INCREMENTAL METHODS FOR WEAKLY CONVEX OPTIMIZATION

XIAO LI*, ZHIHUI ZHU†, ANTHONY MAN-CHO SO‡, AND JASON D LEE§

Abstract. Incremental methods are widely utilized for solving finite-sum optimization problems in machine learning and signal processing. In this paper, we study a family of incremental methods—including incremental subgradient, incremental proximal point, and incremental prox-linear methods—for solving weakly convex optimization problems. Such a problem class covers many nonsmooth nonconvex instances that arise in engineering fields. We show that the three said incremental methods have an iteration complexity of $O\left(\varepsilon^{-\frac{4}{2}}\right)$ for driving a natural stationarity measure to below $\varepsilon$. Moreover, we show that if the weakly convex function satisfies a sharpness condition, then all three incremental methods, when properly initialized and equipped with geometrically diminishing stepsizes, can achieve a local linear rate of convergence. Our work is the first to extend the convergence rate analysis of incremental methods from the nonsmooth convex regime to the weakly convex regime. Lastly, we conduct numerical experiments on the robust matrix sensing problem to illustrate the convergence performance of the three incremental methods.

Key words. nonsmooth nonconvex optimization, incremental subgradient method, incremental proximal-type methods, iteration complexity, sharpness, linear convergence.

AMS subject classifications. 68Q25, 65K10, 90C90, 90C26, 90C06.

1. Introduction. Throughout this paper, we focus on the finite-sum optimization problem

\begin{equation}
\min_{x \in \mathbb{R}^n} \ f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)
\end{equation}

subject to \( x \in C \),

where each component function \( f_i : \mathbb{R}^n \rightarrow \mathbb{R} \) is assumed to be weakly convex and \( C \subseteq \mathbb{R}^n \) is a nonempty closed convex set. Recall that the function \( f \) is said to be \( \tau \)-weakly convex if \( f(\cdot) + \frac{\tau}{2} \| \cdot \|^2 \) is convex for some constant \( \tau \geq 0 \) [55]. In modern data processing applications, the number of components \( m \) can be very large. Thus, it can be computationally expensive to utilize the full information (such as the subgradient) of \( f \) in each update. This observation is precisely one of the main motivations of incremental methods, which update the iterates using a single component function \( f_i \) rather than all the components of \( f \). It has been demonstrated by many experiments that depending on the value of \( m \), an incremental method can significantly outperform its nonincremental counterpart [3].

Incremental methods are practical algorithms, which have a long history and are widely utilized in engineering fields. Most notably, incremental subgradient method and its randomly permuted version (also known as “random shuffling”) are broadly employed in practice for training deep neural networks with a nonsmooth activation function; see e.g., [5, 7, 8, 22, 25, 50]. Despite the broad applications of incremental
methods to nonsmooth optimization problems, their theoretical properties are less explored. The main prior results on convergence rates of incremental methods when used to solve nonsmooth optimization problems are based on convexity assumptions, as stated in [2, Section 4.1.3]. To the authors’ knowledge, there is no convergence rate result for incremental methods even when the objective function in (1.1) is weakly convex (which can be nonsmooth and nonconvex).

Our interest in applying incremental methods to solve problem (1.1) stems from the fact that this problem appears widely in engineering fields such as signal processing and machine learning and almost no incremental methods have been proposed to solve it. As a demonstration, we list below two motivating applications that are instances of problem (1.1). The weak convexity of the objective functions in these applications follows from the fact that they all have the form

\begin{equation}
    f(x) = h(c(x)) = \frac{1}{m} \sum_{i=1}^{m} h_i(c_i(x)) = \frac{1}{m} \sum_{i=1}^{m} \left| y_i - \langle a_i, x \rangle \right|_2,
\end{equation}

where \(h_i : \mathbb{R}^d \to \mathbb{R}\) is a Lipschitz continuous convex function and \(c_i : \mathbb{R}^n \to \mathbb{R}^d\) is a smooth mapping with Lipschitz continuous Jacobian (see, e.g., [17, Lemma 4.2]).

1.1. Motivating applications.

**Application 1: Robust Matrix Sensing [33].** Low-rank matrices are ubiquitous in computer vision, signal processing, machine learning, and data science applications. One fundamental computational task is to recover a positive semidefinite (PSD) matrix \(X^* \in \mathbb{R}^{n \times n}\) with rank\((X^*) = r \ll n\) from a number of corrupted linear measurements

\begin{equation}
    y = A(X^*) + s^*,
\end{equation}

where \(A : \mathbb{R}^{n \times n} \to \mathbb{R}^{m}\) is a linear measurement operator consisting of a set of sensing matrices \(A_1, \ldots, A_m \in \mathbb{R}^{n \times n}\) and \(s^* \in \mathbb{R}^{m}\) is a sparse vector with arbitrary nonzero entries (i.e., outliers). The work [33] proposes to recover the low-rank matrix \(X^*\) by using a factored representation of the matrix variable [9] (i.e., \(X = UU^T\) with \(U \in \mathbb{R}^{n \times r}\)) and employing a \(\ell_1\)-loss function to robustify the solution against outliers. This leads to the optimization formulation

\begin{equation}
\min_{U \in \mathbb{R}^{n \times r}} \frac{1}{m} \|y - A(UU^T)\|_1 = \frac{1}{m} \sum_{i=1}^{m} |y_i - \langle a_i, UU^T \rangle|.
\end{equation}

**Application 2: (Real-Valued) Robust Phase Retrieval [18].** An important problem arising in physics, imaging science, and signal processing is phase retrieval, which aims to recover a signal \(x^* \in \mathbb{C}^n\) from its amplitude measurements. For the purpose of illustration, let us consider a real-valued version of the problem. Suppose that the measurements are given by

\begin{equation}
    y = |Ax^*|^2 + s^*,
\end{equation}

where the operator \(|\cdot|^2\) in (1.5) is applied component-wise to its arguments. Here, \(A \in \mathbb{R}^{m \times n}\) is the measurement matrix and \(s^* \in \mathbb{R}^{m}\) is a sparse vector with arbitrary nonzero entries (i.e., outliers). The work [18] considers the following formulation for recovering both the sign and magnitude of \(x^*\):

\begin{equation}
\min_{x \in \mathbb{R}^n} \frac{1}{m} \|y - |Ax|^2\|_1 = \frac{1}{m} \sum_{i=1}^{m} |y_i - |a_i, x||^2 .
\end{equation}
1.2. Main contributions. In this paper, we study a family of incremental methods—including incremental subgradient, incremental proximal point, and incremental prox-linear methods—for solving problem (1.1). Among them, the incremental prox-linear method is new to our knowledge. We develop a unified framework for analyzing the convergence rates of these methods. In particular, we show that the three incremental methods mentioned above drive a surrogate stationarity measure to zero at a rate of $O(k^{-1/4})$, where $k$ is the iteration index (see Theorem 1). In addition, we show that if problem (1.1) possesses the so-called sharpness property (see Definition 1), then the three incremental methods with properly designed geometrically diminishing stepsizes and a good initialization will converge to the set of weak sharp minima at a linear rate (see Theorem 2). Our work is the first to extend the convergence rate analysis of incremental methods from the nonsmooth convex regime to the weakly convex regime, which covers a large class of nonsmooth nonconvex problems.

Unlike the incremental aggregated gradient methods (see e.g., [6, 23, 36]), which has a linear rate of convergence for smooth strongly convex optimization problems, we do not incorporate any aggregation techniques in our incremental methods. Moreover, it should be noted that the full subgradient method also requires geometrically diminishing stepsizes to guarantee linear convergence, and it is usually outperformed by its incremental counterpart numerically (see Section 5 for an illustration). This further demonstrates the promise of the incremental subgradient method.

Our linear convergence result generalizes the original ones in [3, 38, 39], which concern the incremental subgradient and proximal point methods for nonsmooth convex optimization problems. To obtain the global sublinear convergence result, we adopt the surrogate stationarity measure for weakly convex minimization problems from [13,17].

1.3. Related works.

Incremental gradient method. A popular algorithm for solving problem (1.1) when the components are smooth is the incremental gradient method. Such a method has a long tradition and is extensively studied. The starting work dates back to [58] for solving linear least-squares problems. Then, various works [21, 34, 35, 52, 54] study the convergence of incremental gradient method with different stepsize schemes. Most of them establish asymptotic convergence of the method without providing an explicit convergence rate. For instance, Solodov [52] showed that when applied to a smooth nonconvex optimization problem, every limit point of the sequence of iterates generated by the incremental gradient method with a constant stepsize bounded away from zero is an approximate stationary point. A more recent work [24] shows that if the objective function $f$ is strongly convex and twice continuously differentiable and the iterates are uniformly bounded, then the incremental gradient method with stepsizes diminishing at the rate of $O(1/k)$ will drive the distances between the iterates and the optimal solution to zero at an asymptotic rate of $O(1/k)$. The analysis in [24] was later adopted by several works [22, 25, 37] for showing the convergence of random shuffling (a randomly permuted version of incremental gradient method) when applied to smooth strongly convex optimization problems. A popular variant named incremental aggregated gradient method [6, 23, 36]—which, in each update, evaluates the gradient of a single component function while keeping a memory of the most recent gradients of all the other components to approximate the full gradient—is shown to converge linearly for smooth strongly convex minimization using a constant stepsize.
**Incremental subgradient method.** The incremental subgradient method is widely used to tackle problem (1.1) when the components are nonsmooth. In the nonsmooth setting, almost all the existing convergence rate results require convexity of the components; see the comments in [2, Section 4.1.3]. As reviewed in [2], the incremental subgradient method was first proposed in [29]. After that, Nedić and Bertsekas [38,39] provided convergence results for the incremental subgradient method using several stepsize rules when the components are convex. In particular, they proved that the algorithm converges at a rate of $O(1/\sqrt{k})$ in terms of the function suboptimality gap $f(x_k) - f^*$ when a constant stepsize or Polyak’s dynamic stepsizes is used. Furthermore, under an additional sharpness property (see Definition 1), they proved that the algorithm with Polyak’s dynamic stepsizes will drive the distances between the iterates and the optimal solution set to zero at a linear rate. Later, the works [30, 40, 45] establish convergence results for the incremental $\varepsilon$-subgradient method, in which the exact subgradient oracle is not available. Among these, the work [40] also considers the effect of deterministic noise in the update. There are also many other works studying the incremental subgradient method for the setting where the components are convex in the context of distributed optimization, sensor network optimization, etc; see, e.g., [27, 28,41, 42,46–48, 56]. By contrast, our global iteration complexity and linear convergence results apply to weakly convex minimization, in which the objective function can be nonsmooth and nonconvex. In addition, our linear convergence result builds on a more practical geometrically diminishing stepsize rule (due to Shor) than the Polyak’s stepsize rule, which requires the knowledge of the optimal function value $f^*$.

There are also works considering the incremental subgradient method for nonsmooth nonconvex minimization. In particular, Solodov [53] proposed perturbed subgradient-type methods that include the incremental subgradient method as a special case and showed that the iterates generated by these methods with diminishing stepsizes will converge to the set of first-order stationary points of a Lipschitz continuous and regular function (in the sense of Clarke). Despite the generality of this work, the author did not report any convergence rate result even in the case where the objective function is restricted to be convex. It is also worth mentioning that if the problem is convex and possesses the sharpness property, then the work [53] shows that the iterates generated by the perturbed incremental subgradient method will converge to the set of exact global minimizers as long as the magnitude of the perturbation is smaller than the sharpness parameter. A more recent work [26] studies the case where the components are quasi-convex. Under the stringent assumption that all the components have a common optimal solution, it is shown that the incremental subgradient method asymptotically converges to the optimal solution set, but there is no rate guarantee. By contrast, we establish an explicit rate of convergence of the incremental subgradient method when applied to weakly convex minimization problems.

