A Projection-free Incremental Gradient Method for Large-scale Constrained Optimization

Harshal D. Kaushik∗ and Farazd Yousefian †
Department of Industrial Engineering and Management
Oklahoma State University
Stillwater, OK 74078

Abstract

The problem of minimizing finite sums where each component function is associated with a block of dataset, is very popular in machine learning. Among well-known avenues to address this class of problems is the incremental gradient (IG) methods where a single component function is selected at each iteration in a cyclic (or randomized) manner. When the problem is constrained, the existing IG schemes often require a projection step onto the feasible set at each iteration. This step is computationally costly, in particular, in large-scale applications. To the best of our knowledge, the only variant of the IG method that does not require a projection-type step is the incremental augmented Lagrangian method which can only address the separable equality-constrained optimization problems. Motivated by this gap, we consider the problem of minimizing a strongly convex and smooth finite sum function, subject to nonseparable linear constraints, associated with each component function. By introducing a novel regularization-based relaxation technique, we develop an iteratively regularized IG algorithm that requires no projections. With a proper choice of update rules for the stepsize and the regularization parameters, we derive a sublinear rate of convergence for the objective function and establish an asymptotic convergence for the infeasibility of the generated iterates. We compare the performance of the proposed scheme with that of the classical projected IG schemes and SAGA. The effectiveness of our method is validated by increasing both the dimensionality of the solution space and the number of constraints.

1 Introduction

We consider a large-scale linearly constrained finite sum optimization problem as the following:

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} & \quad f(x) \triangleq \sum_{i=1}^{m} f_i(x) \\
\text{subject to} & \quad A_i x = b_i \quad \text{for all } i \in \{1, \ldots, m\},
\end{align*}
\]

where the component functions \( f_i : \mathbb{R}^n \to \mathbb{R} \) are convex, \( x \in \mathbb{R}^n \) is the (global) decision variable, \( A_i \in \mathbb{R}^{d_i \times n} \) and \( b_i \in \mathbb{R}^{d_i} \) are given parameters for all \( i \), and parameters \( n, m \) and \( p \triangleq \sum_{i=1}^{m} d_i \) are possibly large. Problem (P) arises in a breadth of applications including expected loss minimization in statistical learning [21] where \( f_i \) is associated with a data block, as well as distributed optimization...
One of the popular methods in addressing finite sum problems, in particular in the unconstrained regimes, is the class of incremental gradient (IG) methods where utilizing the additive structure of the problem, the algorithm cycles through the data blocks and updates the local estimates of the optimal solution in a sequential manner [7]. While the first variants of IG schemes find their roots in addressing neural networks as early as in 80s’ [5], the complexity analysis of these schemes has been a trending research topic in the optimization and machine learning communities in the past two decades. In addressing constrained problems, the projected incremental gradient (P-IG) method and its subgradient variant were developed [16]. In the smooth case, it is described as follows: given an initial point \( x_{0,1} \in X \), where \( X \subseteq \mathbb{R}^n \) denotes the constraint set, for each \( k \geq 1 \), consider the following update rule:

\[
\begin{align*}
    x_{k,i+1} &= P_X (x_{k,i} - \gamma_k \nabla f_i (x_{k,i})) \\
    x_{k+1} &:= x_{k,m+1}
\end{align*}
\]

where \( P \) denotes the Euclidean projection operator and is defined as \( P_X (z) \triangleq \arg\min_{x \in X} \| x - z \|_2 \) and \( \gamma_k > 0 \) is the stepsize parameter. Recently, under the assumption of strong convexity and twice continuous differentiability of the objective function, the standard IG method was proved to converge with the rate \( O(1/k) \) [11] in the unconstrained case. This indeed was an important improvement to the prior known rate of \( O(1/\sqrt{k}) \) for the merely convex case. Accelerated variants of IG schemes with provable convergence speeds were developed more recently, including the incremental aggregated gradient method (IAG) [8, 10, 21, and SAGA [9]. While addressing the non-strongly convex case, SAGA with averaging achieves a sublinear convergence rate, assuming strong convexity and smoothness, this is improved for non-averaging variants of SAGA and IAG to a linear convergence rate. Despite the faster rates of convergence in comparison with the standard IG method, the aforementioned methods require an excessive memory of \( O(mn) \) which limits their applications in the large-scale settings. Another existing challenge in the implementation of these scheme lies in addressing constraints. Contending with the presence of constraints, projected, and more generally, proximal variants of the aforementioned IG schemes have been developed. However, the projection operation may become costly in cases where either the number of component functions, i.e., \( m \), the number of constraints, or the dimensionality of the solution space, i.e., \( n \), is large. Recent works are [14, 23, 25, 17] Another avenue in addressing constrained optimization problems lies in Lagrangian duality. Despite the wide-spread application of the theory of duality and Lagrangian relaxation in addressing constrained problems in centralized regimes, there have been a handful of papers in the area of distributed optimization that can cope with large-scale constraints (see [6, 2, 13] and the references therein). Of these, an incremental Augmented Lagrangian method was developed in [6] for solving finite sum problems with separable linear constraints. In contrast with this scheme and also ADMM-related approaches, our goal in this paper lies in addressing the distributed constrained problem [7] under a nonseparable structure where the constraints for each component function \( i \) are associated with the same variable \( x \).

