convexity provides a quadratic lower bound on f around x^0 . For any optimal solution x^* :

$$f^* = f(x^*) \geq f(x^0) + \langle f'(x^0), x^* - x^0 \rangle + \frac{\bar{\mu}}{2} \|x^* - x^0\|^2.$$

The right-hand side is a quadratic in $(x^* - x^0)$, which is minimized at $x^* - x^0 = -f'(x^0)/\bar{\mu}$. The minimum value is $f(x^0) - \frac{1}{2\bar{\mu}} \|f'(x^0)\|^2$, which gives the desired lower bound on f^* .