**Incremental proximal point method.** Bertsekas [3] proposed the incremental proximal point method for solving problem (1.1) with convex components. The method includes the incremental gradient, subgradient, proximal gradient, and proximal subgradient methods as special cases. By utilizing proof techniques similar to those in [38,39], the author proved an $O(1/\sqrt{k})$ convergence rate in terms of the function suboptimality gap when a constant stepsize is used. Such a convergence rate coincides with that of the incremental subgradient method [38,39]. The work [57] considers a slightly more general case where the constraint is an intersection of a large
number of simple convex constraints and proposes an incremental projection-proximal method for solving it. It is shown in [57] that the method has an $O(1/\sqrt{k})$ convergence rate. To our knowledge, the convergence of the incremental proximal point method has only been studied in the convex setting. By contrast, we consider the incremental proximal point method for weakly convex minimization, where the problem can be nonsmooth and nonconvex. In addition, we elucidate the role of sharpness in the linear convergence analysis of the incremental proximal point method, which is not addressed in [3, 57].

**Stochastic methods.** In the past decade, stochastic methods have been extensively studied in optimization and machine learning. There is a substantial literature on stochastic variants of subgradient and proximal-type methods. Actually, the works [3, 38, 39] mentioned earlier also discussed stochastic variants of incremental subgradient and proximal point methods, which randomly select a component function from $\{f_1, \ldots, f_m\}$ in each update rather than sweeping all the components in a cyclic order. Recently, several works [13, 14, 16, 19] have proposed different stochastic methods for solving weakly convex minimization problems. In particular, it is shown in [13] that a family of stochastic methods will drive a certain surrogate stationarity measure to zero at a rate of $O(k^{-1/4})$. Central to the analysis in [13] is the observation that the Moreau envelope of a weakly convex function can be regarded as an approximate Lyapunov function and the algorithm dynamics can drive the gradient of the Moreau envelope to zero. In this paper, we adopt such a surrogate stationarity measure for analyzing the global iteration complexity of our incremental methods. In addition, the work [14], which appeared on arXiv one week before our preliminary technical report [32], establishes the linear convergence of several stochastic methods when applied to sharp weakly convex minimization problems. However, the stochastic algorithms proposed in [14] incorporate a restarting strategy, which results in a computationally heavy inner loop. When the problem is large-scale and/or high-dimensional, those algorithms can be very inefficient. By contrast, when solving sharp weakly convex optimization problems, our linearly convergent incremental methods do not need any restarting scheme, which is much more efficient in practice. It is worth pointing out that the necessity of investigating incremental methods when there are already results for their stochastic counterparts can be seen from three aspects: 1) The analysis of stochastic methods relies heavily on the stochastic nature of the component function selection rules and does not carry over to incremental methods. 2) There are applications—such as source localization problems in sensor networks and distributed empirical risk minimization [23]—where random sampling required by the stochastic methods may not be possible since the access to the components are predetermined in a deterministic order by the problem’s physical nature. 3) The analysis of cyclic updating order in incremental methods can be crucial for understanding more general component function selection rules such as the above mentioned random shuffling algorithm; see, e.g., [22].

For the sake of clarity, we list several representative results on incremental methods and compare them with our results in Table 1.

### 2. Preliminaries

In this section, we review several elements of nonsmooth analysis and specialize them to weakly convex functions, including the subdifferential, subdifferential calculus for the composite form (1.2), and a subgradient inequality. We then present a family of incremental methods for solving problem (1.1).
Table 1: Summary of prior arts. The algorithm used in [38] is the incremental subgradient method, while that used in [3] is the incremental proximal point method. The results in this paper cover the incremental subgradient, proximal point, and prox-linear methods. Here, \( k \) denotes the iteration index and \( T \) denotes the total number of iterations. In Polyak’s rule, \( f^* \) denotes the optimal function value and \( \nabla f(x_k) \in \partial f(x_k) \) denotes a nonzero subgradient. Furthermore, \( \Theta(x_k) \) denotes a surrogate stationarity measure of a weakly convex function \( f \) (see (3.3)), \( \text{dist}(x, Z) := \inf_{z \in Z} \|x - z\|_2 \) denotes the Euclidean distance between the point \( x \) and the non-empty closed set \( Z \), and \( X \) denotes the set of weak sharp minima of problem (1.1) (see Definition 1). We hide numerical constants in the big-oh notation for a cleaner display.

| Paper | Assumptions | Stepsize | Complexity | Stationarity Measure |
|-------|-------------|---------|------------|---------------------|
| [38]  | \( f_i \) convex \( f_i \) Lipschitz | Constant | \( T = O\left( \frac{1}{\epsilon} \right) \) | \( \min_{0 \leq k \leq T} f(x_k) - f^* \leq \epsilon \) |
| [38]  | \( f_i \) convex \( f_i \) Lipschitz \( f \) sharp | Polyak’s rule | \( T = O\left( \log \left( \frac{1}{\epsilon} \right) \right) \) | \( \text{dist}(x_T, X) \leq \epsilon \) |
| [3]   | \( f_i \) convex \( f_i \) Lipschitz \( f \) sharp | Constant | \( T = O\left( \frac{1}{\epsilon^2} \right) \) | \( \min_{0 \leq k \leq T} f(x_k) - f^* \leq \epsilon \) |
| This paper \ Theorem 1 | \( f_i \) weakly convex \( f_i \) Lipschitz \( f \) sharp \ Good initialization | Constant & diminishing \( \Theta(x_k) \leq \epsilon \) | \( T = O\left( \frac{1}{\epsilon} \right) \) | \( \text{dist}(x_T, X) \leq \epsilon \) |
| This paper \ Theorem 2 | \( f_i \) weakly convex \( f_i \) Lipschitz \( f \) sharp \ Good initialization | Geometrically diminishing \( \rho^k \cdot \rho_0 \) | \( T = O\left( \log \left( \frac{1}{\epsilon} \right) \right) \) | \( \text{dist}(x_T, X) \leq \epsilon \) |

2.1. Subdifferential, first-order optimality condition, and a subgradient inequality. Let \( f : \mathbb{R}^n \to \mathbb{R} \) be a \( \tau \)-weakly convex function for some parameter \( \tau \geq 0 \). By [55, Proposition 4.3], there is a convex function \( \phi : \mathbb{R}^n \to \mathbb{R} \) such that \( f(x) = \phi(x) - \frac{\tau}{2} \|x\|_2^2 \) for any \( x \in \mathbb{R}^n \). According to [55, Proposition 4.6], we have

\[
\partial f(x) = \partial \phi(x) - \tau x, \tag{2.1}
\]

where \( \partial \phi(x) \) is the usual convex subdifferential of \( \phi \) at \( x \). Thus, the above subdifferential of the weakly convex function \( f \) is well defined.

It is not always immediate how to explicitly calculate the subdifferential of \( f \) from (2.1), as it may not be easy to obtain the convex function \( \phi \) associated with \( f \). Nevertheless, when \( f \) takes the composite form (1.2), we have

\[
\partial f(x) = \nabla c(x)^\top \partial h(c(x))
\]

by [49, Theorem 10.6] and [55, Proposition 4.5]. An element \( \nabla f(x) \in \partial f(x) \) is called a subgradient of \( f \) at \( x \). Using the definition of the subdifferential of \( f \) in (2.1), we have the following equivalent characterization of the \( \tau \)-weak convexity of \( f \): For all...
\[ x, y \in \mathbb{R}^n, \]

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{\tau}{2} \| y - x \|^2, \quad \forall \nabla f(x) \in \partial f(x); \]

see, e.g., [55, Proposition 4.8].

We call \( x \in C \) a stationary point of problem (1.1) if it satisfies

\[ 0 \in \partial f(x) + N_C(x), \]

where \( N_C(x) = \{ s \in \mathbb{R}^n : \langle s, y - x \rangle \leq 0, \forall y \in C \} \) is the normal cone to the convex set \( C \) at \( x \).

2.2. The incremental methods. Let us now introduce a family of incremental methods—namely the incremental subgradient, incremental proximal point, and incremental prox-linear methods. At each iteration, the incremental methods update \( x_k \) to \( x_{k+1} \) through \( m \) sequential steps by utilizing the components \{\( f_1, \ldots, f_m \)\} sequentially. In each step, only one component function \( f_i \in \{ f_1, \ldots, f_m \} \) is selected for updating. To be more specific, at the \((k+1)\)-st iteration, incremental methods start with \( x_{k,0} = x_k \) and then update \( x_{k,i} \) using \( f_i \) for \( i = 1, \ldots, m \), giving \( x_{k+1} = x_{k,m} \).

Let \( \mu_k \) be the stepsize at the \((k+1)\)-st iteration. The three incremental methods are presented below. They differ from each other in the update of \( x_{k,i} \).

A) Incremental subgradient method

\[ x_{k,i} = P_C(x_{k,i-1} - \mu_k \nabla f_i(x_{k,i-1})) \quad \text{with} \quad \nabla f_i(x_{k,i-1}) \in \partial f_i(x_{k,i-1}), \]

for \( i = 1, \ldots, m \).

B) Incremental proximal point method

\[ x_{k,i} = \arg\min_{x \in C} \left\{ f_i(x) + \frac{1}{2\mu_k} \| x - x_{k,i-1} \|^2 \right\}, \]

for \( i = 1, \ldots, m \).

C) Incremental prox-linear method

When the objective function \( f \) in problem (1.1) has the composite form (1.2), one can ‘inner-linearize’ \( c_i \) at \( x_{k,i-1} \) as

\[ f_i(x; x_{k,i-1}) = h_i \left( c_i(x_{k,i-1}) + \nabla c_i(x_{k,i-1})^T (x - x_{k,i-1}) \right). \]

Then, the incremental prox-linear method updates \( x_{k,i} \) as

\[ x_{k,i} = \arg\min_{x \in C} \left\{ f_i(x; x_{k,i-1}) + \frac{1}{2\mu_k} \| x - x_{k,i-1} \|^2 \right\}, \]

for \( i = 1, \ldots, m \). It is worth noting that subproblem (2.7) admits a closed-form solution when \( h_i \) is the absolute value function and \( C \equiv \mathbb{R}^n \), which is the case for all the concrete examples listed in Subsection 1.1. Indeed, in this case, subproblem (2.7) takes the form

\[ x_{k,i} = \arg\min_{x \in \mathbb{R}^n} \left\{ |\langle a, x \rangle + b| + \frac{1}{2} \| x - x_{k,i-1} \|^2 \right\} \]
for some \(a \in \mathbb{R}^n, b \in \mathbb{R}\), whose solution is given by

\[
x_{k,i} = x_{k,i-1} - P_{[-1,1]}(\sigma)a \quad \text{with} \quad \sigma = \frac{\langle a, x_{k,i-1} \rangle + b}{\|a\|_2^2}.
\]

Here, \(P_{[-1,1]}\) represents the projector onto the line interval \([-1,1]\); see, e.g., [19].