**Main contributions.** We develop an iteratively regularized incremental gradient algorithm where at each iteration, a linear function characterized by the problem parameters is regularized by the gradient mapping of the objective function. Importantly, this scheme does not require any projections
and is suitable for addressing large-scale problems. Through employing a novel regularization-based relaxation technique, with a careful choice of update rules for the stepsize and the regularization parameters, we derive a sublinear rate of convergence for the objective function and establish an asymptotic convergence for the infeasibility of the generated iterates. We compare the performance of the proposed scheme with that of the classical projected IG scheme and SAGA. We validate the effectiveness of the proposed scheme in cases where both the dimensionality of the solution space and the number of constraints are large.

Outline. The remainder of the paper is organized as follows. Section 2 introduces the algorithm outline for addressing problem (P). We also provide the main assumptions needed for the convergence analysis. Section 3 includes the convergence analysis of the proposed scheme. Section 4 contains the numerical implementation where we compare the proposed algorithm with IG schemes such as SAGA and IAG.

Notation and preliminary definitions. A function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is said to be in the class \( C^{k,r}_{\mu,L} \) if \( f \) is \( \mu \)-strongly convex in \( \mathbb{R}^n \), \( k \) times continuously differentiable, and its \( r \)-th derivative is Lipschitz continuous with constant \( L \). For any symmetric square matrix \( B \in \mathbb{R}^{n \times n} \), the spectral norm is denoted by \( \| B \| \) and is defined as the maximum absolute value of eigenvalues of the matrix, i.e., we have \( \| B \| = \max \{|\lambda_{\text{min}}(B)|, |\lambda_{\text{max}}(B)|\} \). Note that, for a positive semidefinite matrix \( B \), we have \( \| B \| = \lambda_{\text{max}}(B) \). For any vector \( x \in \mathbb{R}^n \), we use \( \| x \| \) to denote the \( \ell_2 \)-norm. For a Lipschitz continuous function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) with Lipschitz parameter \( L_f \), we have for any \( L \geq L_f \), \( f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{L}{2}\| x - y \|^2 \) for all \( x, y \in \mathbb{R}^n \). A continuously differentiable function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is \( \mu \)-strongly convex if we have \( f(x) \geq f(y) + \nabla f(y)^T(x - y) + \frac{\mu}{2}\| x - y \|^2 \) for all \( x, y \in \mathbb{R}^n \). For problem (P), we define the matrix \( A \in \mathbb{R}^{p \times n} \) as \( A \triangleq (A_1^T, A_2^T, \ldots, A_m^T)^T \) and the vector \( b \in \mathbb{R}^{p \times 1} \) as \( b \triangleq (b_1^T, b_2^T, \ldots, b_m^T)^T \).

2 Algorithm outline

In this section, we present the proposed algorithm. The general outline of the scheme is shown in Algorithm 1. The main step includes the update \( x_{k,i+1} := x_{k,i} - \gamma_k (A_i^T(A_i x_{k,i} - b_i) + \eta_k \nabla f_i(x_{k,i})) \) where \( \gamma_k > 0 \) and \( \eta_k \) denote the stepsize and regularization parameters. We note that similar to the standard IG scheme, Algorithm 1 is a distributed method both in terms of the objective and constraints in the sense that at a time and for each particular \( i \), only the information of \( A_i, b_i \), and \( f_i \) are used in the update rule. The underlying idea is that the regularization technique serves as a relaxation rule in that the infeasibility of the method is regularized by the the objective function value. Importantly, we allow both of these parameters to be updated iteratively. The main research question lies in deriving suitable update rules for tuning these sequences to establish the convergence and obtain rate statements.

Algorithm 1 Iterative Regularized Incremental Gradient (IR-IG)

1: **Initialization**: Select a point \( x_0 \in \mathbb{R}^n \)
2: for \( k = 0, 1, \ldots, N - 1 \) do
3: \( x_{k,1} := x_k \) and select \( \gamma_k > 0, \eta_k > 0 \)
4: for \( i = 1, \ldots, m \) do
5: \( x_{k,i+1} := x_{k,i} - \gamma_k (A_i^T(A_i x_{k,i} - b_i) + \eta_k \nabla f_i(x_{k,i})) \)
6: end for
7: end for
8: Set \( x_{k+1} \triangleq x_{k,m+1} \)

Next, we state the assumptions on problem (P) that will be employed in the convergence analysis.

**Assumption 2.1** (Properties of problem (P)). Let the following hold:
(a) The function \( f \) is twice differentiable strongly convex over \( \mathbb{R}^n \) with parameter \( \mu_f > 0 \).
(b) The feasible set of problem (P) is nonempty.

Provided that Assumption 2.1 holds, the objective of problem (P) is a strongly convex function over a convex closed set. As such, problem (P) has a unique solution. Throughout, the unique solution to
problem $(P)$ is denoted as $x^*$. Next, we define a regularized problem that has an important role in the convergence analysis of our work.

\textbf{Definition 1 (The regularized problem).} For a regularization parameter $\eta_k > 0$ at iteration $k$, the regularized problem is defined as:

$$\min_{x \in \mathbb{R}^n} f_{\eta_k}(x) = \frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 + \eta_k \sum_{i=1}^{m} f_i(x).$$

$(P_{\eta_k})$

Note that under Assumption 2.1, problem $(P_{\eta_k})$ has also a unique solution denoted by $x^*_{\eta_k}$.

Next, we make an assumption on the boundedness of the iterates from Algorithm 1. This will be useful later in obtaining the upper bound on the gradients. This assumption has been employed in the literature on the convergence of IG schemes, for example in [4],[12], and more recently in [11].

\textbf{Assumption 2.2 (Boundedness of iterates).} There exists a nonempty compact Euclidean ball $X_1 \subset \mathbb{R}^n$ such that $X_1$ contains all the iterates of Algorithm 1.

Later in Lemma 3.1, it will be shown that $x^*_{\eta_k}$ converges to $x^*$ when $\eta_k \to 0$. As such, there exits a ball $X_0 \subset \mathbb{R}^n$ containing the solutions $x^*_{\eta_k}$. Throughout, we denote the convex hull of $X_1$ and $X_0$ by a compact set $X(\subset \mathbb{R}^n)$.

\textbf{Remark 2.1.} Under Assumptions 2.1 and 2.2, the continuous gradient map of $f$ over a compact set $X$ is bounded. Therefore, there exits scalars $C$ and $C_f$ such that: $\max_{1 \leq i \leq m} \sup_{x \in X} \|A_i^T (A_i x - b_i)\| \leq C,$ and $\max_{1 \leq i \leq m} \sup_{x \in X} \|\nabla f_i(x)\| \leq C_f$. Throughout, we denote $C_F \triangleq m C_f$.

\textbf{Remark 2.2.} As a consequence of Assumptions 2.1 and 2.2, the gradients of component functions $f_i$ are Lipschitz continuous (cf. Chapter 8 in [3]). That is, we have $L_{f_i} \triangleq \max_{z \in X} \|\nabla f_i(z)\|$ such that $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_{f_i} \|x-y\|$ for all $x, y \in X$, and $i = 1, 2, \ldots, m$. Similarly, for $\|A_i x - b_i\|^2$ and $i \in \{1, 2, \ldots, m\}$, we can write:

$$\|A_i^T (A_i x - b_i) - A_i^T (A_i y - b_i)\| \leq \|A_i^T A_i\| \|x-y\| \quad \text{for all} \ x, y \in X. \quad (1)$$

Next, from the definition of spectral norm, we have: $\|A_i^T A_i\| = \lambda_{\max}(A_i^T A_i)$. Thus, we obtain $\|A_i^T (A_i x - b_i) - A_i^T (A_i y - b_i)\| \leq \lambda_{\max}(A_i^T A_i) \|x-y\|$ for all $x, y \in X$, implying that the function $\|A_i x - b_i\|^2$ has a Lipschitz gradient with parameter $L_i \triangleq \lambda_{\max}(A_i^T A_i)$.