3. Global Convergence. In this section, we study the iteration complexity of the incremental methods for solving problem (1.1).

3.1. Assumptions and surrogate stationarity measure. The incremental methods presented in the last section exploit different types of problem structure. Thus, we need different assumptions when studying the convergence behaviors of these methods. To start, we state the assumptions needed in our analysis of the incremental subgradient and proximal point methods.

Assumption 1 (incremental subgradient and proximal point methods).

- (bounded subgradients) For \(i \in \{1, \ldots, m\}\), there exist an open convex set \(Q\) that contains \(C\) and a constant \(L_1 > 0\) such that \(\|\nabla f_i(x)\| \leq L_1\) for all \(x \in Q\) and \(\nabla f_i(x) \in \partial f_i(x)\).

- (weak convexity) The component functions in (1.1) are weakly convex with parameter \(\tau_1 \geq 0\).

Note that the bounded subgradients assumption is standard in the analysis of incremental, stochastic, and subgradient-based algorithms; see, e.g., [3, 13, 38, 39, 43]. In many important applications, the set \(C\) is compact. In this case, the bounded subgradients assumption is automatically satisfied due to (2.1) and the fact that \(\partial \phi(x)\) is compact (see, e.g., [1, Proposition B.24]). Even if \(C\) is not compact, it may be possible to intersect it with a ball that is large enough to contain an optimal solution to problem (1.1), so that the bounded subgradients assumption is satisfied.

Though the component functions in (1.2) are automatically weakly convex, the incremental prox-linear method (2.7) explicitly exploits their composite structure. Therefore, in addition to weak convexity, we need to make a slightly stronger assumption, namely quadratic approximation, when analyzing this algorithm.

Assumption 2 (incremental prox-linear method).

- (bounded subgradients) For \(i \in \{1, \ldots, m\}\), there exist an open convex set \(Q\) that contains \(C\) and a constant \(L_2 > 0\) such that \(\|\nabla f_i(x; \bar{x})\| \leq L_2\) for all \(x, \bar{x} \in Q\) and \(\nabla f_i(x; \bar{x}) \in \partial f_i(x; \bar{x})\).

- (quadratic approximation) There exists a constant \(\tau_2 > 0\) such that each component function \(f_i\) satisfies

\[
f_i(x; \bar{x}) - f_i(x) \leq \frac{\tau_2}{2} \|x - \bar{x}\|^2, \quad \forall x, \bar{x} \in \mathbb{R}^n.
\]

Here, \(f_i(x; \bar{x})\) is defined in (2.6).

The quadratic approximation assumption has been used in [18] and [13, 19] to analyze the full and stochastic prox-linear methods, respectively. It is straightforward to verify that this assumption immediately implies the \(\tau_2\)-weak convexity of \(f_i\) by applying the convex subgradient inequality to \(h_i\). To unify the notation in Assumption 1 and Assumption 2, we define

\[
L = \max\{L_1, L_2\} \quad \text{and} \quad \tau = \max\{\tau_1, \tau_2\}.
\]
One of the main challenges in analyzing the convergence rate of an algorithm for nonsmooth nonconvex optimization is to find an appropriate stationarity measure to track the progress of the algorithm. The recent papers [13,17] show that the Moreau envelope of a weakly convex function can be regarded as an approximate Lyapunov function and defines a surrogate stationarity measure for the weakly convex function at hand. Once this Lyapunov function is identified, the convergence analysis of subgradient-type methods for weakly convex minimization (see, e.g., [13]) essentially follows that for convex optimization. In this paper, we adopt the said surrogate stationarity measure for analyzing the global iteration complexity of our incremental methods. For completeness, we briefly introduce the related notions of Moreau envelope and proximal mapping; see [49, Definition 1.22] for details.

For any $\lambda > 0$, the Moreau envelope of $f$ is defined as

$$
\tilde{f}_\lambda(x) := \min_{y \in C} \left\{ f(y) + \frac{1}{2\lambda} \| y - x \|^2 \right\}, \quad x \in C.
$$

The corresponding proximal mapping is defined as

$$
P_{\lambda,f}(x) := \arg\min_{y \in C} \left\{ f(y) + \frac{1}{2\lambda} \| y - x \|^2 \right\}, \quad x \in C.
$$

By the subdifferential calculus of sum of (regular) functions [49, Corollary 10.9], the first-order optimality condition of $P_{\lambda,f}(x)$ in (3.2) implies that

$$
0 \in \partial f(P_{\lambda,f}(x)) + N_{C}(P_{\lambda,f}(x)) + \frac{1}{\lambda}(P_{\lambda,f}(x) - x).
$$

It follows that

$$
\text{dist}(0, \partial f(P_{\lambda,f}(x)) + N_{C}(P_{\lambda,f}(x))) \leq \frac{1}{\lambda} \| x - P_{\lambda,f}(x) \|^2 =: \Theta(x).
$$

One can see from (2.3) and (3.3) that if $\Theta(x) = 0$, then $x \in C$ is a stationary point of problem (1.1). Thus, we use $x \mapsto \Theta(x)$ as the surrogate stationarity measure of problem (1.1) and call $x \in C$ an $\varepsilon$-nearly stationary point if $\Theta(x) \leq \varepsilon$.

3.2. Several useful lemmas. Before stating our main convergence results, let us present several useful lemmas that will be used in our later development.

First, we show that (2.5) and (2.7) can be interpreted as certain subgradient-type update.

**Lemma 1** (subgradient-type update). The following hold for all $k \geq 0$ and $1 \leq i \leq m$:

1. If $x_{k,i}$ is generated by the incremental proximal point method (2.5), then there exists a $\tilde{\nabla} f_i(x_{k,i}) \in \partial f_i(x_{k,i})$ such that

$$
x_{k,i} = P_C(x_{k,i-1} - \mu_k \tilde{\nabla} f_i(x_{k,i})).
$$

2. If $x_{k,i}$ is generated by the incremental prox-linear method (2.7), then there exists $\tilde{\nabla} f_i(x_{k,i}; x_{k,i-1}) \in \partial f_i(x_{k,i}; x_{k,i-1})$ such that

$$
x_{k,i} = P_C(x_{k,i-1} - \mu_k \tilde{\nabla} f_i(x_{k,i}; x_{k,i-1})).
$$
Proof. We provide the proof for (3.4). The proof of (3.5) follows a similar argument. According to the first-order optimality of $x_{k,i}$ in (2.5) and [49, Corollary 10.9], we have
\[
\frac{1}{\mu_k} (x_{k,i-1} - x_{k,i}) \in \partial f_i(x_{k,i}) + N_C(x_{k,i}),
\]
which implies that
\[
(x_{k,i-1} - \mu_k \nabla f_i(x_{k,i})) - x_{k,i} \in N_C(x_{k,i})
\]
for some $\nabla f_i(x_{k,i}) \in \partial f_i(x_{k,i})$. The above inclusion is equivalent to (3.4) by convexity of $C$.

The relations (3.4) and (3.5) are crucial to the analysis of proximal-type algorithms [3, 4]. It is interesting to see that the updates (3.4) and (3.5) are very similar to that of the incremental subgradient method (2.4). The only difference is that the subgradients in (3.4) and (3.5) are evaluated at $x_{k,i}$, while that in (2.4) is evaluated at $x_{k,i-1}$. This observation suggests that we can follow a unified proof strategy for all three incremental methods.

Next, we develop a preliminary recursion for the term $\|x_{k,i} - y\|^2$, where $x_{k,i}$ is the iterate generated by one of the incremental methods and $y \in C$ is arbitrary.

Lemma 2 (preliminary recursion). For any $y \in C$, $k \geq 0$, and $1 \leq i \leq m$, we have
\[
\|x_{k,i} - y\|^2 \leq \|x_{k,i-1} - y\|^2 - 2\mu_k (f_i(z_i) - f_i(y)) + \tau \mu_k \|z_i - y\|^2 + \gamma \|x_{k,i} - x_{k,i-1}\|^2,
\]
where
(a) $z_i = x_{k,i-1}$, $\gamma = 1$, $\theta = 2$

if $x_{k,i}$ is generated by the incremental subgradient method (2.4);
(b) $z_i = x_{k,i}$, $\gamma = 0$

if $x_{k,i}$ is generated by the incremental proximal point method (2.5);
(c) $z_i = x_{k,i-1}$, $\gamma = 2\mu_k L$, $\theta = 1$

if $x_{k,i}$ is generated by the incremental prox-linear method (2.7).

Proof. We decompose the error $\|x_{k,i-1} - y\|^2 = \|x_{k,i-1} - x_{k,i} + x_{k,i} - y\|^2$. Expanding the right-hand side, we have
\[
\|x_{k,i} - y\|^2 = \|x_{k,i-1} - y\|^2 - 2 (x_{k,i-1} - x_{k,i}, x_{k,i} - y) - \|x_{k,i} - x_{k,i-1}\|^2.
\]

We first show (a). By the first-order optimality of $x_{k,i}$ in (2.4), there exists an $s \in N_C(x_{k,i})$ such that $x_{k,i-1} - x_{k,i} = \mu_k (\nabla f_i(x_{k,i-1}) + s)$. Plugging this relation...
This completes the proof of Lemma 2.

Assumption 2 and [49, Theorem 9.13]. Substituting (3.10) into (3.9) yields

where the second line utilizes the quadratic approximation assumption in Assumption 1 and applying the weakly convex inequality (2.2) to (3.8), we have

\[
\|x_{k,i} - y\|^2 = \|x_{k,i-1} - y\|^2 - 2\mu_k \langle \nabla f_i(x_{k,i-1}) + s, x_{k,i} - y \rangle - \|x_{k,i} - x_{k,i-1}\|^2
\]

\[
\leq \|x_{k,i-1} - y\|^2 - 2\mu_k \langle \nabla f_i(x_{k,i-1}), x_{k,i-1} - y \rangle - 2\mu_k \langle \nabla f_i(x_{k,i-1}) + s, x_{k,i} - x_{k,i-1} \rangle - \|x_{k,i} - x_{k,i-1}\|^2,
\]

\[
= \|x_{k,i-1} - y\|^2 - 2\mu_k \langle \nabla f_i(x_{k,i-1}), x_{k,i-1} - y \rangle + \|x_{k,i} - x_{k,i-1}\|^2
\]

where we have used the fact that \(\langle s, x_{k,i} - y \rangle \leq 0\) in the inequality. By using the weak convexity assumption in Assumption 1 and applying the weak convex inequality (2.2) to (3.8), we have

\[
\|x_{k,i} - y\|^2 \leq \|x_{k,i-1} - y\|^2 - 2\mu_k (f_i(x_{k,i-1}) - f_i(y)) + \tau \mu_k \|x_{k,i} - y\|^2.
\]

We now show (b). By the first-order optimality of \(x_{k,i}\) in (2.5) and [49, Corollary 10.9], there exist \(\nabla f_i(x_{k,i}) \in \partial f_i(x_{k,i})\) and \(s \in N_C(x_{k,i})\) such that \(x_{k,i-1} - x_{k,i} = \mu_k (\nabla f_i(x_{k,i}) + s)\). Invoking this relation in (3.7) gives

\[
\|x_{k,i} - y\|^2 = \|x_{k,i-1} - y\|^2 - 2\mu_k \langle \nabla f_i(x_{k,i}), x_{k,i} - y \rangle - \|x_{k,i} - x_{k,i-1}\|^2
\]

\[
\leq \|x_{k,i-1} - y\|^2 - 2\mu_k \langle \nabla f_i(x_{k,i}), x_{k,i} - y \rangle - \|x_{k,i} - x_{k,i-1}\|^2
\]

\[
\leq \|x_{k,i-1} - y\|^2 - 2\mu_k (f_i(x_{k,i}) - f_i(y)) + \tau \mu_k \|x_{k,i} - y\|^2,
\]

where the last inequality is due to the weak convexity assumption in Assumption 1 and (2.2).

Finally, we show (c). The first-order optimality of \(x_{k,i}\) in (2.7) and [49, Corollary 10.9] ensure the existence of \(\nabla f_i(x_{k,i}, x_{k,i-1}) \in \partial f_i(x_{k,i}; x_{k,i-1})\) and \(s \in N_C(x_{k,i})\) such that \(x_{k,i-1} - x_{k,i} = \mu_k (\nabla f_i(x_{k,i}; x_{k,i-1}) + s)\). This, together with (3.7), gives

\[
\|x_{k,i} - y\|^2 \leq \|x_{k,i-1} - y\|^2 - 2\mu_k \langle \nabla f_i(x_{k,i}; x_{k,i-1}), x_{k,i} - y \rangle - \|x_{k,i} - x_{k,i-1}\|^2
\]

Recalling that \(x \mapsto f_i(x; x_{k,i-1})\) is convex, we have

\[
- \langle \nabla f_i(x_{k,i}; x_{k,i-1}), x_{k,i} - y \rangle \leq f_i(y; x_{k,i-1}) - f_i(x_{k,i}; x_{k,i-1})
\]

\[
\leq f_i(y) + \frac{\tau}{2} \|x_{k,i-1} - y\|^2 - f_i(x_{k,i-1}) + f_i(x_{k,i-1}) - f_i(x_{k,i}; x_{k,i-1})
\]

\[
\leq - (f_i(x_{k,i-1}) - f_i(y)) + \frac{\tau}{2} \|x_{k,i-1} - y\|^2 + L \|x_{k,i} - x_{k,i-1}\|^2,
\]

where the second line utilizes the quadratic approximation assumption in Assumption 2 and the last inequality follows from the bounded subgradients assumption in Assumption 2 and [49, Theorem 9.13]. Substituting (3.10) into (3.9) yields

\[
\|x_{k,i} - y\|^2 \leq \|x_{k,i-1} - y\|^2 - 2\mu_k (f_i(x_{k,i-1}) - f_i(y)) + \tau \mu_k \|x_{k,i-1} - y\|^2
\]

\[
+ 2\mu_k L \|x_{k,i} - x_{k,i-1}\|^2.
\]

This completes the proof of Lemma 2.
Lemma 3 (inner step length). For any $k \geq 0$, let $\{x_{k,i}\}_{i=1}^{m}$ be the sequence generated by any of the three incremental methods. Then, we have

$$
\|x_{k,i} - x_{k,j}\|_2 \leq |i - j| \mu_k L, \quad \forall i, j \in \{1, \ldots, m\}.
$$

Proof. Without loss of generality, we assume that $i \geq j$. According to Lemma 1, the updates of the three incremental methods can be written in a unified manner as $x_{k,i} = P_C(x_{k,i-1} - \mu_k S_{k,i})$, where $S_{k,i} = \nabla f_i(x_{k,i-1}) \in \partial f_i(x_{k,i-1})$ for the incremental subgradient method, $S_{k,i} = \nabla f_i(x_{k,i}) \in \partial f_i(x_{k,i})$ for the incremental proximal point method, and $S_{k,i} = \nabla f_i(x_{k,i}; x_{k,i-1}) \in \partial f_i(x_{k,i}; x_{k,i-1})$ for the incremental prox-linear method. Now, we prove (3.11) by induction. It is trivial to verify that (3.11) is true when $i = j$. Assuming that (3.11) holds for $i = l$, we have

$$
\|x_{k,l+1} - x_{k,j}\|_2 = \|P_C(x_{k,l} - \mu_k S_{k,l+1}) - x_{k,j}\|_2
\leq \|x_{k,l} - \mu_k S_{k,l+1} - x_{k,j}\|_2
\leq \|x_{k,l} - x_{k,j}\|_2 + \mu_k \|S_{k,l+1}\|_2
\leq (l + 1 - j) \mu_k L,
$$

where the first inequality is due to the fact that the projector $P_C$ is nonexpansive; the last inequality is due to $\|S_{k,l}\|_2 \leq L$, which follows from the bounded subgradients assumptions in Assumption 1 and Assumption 2.

3.3. Global sublinear convergence result. The following proposition provides an important recursion shared by all three incremental methods.

Proposition 1 (key recursion for global convergence). Suppose that Assumption 1 is valid when considering the incremental subgradient and proximal point methods, while Assumption 2 is valid when considering the incremental prox-linear method. Let $\{x_k\}_{k \geq 0}$ be the sequence generated by any of the three incremental methods for solving problem (1.1) with arbitrary initialization. Then, for any $\lambda < \frac{1}{2\tau}$ in (3.1), we have

$$
f_\lambda(x_{k+1}) \leq f_\lambda(x_k) - \left(\frac{1}{2\lambda} - \tau\right) \frac{1}{\lambda} m \mu_k \|x_k - P_{\lambda, f}(x_k)\|^2_2
\leq \left(1 + \frac{1}{\lambda} \frac{1}{1/\lambda - \tau}\right) \frac{1}{\lambda} m^2 \mu_k^2 L^2 + \left(1 + \frac{1}{\lambda} \frac{1}{1/\lambda - \tau}\right)^2 \frac{1}{\lambda} \tau m^3 \mu_k^3 L^2.
$$

Proof. From the optimality of $P_{\lambda, f}(x_{k,i})$ in (3.2) and the fact that $P_{\lambda, f}(x_{k,i-1}) \in C$, we have

$$
f_\lambda(x_{k,i}) \leq f(P_{\lambda, f}(x_{k,i-1})) + \frac{1}{2\lambda} \|x_{k,i} - P_{\lambda, f}(x_{k,i-1})\|^2_2.
$$

This, together with Lemma 2 (by letting $y = P_{\lambda, f}(x_{k,i-1})$) and the definition of the Moreau envelope $f_\lambda(x_{k,i-1}) = f(P_{\lambda, f}(x_{k,i-1})) + \frac{1}{2\lambda} \|x_{k,i-1} - P_{\lambda, f}(x_{k,i-1})\|^2_2$, gives

$$
f_\lambda(x_{k,i}) \leq f_\lambda(x_{k,i-1}) - \frac{\mu_k}{\lambda} (f_i(z_i) - f_i(P_{\lambda, f}(x_{k,i-1})))
\leq \frac{\tau \mu_k}{2\lambda} \|z_i - P_{\lambda, f}(x_{k,i-1})\|^2_2 + \frac{\gamma}{2\lambda} \|x_{k,i} - x_{k,i-1}\|^2_2.
$$
Summing the inequality in (3.13) over \(i = 1, \ldots, m\) yields

\[
\begin{align*}
\Delta_k(x_{k+1}) & \leq \frac{\mu_k}{\lambda} \sum_{i=1}^{m} \left( f_i(z_i) - f_i(P_{\lambda, f}(x_{k,i-1})) \right) \\
& \quad + \Delta_1 + \Delta_2 + \Delta_3.
\end{align*}
\]

(3.14)

where the last line utilizes the same Lipschitz continuous property of \(f\) as in (3.15). We now bound \(\Delta_1, \Delta_2, \) and \(\Delta_3\) in (3.14) for different incremental methods.

**Part I: Incremental subgradient method.** According to Lemma 2, we have \(z_i = x_{k,i-1}, \gamma = 1, \theta = 2\) in (3.14) for the incremental subgradient method. For \(\Delta_1\), we have

\[
\begin{align*}
\Delta_1 &= \sum_{i=1}^{m} \left[ f_i(x_{k,i-1}) - f_i(x_k) + f_i(P_{\lambda, f}(x_k)) \\
& \quad - f_i(P_{\lambda, f}(x_{k,i-1})) + f_i(x_k) - f_i(P_{\lambda, f}(x_k)) \right] \\
& \geq m[f(x_k) - f(P_{\lambda, f}(x_k))] \\
& \quad - L \sum_{i=1}^{m} \left[ \|x_{k,i-1} - x_k\|_2 + \|P_{\lambda, f}(x_{k,i-1}) - P_{\lambda, f}(x_k)\|_2 \right] \\
& \quad \geq m[f(x_k) - f(P_{\lambda, f}(x_k))] - L \left( 1 + \frac{1}{\lambda - \tau} \right) \sum_{i=1}^{m} \|x_{k,i-1} - x_k\|_2 \\
& \quad \geq m[f(x_k) - f(P_{\lambda, f}(x_k))] - \left( 1 + \frac{1}{\lambda - \tau} \right) \frac{m^2 - m}{2} \mu_k L^2,
\end{align*}
\]

(3.15)

where the first inequality utilizes the bounded subgradients assumption in Assumption 1 and [49, Theorem 9.13]; the second inequality follows from [49, Proposition 12.19], which states that for any \(\tau\)-weakly convex function \(f, P_{\lambda, f}\) is Lipschitz continuous with constant \((1/\lambda)/((1/\lambda) - \tau)\) if \(\lambda < \frac{1}{\tau}\); the last inequality is because of Lemma 3.

Similarly,

\[
\begin{align*}
\Delta_2 &= \sum_{i=1}^{m} \|x_{k,i-1} - x_k + P_{\lambda, f}(x_k) - P_{\lambda, f}(x_{k,i-1}) + x_k - P_{\lambda, f}(x_k)\|_2^2 \\
& \leq 2 \sum_{i=1}^{m} \left( \left( 1 + \frac{1}{\lambda - \tau} \right)^2 \|x_{k,i-1} - x_k\|_2^2 + \|x_k - P_{\lambda, f}(x_k)\|_2^2 \right),
\end{align*}
\]

(3.16)

where the last line utilizes the same Lipschitz continuous property of \(P_{\lambda, f}\) as in (3.15). By upper bounding \(\|x_{k,i-1} - x_k\|_2\) using Lemma 3, one can see that

\[
\Delta_2 \leq \left( 1 + \frac{1}{\lambda - \tau} \right)^2 \frac{(m-1)m(2m-1)}{3} \mu_k^2 L^2 + 2m \|x_k - P_{\lambda, f}(x_k)\|_2^2.
\]

(3.17)

For \(\Delta_3\), Lemma 3 gives

\[
\Delta_3 \leq \frac{1}{2\lambda} m \mu_k^2 L^2.
\]

(3.18)
Substituting (3.15), (3.17), and (3.18) into (3.14) yields

\begin{equation}
\begin{array}{c}
f_{\lambda}(x_{k+1}) \leq f_{\lambda}(x_k) - \frac{m \mu_k}{\lambda} (f(x_k) - f(P_{\lambda, f}(x_k))) - \tau \|x_k - P_{\lambda, f}(x_k)\|_2^2 \\
+ \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right) \frac{1}{\lambda} m^2 + m \mu_k^2 L^2 + \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right) \frac{2}{\lambda} \tau (m-1)m(2m-1)/6 \mu_k^3 L^2,
\end{array}
\end{equation}

where we have enlarged the term \( \frac{1}{\lambda} m \mu_k^2 L^2 \) in (3.18) to \( \frac{1}{\lambda} \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right) m \mu_k^2 L^2 \). Note that

\begin{equation}
f(x_k) - f(P_{\lambda, f}(x_k)) - \tau \|x_k - P_{\lambda, f}(x_k)\|_2^2
\end{equation}

\begin{equation}
= f(x_k) - \left(f(P_{\lambda, f}(x_k)) + \frac{1}{2\lambda} \|x_k - P_{\lambda, f}(x_k)\|_2^2\right) + \left(\frac{1}{2\lambda} - \tau\right) \|x_k - P_{\lambda, f}(x_k)\|_2^2
\end{equation}

\begin{equation}
\geq \left(\frac{1}{2\lambda} - \tau\right) \|x_k - P_{\lambda, f}(x_k)\|_2^2,
\end{equation}

where the inequality is from the definition of the Moreau envelope. Plugging the above inequality into (3.19) provides

\begin{equation}
f_{\lambda}(x_{k+1}) \leq f_{\lambda}(x_k) - \left(\frac{1}{2\lambda} - \tau\right) \frac{m \mu_k}{\lambda} \|x_k - P_{\lambda, f}(x_k)\|_2^2
\end{equation}

\begin{equation}
+ \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right) \frac{1}{\lambda} m^2 + m \mu_k^2 L^2 + \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right) \frac{2}{\lambda} \tau m^3 \mu_k^3 L^2.
\end{equation}

Note that we have enlarged \( \frac{m^2 + m}{2} \) to \( m^2 \) and \( \frac{(m-1)m(2m-1)}{6} \) to \( m^3 \) in order for all three incremental methods to satisfy this recursion. This yields the desired result for the incremental subgradient method.