3 Convergence analysis

Here, we provide the convergence analysis and derive the convergence rate statements for Algorithm 1. We begin this section with Lemma 3.1 where we establish the convergence of the Tikhonov trajectory to the optimal solution of problem $(P)$. We use Lemma 3.1 later to derive the main rate result of our work presented in Theorem 3.1.

\textbf{Lemma 3.1 (The properties of Tikhonov trajectory).} Consider problem $(P_{\eta_k})$. Let Assumptions 2.1 and 2.2 hold. Let $\eta_k$ be a nonincreasing sequence such that $\lim_{k \to \infty} \eta_k = 0$. Then:

(a) The limit point of the Tikhonov trajectory $\{x^*_{\eta_k}\}$ when $k$ goes to infinity exists and is equal to $x^*$.

(b) For some $C_F > 0$ and all $k \geq 1$, we have: $\left\|x^*_{\eta_{k+1}} - x^*_{\eta_k}\right\| \leq \frac{C_F}{\mu_f} \left|1 - \frac{\eta_{k+1}}{\eta_k}\right|.$

\textbf{Proof.} (a) Consider problem $(P_{\eta_k})$. From Definition 1 $x^*_{\eta_k}$ is a unique solution to problem $(P_{\eta_k})$. Then for any $x \in \mathbb{R}^n$ at iteration $k$, we have: $f_{\eta_k}(x^*_{\eta_k}) \leq f_{\eta_k}(x)$. Writing this for a feasible $x$, we have:

$$f_{\eta_k}(x^*_{\eta_k}) \leq \eta_k f(x) + \frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2$$

for all $k$. Then, we have:

$$\left\|x^*_{\eta_{k+1}} - x^*_{\eta_k}\right\| \leq \frac{C_F}{\mu_f} \left|1 - \frac{\eta_{k+1}}{\eta_k}\right|.$$
from feasibility of \( x \), term 1 is zero, therefore, we have: 
\[ f_{\eta_k} \left( x_{\eta_k}^* \right) \leq \eta_k f(x). \]
Substituting for \( f_{\eta_k} \),
\[
\eta_k f \left( x_{\eta_k}^* \right) + \frac{1}{2} \sum_{i=1}^{m} \left\| A_i x_{\eta_k}^* - b_i \right\|^2 \leq \eta_k f(x).
\]
Note that term 2 is nonnegative, therefore, \( \eta_k f \left( x_{\eta_k}^* \right) \leq \eta_k f(x) \). Next, taking into account the nonnegativity of \( \eta_k \), we have:
\[
f \left( x_{\eta_k}^* \right) \leq f(x).
\]
(2)

From Assumption 2.2, the sequence of the solution set \( \{ x_{\eta_k}^* \} \) is bounded. Using Bolzano Weierstrass theorem, we must have a converging subsequence \( \{ x_{\eta_{k_j}}^* \} \) to limit point \( \hat{x}^* \). Next, taking into account the continuity of \( f \), we can write the following limits from equation (2),
\[
f(x) \geq f \left( \lim_{k \to \infty} \{ x_{\eta_{k_j}}^* \} \right) = f(\hat{x}^*).
\]
(3)

Note that above equation is true for any \( x \) such that \( Ax = b \). The solution to problem (P), \( x^* \) in fact must hold the feasibility. Therefore, from equation (3), we have:
\[
f(x^*) \geq f(\hat{x}^*).
\]
(4)

Next, consider any other subsequence \( \{ x_{\eta_{k_j}}^* \} \) of bounded sequence \( \{ x_{\eta_k}^* \} \), converging to limit point \( \hat{x}^* \). Now we know that \( \{ x_{\eta_{k_j}}^* \} \) is the solution to regularized problem \( P_{\eta_{k_j}} \). It must satisfy the optimality condition,
\[
\sum_{i=1}^{m} A_i^T \left( A_i x_{\eta_{k_j}}^* - b_i \right) + \eta_{k_j} \nabla f \left( x_{\eta_{k_j}}^* \right) = 0.
\]
Taking limits and noting \( \lim (\eta_{k_j}) \to 0 \), we have:
\[
\sum_{i=1}^{m} A_i^T \left( A_i \hat{x}^* - b_i \right) = 0.
\]
(5)

From above, we can conclude that the limit point of \( \{ x_{\eta_{k_j}}^* \} \) is feasible.