**Part II: Incremental proximal point method.** In this case, we have \( z_i = x_{k,i} \) and \( \gamma = 0 \) (i.e., \( \Delta_3 = 0 \)) in (3.14). Using similar arguments as (3.15) and (3.16), we have

\begin{equation}
\Delta_1 \geq m[f(x_k) - f(P_{\lambda, f}(x_k))] - L \sum_{i=1}^m \|x_{k,i} - x_k\|_2 - L \frac{1}{1/\lambda - \tau} \sum_{i=1}^m \|x_{k,i-1} - x_k\|_2
\end{equation}

\begin{equation}
\geq m[f(x_k) - f(P_{\lambda, f}(x_k))] - \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right) \frac{m^2 + m}{2} \mu_k L^2
\end{equation}

and

\begin{equation}
\Delta_2 = \sum_{i=1}^m \|x_{k,i} - x_k + P_{\lambda, f}(x_k) - P_{\lambda, f}(x_{k,i-1}) + x_k - P_{\lambda, f}(x_k)\|_2^2
\end{equation}

\begin{equation}
\leq 2 \sum_{i=1}^m \left(\|x_{k,i} - x_k\|_2^2 + \frac{1/\lambda}{1/\lambda - \tau} \|x_{k,i-1} - x_k\|_2^2\right)^2 + 2m \|x_k - P_{\lambda, f}(x_k)\|_2^2
\end{equation}

\begin{equation}
\leq \left(1 + \frac{1/\lambda}{1/\lambda - \tau}\right)^2 \frac{m(m+1)(2m+1)}{3} \mu_k^3 L^2 + 2m \|x_k - P_{\lambda, f}(x_k)\|_2^2.
\end{equation}

Substituting (3.22) and (3.23) into (3.14) and applying (3.20) lead to exactly the same recursion as (3.21), where we have enlarged \( \frac{m^2 + m}{2} \) to \( m^2 \) and \( \frac{(m-1)m(2m-1)}{6} \) to \( m^3 \).
Part III: Incremental prox-linear method. According to Lemma 2, we have $z_i = x_{k,i-1}$, $\gamma = 2\mu_k L$, and $\theta = 1$ in (3.14) for the incremental prox-linear method. In this case, the bounds for $\Delta_1$ and $\Delta_2$ are exactly the same as (3.15) and (3.17), respectively. For $\Delta_3$, (3.11) gives $\Delta_3 \leq \frac{1}{2} m^2 \mu_k^2 L^2$. Substituting the bounds for $\Delta_1$, $\Delta_2$, and $\Delta_3$ into (3.14) and applying (3.20), we obtain exactly the same recursion as (3.21).

Equipped with the above proposition, we are ready to establish our global convergence result.

Theorem 1 (global convergence). Under the setting of Proposition 1, the following hold:

(a) If we choose the constant stepsize $\mu_k = \frac{1}{m\tau \sqrt{T+1}}$ for $k \geq 0$ with $T$ being the total number of iterations, then

$$\min_{0 \leq k \leq T} \Theta^2(x_k) \leq \frac{C_1}{(\frac{1}{2\lambda} - \tau) \sqrt{T+1}} + \frac{C_2}{(\frac{1}{2\lambda} - \tau)(T+1)}.$$  

(b) If we choose the diminishing stepsizes $\mu_k = \frac{1}{m\tau \sqrt{k+1}}$ for $k \geq 0$, then

$$\min_{0 \leq k \leq T} \Theta^2(x_k) \leq \frac{C_1}{(\frac{1}{2\lambda} - \tau) \sqrt{T+1}} + \frac{C_2(\ln(T+1) + 1)}{(\frac{1}{2\lambda} - \tau)(T+1)}.$$  

Here, $C_1 = \frac{1}{\tau} (f_\lambda(x_0) - \min f_\lambda) + \left(1 + \frac{1}{\lambda \sqrt{T+1}}\right) \frac{1}{\sqrt{T+1}} L^2$ and $C_2 = \left(1 + \frac{1}{1/\lambda - \tau}\right)^2 \frac{1}{\lambda \tau} L^2$.

Proof. Unrolling the recursion (3.12) in Proposition 1 and invoking the definition $\Theta(x_k) = \frac{1}{\lambda} \|x_k - P_{\lambda f}(x_k)\|_2$ from (3.3), we obtain

(3.24)

$$\left(\frac{1}{2\lambda} - \tau\right) \left(\sum_{k=0}^T \mu_k \right) m \lambda \min_{0 \leq k \leq T} \Theta^2(x_k) \leq f_\lambda(x_0) - \min f_\lambda$$  

$$+ \left(1 + \frac{1}{\lambda \sqrt{T+1}}\right) \frac{1}{\lambda} m^2 \left(\sum_{k=0}^T \mu_k^2 \right) L^2 + \left(1 + \frac{1}{1/\lambda - \tau}\right)^2 \frac{\tau}{\lambda} m^3 \left(\sum_{k=0}^T \mu_k^3\right) L^2.$$  

Dividing $\left(\frac{1}{2\lambda} - \tau\right) \left(\sum_{k=0}^T \mu_k \right) m \lambda$ on both sides of (3.24) gives

$$\min_{0 \leq k \leq T} \Theta^2(x_k) \leq \frac{1}{\lambda} (f_\lambda(x_0) - \min f_\lambda) + \left(1 + \frac{1}{\lambda \sqrt{T+1}}\right) \frac{1}{\lambda} m^2 \left(\sum_{k=0}^T \mu_k^2 \right) L^2$$  

$$+ \left(1 + \frac{1}{1/\lambda - \tau}\right)^2 \frac{\tau}{\lambda} m^3 \left(\sum_{k=0}^T \mu_k^3\right) L^2$$

$$\leq \frac{1}{\lambda} (f_\lambda(x_0) - \min f_\lambda) + \left(1 + \frac{1}{\lambda \sqrt{T+1}}\right) \frac{1}{\lambda} m^2 \left(\sum_{k=0}^T \mu_k^2 \right) L^2$$  

$$+ \left(1 + \frac{1}{1/\lambda - \tau}\right)^2 \frac{\tau}{\lambda} m^3 \left(\sum_{k=0}^T \mu_k^3\right) L^2$$

(3.25)

The result in (a) follows by taking the constant stepsize $\mu_k = \frac{1}{m\tau \sqrt{T+1}}$ in (3.25). The result in (b) follows by taking the diminishing stepsizes $\mu_k = \frac{1}{m\tau \sqrt{k+1}}$ in (3.25) and recognizing that $\sum_{k=0}^T \frac{1}{\sqrt{k+1}} > \sqrt{T+1}$ and $\sum_{k=0}^T \frac{1}{k+1} < \ln(T+1) + 1$.  

Theorem 1 implies that the iteration complexity of the incremental methods for computing an $\varepsilon$-nearly stationary point of problem (1.1) is $O(\varepsilon^{-2})$, which matches that of their stochastic counterparts [13].
4. Linear convergence for sharp weakly convex optimization. The global sublinear convergence result discussed in the last section does not rely on any specific structure of problem (1.1) besides weak convexity (or the slightly stronger quadratic approximation property in Assumption 2). Nonetheless, many applications give rise to instances of problem (1.1) that have additional structures, which can potentially be exploited by our incremental methods to achieve faster convergence rates. In this section, we focus on a structural property called \textit{sharpness} and show that all three incremental methods for solving problem (1.1) will achieve a local linear rate of convergence when the problem possesses the sharpness property.

4.1. Sharpness: Weak sharp minima. We start by defining the sharpness property through the notion of weak sharp minima.

\textbf{Definition 1 (sharpness; cf. [10]).} Consider problem (1.1). We say that $\mathcal{X} \subseteq \mathcal{C}$ is a set of weak sharp minima for the function $f : \mathbb{R}^n \to \mathbb{R}$ over $\mathcal{C}$ with parameter $\alpha > 0$ if for any $x \in \mathcal{C}$, we have

$$f(x) - f(y) \geq \alpha \text{dist}(x, \mathcal{X})$$

for all $y \in \mathcal{X}$. We say that problem (1.1) possesses the sharpness property if it has a set of weak sharp minima.

The notion of sharpness plays an important role in establishing linear convergence results for subgradient-based methods and superlinear convergence results for proximal-based methods; see, e.g., [15, 20, 38, 51].

As it turns out, many concrete applications give rise to sharp instances of problem (1.1). For instance, it is shown in [33] that the robust matrix sensing (RMS) problem (1.4) possesses the sharpness property under certain statistical conditions. The work [33] also reveals a more general implication, namely once the measurement operator $\mathcal{A}$ possesses the so-called $\ell_1/\ell_2$-RIP (see [33] for the definition), then the associated problem will have the sharpness property. Such an implication opens up the possibility of establishing the sharpness property of a wide class of signal recovery problems, such as robust blind deconvolution [12] and robust phase retrieval problems [11, 18].

Now, suppose that problem (1.1) possesses the sharpness property with parameter $\alpha > 0$. Consider the case where all the component functions $f_1, \ldots, f_m$ in problem (1.1) satisfy the bounded subgradients assumption in Assumption 1 with parameter $L_1$ on an open convex set $Q$ that contains $\mathcal{C}$. Let

$$L_1 := \sup \left\{ \| \tilde{\nabla} f(x) \|_2 : x \in Q, \tilde{\nabla} f(x) \in \partial f(x) \right\}.$$ 

It is clear that $L_1 \geq \bar{L}_1$. According to [49, Theorem 9.13], $f$ is Lipschitz continuous on $Q$ with parameter $\bar{L}_1$. This, together with Definition 1 (take $y \in \mathcal{P}_{\mathcal{X}}(x)$), gives

$$\alpha \text{dist}(x, \mathcal{X}) \leq f(x) - f(y) \leq \bar{L}_1 \text{dist}(x, \mathcal{X}).$$

Consequently, we have $\alpha \leq \bar{L}_1 \leq L_1$.