Similar sets of arguments that we used for obtaining (4), can be used again and equation (4) can be obtained for “any” limit point of the sequence \( \{ x_{\eta_{k_j}}^* \} \).

Now we know that solution to problem (P) is unique and equations (4) and (5) can be written for “any” limit point of sequence \( \{ x_{\eta_k}^* \} \). Therefore, we can say that all the subsequences \( \{ x_{\eta_{k_j}}^* \} \) converge to the same limit point: \( x^* \).

(b) Consider problem \( P_{\eta_1} \). From the convexity of \( \frac{1}{2} \sum_{i=1}^{m} \| A_i x - b_i \|^2 \) and strong convexity of \( \sum_{i=1}^{m} f_i(x) \), for any consecutive iterations \( k \) and \( k - 1 \), we can easily obtain the following:
\[
\sum_{i=1}^{m} A_i^T A_i \left( x_{\eta_{k-1}}^* - x_{\eta_k}^* \right)^T \left( x_{\eta_{k-1}}^* - x_{\eta_k}^* \right) \geq 0,
\]
(6)
\[
\sum_{i=1}^{m} \left( \nabla f_i \left( x_{\eta_{k-1}}^* \right) - \nabla f_i \left( x_{\eta_k}^* \right) \right)^T \left( x_{\eta_{k-1}}^* - x_{\eta_k}^* \right) \geq \mu_f \| x_{\eta_{k-1}}^* - x_{\eta_k}^* \|^2.
\]
(7)

Equations (6) and (7) can be easily verified as following:

Consider any two vectors \( x, y \in \mathbb{R}^n \). From convexity of \( \frac{1}{2} \sum_{i=1}^{m} \| A_i x - b_i \|^2 \), we have:
\[
\frac{1}{2} \sum_{i=1}^{m} \| A_i y - b_i \|^2 \geq \frac{1}{2} \sum_{i=1}^{m} \| A_i x - b_i \|^2 + \sum_{i=1}^{m} A_i^T (A_i x - b_i)^T (y - x),
\]
by exchanging the vectors $x$ and $y$ in the above expression, we have:

$$\frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 \geq \frac{1}{2} \sum_{i=1}^{m} \|A_i y - b_i\|^2 + \sum_{i=1}^{m} A_i^T (A_i y - b_i)^T (x - y).$$

Adding the above two expressions, we have:

$$\sum_{i=1}^{m} A_i^T A_i (x - y)^T (x - y) \geq 0,$$

substituting for $x := x_{n_{k-1}}^*$ and $y := x_{m_{k-1}}^*$ in the above equation, we obtain (6). In a similar way, using the strong convexity of $f$, we obtain (7) for any two vectors $x_{n_{k}}, x_{m_{k-1}} \in \mathbb{R}^n$.

From Definition [f] $f_{n_{k}}$ is a strongly convex function over $\mathbb{R}^n$. Therefore at iteration $k$, we have:

$$\sum_{i=1}^{m} A_i^T A_i (x_{n_{k}}^* - b_i) + \eta_k \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*) = 0. \tag{8}$$

Similarly at iteration $k - 1$, for the solution $x_{n_{k-1}}^*$, we have the following:

$$\sum_{i=1}^{m} A_i^T A_i (x_{n_{k-1}}^* - b_i) + \eta_{k-1} \sum_{i=1}^{m} \nabla f_i(x_{n_{k-1}}^*) = 0. \tag{9}$$

Next, multiplying by $(x_{n_{k-1}}^* - x_{n_{k}}^*)$ to equation (8) and $(x_{m_{k-1}}^* - x_{m_{k}}^*)$ to equation (9), we have:

$$\sum_{i=1}^{m} A_i^T A_i (x_{n_{k}}^* - x_{n_{k-1}}^*)^T (x_{n_{k-1}}^* - x_{n_{k}}^*)$$

$$+ \left( \eta_k \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*) - \eta_{k-1} \sum_{i=1}^{m} \nabla f_i(x_{n_{k-1}}^*) \right) (x_{n_{k-1}}^* - x_{n_{k}}^*) = 0.$$