Next, consider the case where all the component functions $f_1, \ldots, f_m$ in problem (1.1) satisfy the bounded subgradients assumption in Assumption 2 with parameter $L_2$ on an open convex set $Q$ that contains $\mathcal{C}$. Let

$$L_2 := \sup \left\{ \| \tilde{\nabla} f(x; \mathbf{\bar{x}}) \|_2 : x, \mathbf{\bar{x}} \in Q; \tilde{\nabla} f(x; \mathbf{\bar{x}}) \in \partial f(x; \mathbf{\bar{x}}) \right\}.$$
It is clear that $L_2 \geq \tilde{L}_2$. By [49, Theorem 9.13], $x \mapsto f(x; \bar{x})$ is Lipschitz continuous on $Q$ with parameter $\tilde{L}_2$ for any $\bar{x} \in Q$, which gives

$$f(\bar{x}) - f(x; \bar{x}) = f(x; \bar{x}) - f(x; \bar{x}) \leq \tilde{L}_2\|x - \bar{x}\|_2.$$ 

According to Assumption 2, we have

$$f(x; \bar{x}) - f(x) \leq \frac{\tau}{2} \|x - \bar{x}\|_2.$$ 

It follows that for any $x \in Q$ satisfying $x \neq x$,

$$f(x; \bar{x}) - f(x) \leq \tilde{L}_2 + \frac{\tau}{2} \|x - \bar{x}\|_2.$$ 

Since the function $f$ is weakly convex, its subdifferential can be equivalently characterized as

$$\partial f(x) = \left\{ \tilde{\nabla} f(x) \in \mathbb{R}^n : \liminf_{\bar{x} \to x} \frac{f(\bar{x}) - f(x) - \langle \tilde{\nabla} f(x), \bar{x} - x \rangle}{\|\bar{x} - x\|_2} \geq 0 \right\};$$

see, e.g., [31]. Upon taking $\bar{x} = x + t\tilde{\nabla} f(x)$ with $t \downarrow 0$ and plugging (4.2) into (4.3), we get

$$\|\tilde{\nabla} f(x)\|_2 \leq \tilde{L}_2, \quad \forall x \in Q, \tilde{\nabla} f(x) \in \partial f(x).$$

By [49, Theorem 9.13], this implies that $f$ is Lipschitz continuous with parameter $\tilde{L}_2$ on $Q$. It then follows from (4.1) that $\alpha \leq \tilde{L}_2 \leq L_2$.

Summarizing the above discussion, since $L = \max\{L_1, L_2\}$, we obtain the following intrinsic relation:

$$\alpha \leq L.$$

### 4.2. The key recursion.

The following important recursion, which is shared by all three incremental methods, allows us to exploit the sharpness property in our convergence analysis of these methods.

**Proposition 2** (key recursion for linear convergence). Under the setting of Proposition 1, for any $x^* \in C$, we have

$$\|x_{k+1} - x^*\|_2^2 \leq (1 + 2m\tau\mu_k)\|x_k - x^*\|_2^2 - 2m\mu_k(f(x_k) - f(x^*)) + 2m^2\mu_k^2L^2 + 2\tau m^3\mu_k^3L^2.$$ 

**Proof.** Letting $y = x^*$ in Lemma 2 and summing (3.6) over $i = 1, \ldots, m$ give

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - 2\mu_k \sum_{i=1}^m (f_i(z_i) - f_i(x^*))$$

$$+ \tau\mu_k \sum_{i=1}^m \|z_i - x^*\|_2^2 + \gamma \sum_{i=1}^m \|x_{k,i} - x_{k,i-1}\|_2^2.$$ 

We now derive bounds for $\Delta_1$, $\Delta_2$, and $\Delta_3$ in (4.6) for different incremental methods.
Note that we have enlarged $m$ for this case. To bound $\Delta_1$ and $\Delta_2$, following the derivations in (3.15)–(3.18), we have

$$\Delta_1 = \sum_{i=1}^{m} (f_i(x_{k,i-1}) - f_i(x_k) + f_i(x_k) - f_i(x^*))$$

$$\geq m(f(x_k) - f(x^*)) - \frac{m^2 - m}{2} \mu_k L^2,$$

(4.7)

$$\Delta_2 = \sum_{i=1}^{m} \|x_{k,i-1} - x_k + x_k - x^*\|_2^2$$

$$\leq 2m\|x_k - x^*\|_2^2 + \frac{(m-1)m(2m-1)}{3} \mu_k^2 L^2,$$

(4.8)

and

$$\Delta_3 \leq m\mu_k^2 L^2.$$ (4.9)

Substituting (4.7)–(4.9) into (4.6) yields

$$\|x_{k+1} - x^*\|_2^2 \leq (1 + 2m\tau\mu_k)\|x_k - x^*\|_2^2 - 2m\mu_k(f(x_k) - f(x^*))$$

$$+ 2m^2\mu_k^2 L^2 + 2\tau m^3\mu_k^3 L^2.$$ (4.10)

Note that we have enlarged $m^2\mu_k^2 L^2$ to $2m^2\mu_k^2 L^2$ and $\frac{(m-1)m(2m-1)}{3}$ to $2m^3$ in order for all three incremental methods to satisfy this recursion.

**Part II: Incremental proximal point method.** Recall that $z_i = x_{k,i}$ and $\gamma = 0$ (i.e., $\Delta_3 = 0$) in (4.6) for this case. Following the analysis in (3.22)–(3.23), we have

$$\Delta_1 \geq m(f(x_k) - f(x^*)) - \frac{m^2 + m}{2} \mu_k L^2,$$

(4.11)

$$\Delta_2 \leq 2m\|x_k - x^*\|_2^2 + \frac{m(m + 1)(2m + 1)}{3} \mu_k^2 L^2.$$ (4.12)

Substituting (4.11) into (4.6) yields the same recursion as (4.10), where we have enlarged $m^2 + m$ to $2m^2$ and $\frac{m(m+1)(2m+1)}{3}$ to $2m^3$.

**Part III: Incremental prox-linear method.** Recall that $z_i = x_{k,i-1}$, $\gamma = 2\mu_k L$, $\theta = 1$ in (4.6) for this case. Thus, the bounds for $\Delta_1$ and $\Delta_2$ will be the same as those in (4.7) and (4.8), respectively. In addition, we have $\Delta_3 \leq 2m\mu_k^2 L^2$ from (3.11). Substituting the bounds for $\Delta_1$, $\Delta_2$, and $\Delta_3$ into (4.6) yields the same recursion as (4.10).

**4.3. Linear convergence result.** It is known that subgradient-based methods with a constant stepsize for nonsmooth optimization may never converge to the optimal solution set. In order to get exact convergence, diminishing stepsizes are generally needed [20, 51]. The work [38] analyzes the incremental subgradient method using Polyak’s dynamic stepsizes $\mu_k = (f(x_k) - f^*)/\|\nabla f(x_k)\|^2$ for some $\nabla f(x_k) \in \partial f(x_k)$ with $\nabla f(x_k) \neq 0$, which can be regarded as an adaptive diminishing stepsize rule. However, the implementation of Polyak’s dynamic stepsizes requires the knowledge of the optimal function value $f^*$, which makes it impractical. Motivated by previous works on full subgradient methods [15, 20, 51], in this section, we show that by using
have

where the initial point \( x(1) \) sequence generated by any of the three incremental methods for solving problem (1.1) (see Definition 1). Let \( \{x_k\}_{k \geq 0} \) be the sequence generated by any of the three incremental methods for solving problem (1.1), where the initial point \( x_0 \) satisfies

\[
dist(x_0, \mathcal{X}) < \frac{\alpha}{\tau}
\]

and the stepsizes are chosen as

\[
\mu_k = \rho^k \cdot \mu_0
\]

with

\[
\mu_0 < \frac{\alpha^2}{6mL^2} \left( 1 - \left( \max \left\{ \frac{2\tau}{\alpha} \dist(x_0, \mathcal{X}) - 1, 0 \right\} \right)^2 \right),
\]

\[
\rho \in [\rho_{\min}, 1), \quad \rho_{\min} := \sqrt{1 + 2m \left( \tau - \frac{\alpha}{e_0} \right) \mu_0 + \frac{3m^2L^2}{\epsilon_0^2} \mu_0^2},
\]

and

\[
e_0 := \max \left\{ \dist(x_0, \mathcal{X}), \frac{3mL^2\mu_0}{\alpha} \right\}.
\]

Then, we have

\[
dist(x_k, \mathcal{X}) \leq \rho^k \cdot e_0, \quad \forall \ k \geq 0.
\]

**Proof.** We first show that \( \rho_{\min} \in (0, 1) \), so that \( \rho \) is well defined. Note that \( \rho_{\min} = \sqrt{1 + v(\mu_0)} \) with \( v(\mu_0) = 2m \left( \tau - \frac{\alpha}{e_0} \right) \mu_0 + \frac{3m^2L^2}{\epsilon_0^2} \mu_0^2 \) being a quadratic function of \( \mu_0 \).

To establish \( \rho_{\min} \in (0, 1) \), it suffices to show that \( v(\mu_0) \in (-1, 0) \). On one hand, we have \( v(\mu_0) > -\frac{2m\alpha\mu_0}{\epsilon_0} \geq -\frac{2m\alpha^2}{3\tau^2} > -1 \) by the definition of \( e_0 \) and the fact that \( L \geq \alpha \) (see (4.4)). On the other hand, we prove that \( v(\mu_0) < 0 \), or equivalently,

\[
(4.13) \quad \mu_0 < \frac{2(\alpha e_0 - \tau e_0^2)}{3mL^2}.
\]

Towards establishing (4.13), we consider the cases \( \tau = 0 \) and \( \tau > 0 \) separately. In the case where \( \tau = 0 \) (i.e., \( f \) is convex), (4.13) can be ensured by the definition of \( e_0 \). In the case where \( \tau > 0 \), since \( \mu_0 < \frac{\alpha^2}{6mL^2\tau} \), we have \( \frac{3mL^2\mu_0}{\epsilon_0^2} < \frac{\alpha^2}{2\tau} \). If \( \dist(x_0, \mathcal{X}) \geq \frac{\alpha}{2\tau} \), then \( e_0 = \dist(x_0, \mathcal{X}) \geq \frac{\alpha}{2\tau} \) and the inequality defining \( \mu_0 \) reduces to

\[
\mu_0 < \frac{2(\alpha \dist(x_0, \mathcal{X}) - \tau \dist(x_0, \mathcal{X})^2)}{3mL^2},
\]

which is exactly (4.13). If \( \dist(x_0, \mathcal{X}) < \frac{\alpha}{2\tau} \), then \( e_0 < \frac{\alpha}{2\tau} \), \( \mu_0 < \frac{\alpha^2}{6mL^2\tau} \), and (4.13) is equivalent to

\[
2\tau e_0^2 - 2\alpha e_0 + 3mL^2\mu_0 < 0.
\]

The above is a quadratic inequality in \( e_0 \), which is further equivalent to

\[
(4.14) \quad \frac{\alpha - \sqrt{\alpha^2 - 6m\tau L^2 \mu_0}}{2\tau} < e_0 < \frac{\alpha + \sqrt{\alpha^2 - 6m\tau L^2 \mu_0}}{2\tau}.
\]
The second inequality in (4.14) is ensured by $e_0 < \frac{\sqrt{\tau}}{2\tau}$, while the first inequality follows from the definition of $e_0$ and the fact that
\[
\frac{\alpha}{\sqrt{\alpha^2 - 6m\tau L^2 \mu_0}} = \frac{6m\tau L^2 \mu_0}{2\tau(\alpha + \sqrt{\alpha^2 - 6m\tau L^2 \mu_0})} < \frac{3mL^2 \mu_0}{\alpha} \leq e_0. \quad \text{Hence, we have proved that} \quad \rho_{\min} \in (0, 1).
\]

We now prove (4.12) by induction. It is trivially true when $k = 0$ due to the definition of $e_0$. Assuming that $\text{dist}(x_k, X) \leq \rho^k \cdot e_0$ for some $k \geq 0$, it remains to show that $\text{dist}(x_{k+1}, X) \leq \rho^{k+1} \cdot e_0$.

By applying Proposition 2 with $x^* = P_X(x_k)$, using the fact that $\text{dist}(x_{k+1}, X) \leq \|x_{k+1} - x^*\|_2$ and $\rho < 1$, and utilizing the sharpness property, we have
\[
\text{dist}(x_{k+1}, X) \leq (1 + 2m\tau \mu_0) \text{dist}(x_k, X) - 2m\alpha \mu_k \text{dist}(x_k, X) + 2m^2 \mu_k^2 L^2 + 2\tau m^3 \mu_k^3 L^2.
\]
(4.15)

Observe that the right-hand side of (4.15) is a quadratic function of $\text{dist}(x_k, X)$. Since $0 \leq \text{dist}(x_k, X) \leq \rho^k \cdot e_0$, it follows that the right-hand side of (4.15) achieves its maximum at $\text{dist}(x_k, X) = \rho^k \cdot e_0$, as one has $\frac{2m\alpha \mu_0}{1 + 2m\tau \mu_0} < 2m\alpha \mu_0 < e_0$ by the definition of $e_0$ and the fact that $L \geq \alpha$ (see (4.4)). Upon plugging $\text{dist}(x_k, X) = \rho^k \cdot e_0$ and $\mu_k = \rho^k \cdot \mu_0$ into (4.15), we obtain
\[
\text{dist}(x_{k+1}, X) \leq \rho^{2k} e_0^2 + 2m\tau \mu_0 \rho^{2k} e_0^2 - 2m\alpha \mu_0 \rho^{2k} e_0^2 + 2m^2 \mu_k^2 L^2 + 2\tau m^3 \mu_k^3 L^2 \rho^{2k}
\]
\[
= \rho^{2k} e_0^2 \left(1 + 2m \left(\tau - \frac{\alpha}{e_0}\right) \mu_0 + \frac{2m^2 L^2 + 2\tau m^3 \mu_0 L^2}{e_0^2 \mu_0^2}\right)
\]
\[
< \rho^{2k} e_0^2 \left(1 + 2m \left(\tau - \frac{\alpha}{e_0}\right) \mu_0 + \frac{3m^2 L^2}{e_0^2 \mu_0^2}\right)
\]
\[
= \rho^{2k} e_0^2 \cdot \rho_{\min}^2
\]
\[
\leq \rho^{2k + 2} e_0^2,
\]
where in the first inequality we use $\rho < 1$ and in the second inequality we use $\tau \mu_0 < \frac{1}{6m}$ by (4.4) and the definition of $\mu_0$. This completes the inductive step and hence the proof of the theorem.

Before we leave this section, let us make two remarks. First, Theorem 2 provides a unified analysis of the incremental methods in both the convex (i.e., $\tau = 0$) and nonconvex (i.e., $\tau > 0$) cases. In the convex case, the initial distance $\text{dist}(x_0, X)$ and initial stepsize $\mu_0$ can be chosen arbitrarily. The diminishing stepsize rule in Theorem 2 is practical in the sense of implementation. The initial stepsize $\mu_0$ and decay factor $\rho$ can be computed with the knowledge of the problem parameters $\alpha, \tau, L$ and the initial distance. Even in the case where it is difficult to estimate these parameters accurately, one can still obtain proper $\mu_0$ and $\rho$ for specific applications. This is in sharp contrast to Polyak’s dynamic stepsize rule, which requires the knowledge of the optimal function value $f^*$. Such information is often hard to obtain in practice.

Second, we note that the three incremental methods may have different performance in practice, although Theorem 2 has the same convergence guarantee for them. In other words, Theorem 2 only provides sufficient (but possibly not necessary) conditions for the methods to converge and (4.12) only provides an upper bound on the convergence rate. In the next section, we show the (different) performance of the three incremental methods for solving the RMS problem (1.4). As will be seen, the incremental prox-linear methods often performs best.
5. Experiments. In this section, we conduct a series of experiments on the three incremental methods for solving the RMS problem (1.4).\footnote{Our code is available at https://github.com/lixiao0982/Incremental-methods} We list below the abbreviations of all the algorithms that appear in this section.

\begin{center}
\begin{tabular}{ll}
SGM: & Full subgradient method \\
SGD: & Stochastic subgradient method \\
SPL: & Stochastic prox-linear method \\
ISG: & Incremental subgradient method \\
IPP: & Incremental proximal point method \\
IPL: & Incremental prox-linear method \\
\end{tabular}
\end{center}

We generate $U^* \in \mathbb{R}^{n \times r}$ and the $m$ sensing matrices $A_1, \ldots, A_m \in \mathbb{R}^{n \times n}$ (which defines the linear operator $A$) with i.i.d. standard Gaussian entries. The ground-truth low-rank matrix is then generated by $X^* = U^* U^*^T$. We generate the outlier vector $s^* \in \mathbb{R}^m$ by first randomly selecting $pn$ locations, where $p$ is the ratio of outliers. Then, each of the selected location is filled with an i.i.d. mean 0 and variance 10 Gaussian, while the remaining locations are set to 0. According to (1.3), the linear measurement $y \in \mathbb{R}^m$ is generated by $y_i = \langle A_i, X^* \rangle + s^*_i$, $i = 1, \ldots, m$. We set $n = 50$, $r = 5$, $m = 5nr$, $p = 0.3$. As stated in [33, Proposition 2], the RMS problem possesses the sharpness property (see Definition 1) under this setting. Furthermore, its optimal solution set is precisely $\mathcal{U} = \{U^* R : R \in \mathbb{R}^{r \times r}, RR^T = I\}$. The bounded subgradients assumption of the RMS problem (1.4) on any bounded subset can be verified using [33, Propsition 4]. We have now checked the assumptions in our Theorem 2 (i.e., weak convexity, sharpness, and bounded subgradients) for the RMS problem. According to Theorem 2, in order to achieve linear convergence for the three incremental methods, we need to utilize the geometrically diminishing stepsizes $\mu_k = \rho^k \mu_0$, $k \geq 0$. For a fair comparison, all the algorithms are initialized with the same point, whose entries are i.i.d. standard Gaussians (i.e., random initialization).

It is worth mentioning that though Theorem 2 requires a special initialization, the random initialization works as good as carefully designed ones (see, e.g., [33, Theorem 4] for an initialization strategy) in our experiments.

Recall from Theorem 2 that the three incremental methods converge linearly at the rate of $O(\rho^k)$, where $\rho$ is exactly the decay factor of the geometrically diminishing stepsizes. Thus, the smaller decay factor $\rho$ an algorithm can choose, the faster it converges. Based on this observation, we evaluate an algorithm by numerically testing the smallest possible $\rho$ it can work with.

5.1. Efficiency of incremental methods. We first compare the performance of the three incremental methods studied in this paper. The result is shown in Figure 1. We run all the incremental methods for 500 iterations, where the initial stepsize $\mu_0$ and the decay factor $\rho$ of the geometrically diminishing stepsizes are selected from $\{1/m, 30/m, 60/m, \ldots, 240/m\}$ and $\{0.65, 0.7, 0.75, \ldots, 0.95, 0.99\}$, respectively. For each pair of parameters $(\rho, \mu_0)$, we average the distances to the optimal solution set of the last 5 iterates. The assessed algorithm is deemed successful if this averaged distance is no more than $10^{-8}$ and failed otherwise. In Figures 1a to 1c, each plot represents the success probability of 25 independent trials for different pairs $(\rho, \mu_0)$. The block with white color indicates that the algorithm is 100% successful under the corresponding parameter settings, while the block with black color implies that all 25 runs failed. In between, whiter color implies a higher successful rate. We can ob-
serve from Figures 1a to 1c that the three incremental methods have nearly the same lower bound for the decay factor $\rho$ (i.e., $\rho = 0.8$) when solving the RMS problem. Nonetheless, IPL and IPP are much more robust with respect to the selection of the initial stepsize $\mu_0$. In our test, both IPL and IPP work even with a very large $\mu_0$, say $\mu_0 = 10^4/m$. Such a robustness property is crucial in practice, as it allows much less hand-tuning.

To have a more detailed interpretation of Figures 1a to 1c, we show in Figure 1d the convergence progress of ISG for several representative pairs of parameters $(\rho, \mu_0)$ used in Figure 1a (the interpretation of Figures 1b and 1c will be similar). From the convergence plot, we can draw the following conclusions: 1) If we set $\rho = 0.65$ (say $\rho = 0.65$ and $\mu_0 = 1/m$), then ISG stops progressing in a few iterations. 2) If we set $\rho = 0.99$ and choose $\mu_0$ carefully (say $\rho = 0.99$ and $\mu_0 = 1/m$), then ISG converges very slowly and drives the distance to below $10^{-8}$ after nearly 5000 iterations. 3) If we set $\rho \in [0.8, 0.95]$ and choose $\mu_0$ appropriately (say $\rho = 0.8$ and $\mu_0 = 30/m$ or $\rho = 0.9$ and $\mu_0 = 1/m$), then ISG drives the distance to below $10^{-8}$ within 500 iterations. We can also observe that smaller $\rho$ leads to faster convergence of ISG, which corroborates the results in Theorem 2. 4) If we set $\rho \in [0.8, 0.99]$ but choose a large $\mu_0$ (say $\rho = 0.95$ and $\mu_0 = 60/m$), IGD diverges. It is interesting to mention that in our test, IPL and IPP will not diverge even with a large $\mu_0$ (say $\rho \in [0.8, 0.99]$ and $\mu_0 \in [1/m, 10^4/m]$). This indicates that IPL and IPP are much more robust than ISG with respect to the choice of the initial stepsize $\mu_0$.

We end this subsection by commenting on IPL. It can be observed from Figures 1b and 1c that IPL and IPP have nearly the same performance for different choices of the pair of stepsize parameters $(\rho, \mu_0)$. However, IPP requires an inner solver for its subproblem (2.5), which renders it computationally inefficient. By contrast, there exists a closed-form solution for the subproblem (2.7) of IPL when $h_i$ in (2.6) is the absolute value function, which is the case for the RMS problem. This implies that the update of IPL is as efficient as that of ISG and much more efficient than the update of IPP. Recall that IPL is much more robust than ISG with respect to the choice of the initial stepsize $\mu_0$. Thus, we recommend using IPL whenever its subproblem has a closed-form solution. In the sequel, we will omit the results of IPP as its performance is almost the same as IPL.

5.2. Comparison with SGM and stochastic methods. In this subsection, we compare the incremental methods with SGM and stochastic methods. We apply the same geometrically diminishing stepsizes to both SGM and stochastic methods. All the experimental settings remain the same as those in the last subsection. The result is displayed in Figure 2.

One can observe from Figures 2a and 2b that ISG can tolerate a much smaller choice of $\rho$ than SGM. The smallest workable $\rho$ for SGM is 0.93, while ISG can set $\rho = 0.8$; see Figure 2c for a comparison of the convergence speeds of ISG and SGM when their smallest workable decay factors $\rho$ are used.

In the comparison with stochastic methods, if a stochastic method runs for $T$ iterations, we actually count the iteration number as $T/m$. As observed from Figures 2d and 2e and Figures 2g and 2h, incremental methods can choose a smaller $\rho$ than their stochastic counterparts. In particular, incremental methods work with $\rho = 0.8$, while the smallest workable $\rho$ for the stochastic methods is 0.9. Figures 2f and 2i compare the convergence speeds of incremental and stochastic methods using the smallest possible $\rho$ of each algorithm. The superiority of the incremental methods is clear.
INCREMENTAL METHODS FOR WEAKLY CONVEX OPTIMIZATION

6. Conclusion. In this work, we introduced a family of incremental methods for solving the weakly convex minimization problem (1.1) and developed a unified framework for analyzing their convergence rates. In particular, we showed that these incremental methods have an iteration complexity of $O(\varepsilon^{-4})$ for computing an $\varepsilon$-nearly stationary point of problem (1.1), and that when applied to a sharp instance of problem (1.1) using a good initialization and geometrically diminishing stepsizes, they converge linearly to an optimal solution. Our work is the first to extend the convergence rate analysis of incremental methods from the nonsmooth convex regime to the weakly convex regime, which covers a large class of nonsmooth nonconvex problems. We also conducted a series of numerical experiments on the RMS problem (1.4) to demonstrate the efficacy of the incremental methods. We found that the incremental subgradient method has the simplest implementation, while the incremental proximal point and prox-linear methods are more robust with respect to the choice of the initial stepsize. Moreover, our numerical results suggested that the incremental prox-linear method would be the method of choice when its subproblem can be solved analytically, which is the case for the two motivating applications in Subsection 1.1.
Fig. 2: Comparison between incremental methods and SGM and stochastic methods for solving the RMS problem.

Our experiments clearly showed the superiority of incremental methods over their stochastic counterparts or the full subgradient method, which explains in part why incremental methods are so widely used in practice.

REFERENCES

[1] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[2] D. P. Bertsekas, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, Optimization for Machine Learning, 2010 (2011), p. 3.

[3] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Mathematical programming, 129 (2011), p. 163.

[4] D. P. Bertsekas, *Incremental aggregated proximal and augmented Lagrangian algorithms*, arXiv preprint arXiv:1509.09257, (2015).

[5] D. P. Bertsekas, *Feature-based aggregation and deep reinforcement learning: A survey and some new implementations*, arXiv preprint arXiv:1804.04577, (2018).

[6] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM Journal on Optimization, 18 (2007), pp. 29-51.
[7] L. Bottou, Curiously fast convergence of some stochastic gradient descent algorithms, in Proceedings of the Symposium on Learning and Data Science, Paris, 2009.
[8] L. Bottou, Stochastic gradient descent tricks, in Neural Networks: Tricks of the Trade, Springer, 2012, pp. 421–436.
[9] S. Burer and R. D. Monteiro, A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization, Mathematical Programming, 95 (2003), pp. 329–357.
[10] J. V. Burke and M. C. Ferris, Weak sharp minima in mathematical programming, SIAM Journal on Control and Optimization, 31 (1993), pp. 1340–1359.
[11] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, Low-rank matrix recovery with composite optimization: Good conditioning and rapid convergence. Accepted for publication in Foundations of Computational Mathematics, 2021.
[12] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, Composite optimization for robust blind deconvolution, arXiv preprint arXiv:1901.01624, (2019).
[13] D. Davis and D. Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, SIAM Journal on Optimization, 29 (2019), pp. 207–239.
[14] D. Davis, D. Drusvyatskiy, and V. Charisopoulos, Stochastic algorithms with geometric step decay converge linearly on sharp functions, arXiv preprint arXiv:1907.09547, (2019).
[15] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, Subgradient methods for sharp weakly convex functions, Journal of Optimization Theory and Applications, 179 (2018), pp. 962–982.
[16] D. Davis and B. Grimmer, Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems, SIAM Journal on Optimization, 29 (2019), pp. 1908–1930.
[17] D. Drusvyatskiy and C. Paquette, Efficiency of minimizing compositions of convex functions and smooth maps, Mathematical Programming, 178 (2019), pp. 503–558.
[18] J. C. Duchi and F. Ruan, Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval, Information and Inference: A Journal of the IMA, 8 (2018), pp. 471–529.
[19] J. C. Duchi and F. Ruan, Stochastic methods for composite and weakly convex optimization problems, SIAM Journal on Optimization, 28 (2018), pp. 3229–3259.
[20] J.-L. Goffin, On convergence rates of subgradient optimization methods, Mathematical programming, 13 (1977), pp. 329–347.
[21] L. Grippo, Convergent on-line algorithms for supervised learning in neural networks, IEEE Transactions on Neural Networks, 11 (2000), pp. 1284–1299.
[22] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo, Why random reshuffling beats stochastic gradient descent, Mathematical Programming, 186 (2021), pp. 49–84.
[23] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo, On the convergence rate of incremental aggregated gradient algorithms, SIAM Journal on Optimization, 27 (2017), pp. 1035–1048.
[24] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo, Convergence rate of incremental gradient and incremental Newton methods, SIAM Journal on Optimization, 29 (2019), pp. 2542–2565.
[25] J. Z. HaoChen and S. Sra, Random shuffling beats SGD after finite epochs, in International Conference on Machine Learning, 2019, pp. 2624–2633.
[26] Y. Hu, C. K. W. Yu, and X. Yang, Incremental quasi-subgradient methods for minimizing the sum of quasi-convex functions, Journal of Global Optimization, 75 (2019), pp. 1003–1028.
[27] H. Iiduka and K. Hishinuma, Acceleration method combining broadcast and incremental distributed optimization algorithms, SIAM Journal on optimization, 24 (2014), pp. 1840–1863.
[28] B. Johansson, M. Rabi, and M. Johansson, A randomized incremental subgradient method for distributed optimization in networked systems, SIAM Journal on Optimization, 20 (2010), pp. 1157–1170.
[29] V. Kharin, Decomposition into functions in the minimization problem, Avtomatika i Telemekhanika, (1979), pp. 66–79.
[30] K. C. Kiwiel, Convergence of approximate and incremental subgradient methods for convex optimization, SIAM Journal on Optimization, 14 (2004), pp. 807–840.
[31] J. Li, A. M.-C. So, and W.-K. Ma, Understanding notions of stationarity in nonsmooth optimization, IEEE Signal Processing Magazine, 37 (2020), pp. 18–31.
[32] X. Li, Z. Zhu, A. M.-C. So, and J. D. Lee, Incremental methods for weakly convex optimization, arXiv preprint arXiv:1907.11687, (2019).
[33] X. Li, Z. Zhu, A. M.-C. So, and R. Vidal, Nonconvex robust low-rank matrix recovery, SIAM Journal on Optimization, 30 (2020), pp. 660–686.
[34] Z.-Q. Luo and P. Tseng, Analysis of an approximate gradient projection method with applications to the backpropagation algorithm, Optimization Methods and Software, 4 (1994),
[35] O. Mangasarian and M. Solodov, Serial and parallel backpropagation convergence via nonmonotone perturbed minimization, Optimization Methods and Software, (1994).

[36] A. Mokhtari, M. Gürbüzbalaban, and A. Ribeiro, Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate, SIAM Journal on Optimization, 28 (2018), pp. 1420–1447.

[37] D. Nagaraj, P. Jain, and P. Netrapalli, SGD without replacement: Sharper rates for general smooth convex functions, in International Conference on Machine Learning, 2019, pp. 4703–4711.

[38] A. Nedić and D. Bertsekas, Convergence rate of incremental subgradient algorithms, in Stochastic Optimization: Algorithms and Applications, Springer, 2001, pp. 223–264.

[39] A. Nedić and D. P. Bertsekas, Incremental subgradient methods for nondifferentiable optimization, SIAM Journal on Optimization, 12 (2001), pp. 109–138.

[40] A. Nedić and D. P. Bertsekas, The effect of deterministic noise in subgradient methods, Mathematical programming, 125 (2010), pp. 75–99.

[41] A. Nedić, D. P. Bertsekas, and V. S. Borkar, Distributed asynchronous incremental subgradient methods, Studies in Computational Mathematics, 8 (2001), pp. 381–407.

[42] A. Nedić and A. Ozdaglar, Distributed subgradient methods for multi-agent optimization, IEEE Transactions on Automatic Control, 54 (2009), pp. 48–61.

[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on optimization, 19 (2009), pp. 1574–1609.

[44] Yu. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, vol. 87, Springer Science & Business Media, 2013.

[45] E. S. H. Neto and A. R. De Pierro, Incremental subgradients for constrained convex optimization: A unified framework and new methods, SIAM Journal on Optimization, 20 (2010), pp. 1547–1572.

[46] M. Rabbat and R. Nowak, Distributed optimization in sensor networks, in Proceedings of the 3rd international symposium on Information processing in sensor networks, ACM, 2004, pp. 20–27.

[47] M. G. Rabbat and R. D. Nowak, Quantized incremental algorithms for distributed optimization, IEEE Journal on Selected Areas in Communications, 23 (2005), pp. 798–808.

[48] S. S. Ram, A. Nedić, and V. V. Veeravalli, Incremental stochastic subgradient algorithms for convex optimization, SIAM Journal on Optimization, 20 (2009), pp. 691–717.

[49] R. T. Rockafellar and R. J.-B. Wets, Variational Analysis, vol. 317, Springer Science & Business Media, 2009.

[50] O. Shamir, Without-replacement sampling for stochastic gradient methods, in Advances in Neural Information Processing Systems, 2016, pp. 46–54.

[51] N. Z. Shor, Minimization Methods for Non-Differentiable Functions, Springer-Verlag, Berlin, Heidelberg, 1985.

[52] M. V. Solodov, Incremental gradient algorithms with stepsizes bounded away from zero, Computational Optimization and Applications, 11 (1998), pp. 23–35.

[53] M. V. Solodov and S.扎洛夫, Error stability properties of generalized gradient-type algorithms, Journal of Optimization Theory and Applications, 98 (1998), pp. 663–680.

[54] P. Tseng, An incremental gradient(1-projection) method with momentum term and adaptive stepsize rule, SIAM Journal on Optimization, 8 (1998), pp. 506–531.

[55] J.-P. Vial, Strong and weak convexity of sets and functions, Mathematics of Operations Research, 8 (1983), pp. 231–259.

[56] M. Wang and D. P. Bertsekas, Incremental constraint projection methods for variational inequalities, Mathematical Programming, 150 (2015), pp. 321–363.

[57] M. Wang and D. P. Bertsekas, Stochastic first-order methods with random constraint projection, SIAM Journal on Optimization, 26 (2016), pp. 681–717.

[58] B. Widrow and M. E. Hoff, Adaptive switching circuits, tech. report, Stanford Univ Ca Stanford Electronics Labs, 1960.