From equation (6), term 1 is nonpositive. Therefore, we have

$$\left( \eta_k \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*) - \eta_{k-1} \sum_{i=1}^{m} \nabla f_i(x_{n_{k-1}}^*) \right) (x_{n_{k-1}}^* - x_{n_{k}}^*) \geq 0.$$

Adding and subtracting $\eta_k \sum_{i=1}^{m} \nabla f_i(x_{n_{k-1}}^*) (x_{n_{k-1}}^* - x_{n_{k}}^*)$, we have:

$$(\eta_k - \eta_{k-1}) \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*)^T (x_{n_{k}}^* - x_{n_{k-1}}^*) \geq \eta_k \left( \sum_{i=1}^{m} \nabla f_i(x_{n_{k-1}}^*) - \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*) \right) (x_{n_{k-1}}^* - x_{n_{k}}^*).$$

From equation (7), bounding term 2 in the above, we have:

$$(\eta_k - \eta_{k-1}) \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*)^T (x_{n_{k}}^* - x_{n_{k-1}}^*) \geq \eta_k \mu f \|x_{n_{k-1}}^* - x_{n_{k}}^*\|^2,$$

by squaring and adding all the components and using the Cauchy Schwartz inequality, we have:

$$\eta_k \sum_{i=1}^{m} \nabla f_i(x_{n_{k}}^*) \|x_{n_{k-1}}^* - x_{n_{k}}^*\| \geq \eta_k \mu f \|x_{n_{k-1}}^* - x_{n_{k}}^*\|^2,$$
now, from Remark 2.1, bounding term 3 and slight rearrangement, we have:

$$mC_f \frac{1 - \eta_{k-1}}{\eta_k} \geq \|x_{\eta_{k-1}}^* - x_{\eta_k}^*\|,$$

substituting $C_F \triangleq mC_f$, we have the result.

We make Assumption 3.1 on the update rules for the stepsize $\gamma_k$ and the regularization parameter $\eta_k$. We later give specific examples of the sequences that satisfy Assumption 3.1. This is an step towards answering the main research question that is finding suitable update rules for the stepsize and the regularization parameter to be used in order to derive the convergence rate statements. The answer will be presented in Theorem 3.1.

**Assumption 3.1** (Conditions on the update rules for the stepsize and the regularization parameter). Let $\{\gamma_k\}$ and $\{\eta_k\}$ be strictly positive and nonincreasing sequences. Let the following hold:

(a) There exist $\gamma_0$ and $\eta_0$ such that $\gamma_0 \eta_0 \mu_f < 1$, where $\mu_f$ is the strong convexity parameter $f$.

(b) There exist a scalar $B_1 > 0$ and an integer $k_1$ such that for $k \geq k_1$, $1 - \frac{\eta_{k-1}}{\eta_k} \leq B_1 \gamma_k^2$.

(c) There exist $\rho \in (0, 1)$ and an integer $k_2$ such that for $k \geq k_2$, $\frac{\gamma_{k-1}}{\eta_{k-1}} \leq \frac{\gamma_k}{\eta_k} (1 + \rho \gamma_k \eta_k \mu_f)$.

Using the specified assumptions on the sequences $\{\gamma_k\}$ and $\{\eta_k\}$, next we find an upper bound on the error that relates the generated sequence $\{x_k\}$ with the Tikhonov trajectory.

**Proposition 3.1** (Partial error bound w.r.t. Tikhonov trajectory). Consider problem (P). Let Assumptions 2.1, 2.2 and 3.1 hold. Let $x_k^\star$ denote the unique optimal solution to problem (P_{\eta_k}). Let $x_k$ be generated by Algorithm 1. Then, $\|x_{k+1} - x_k^\star\| \leq \tau \frac{B_2 \gamma_k}{\eta_k^2 \mu_f^2}$, where $\tau \triangleq \max \left\{ \frac{\|x_2 - x_1^\star\| \mu_f \eta_1}{B_2 \gamma_2^2}, \frac{1}{1 - \rho} \right\}$, $\rho \in (0, 1)$ and $B_2 > 0$.

**Proof.** Writing the update from Algorithm 1 in a compact form for iteration $k$,

$$x_{k+1} := x_k - \gamma_k \left( \sum_{i=1}^{m} \{ A_i^T (A_i x_{k,i} - b_i) + \eta_k \nabla f_i(x_{k,i}) \} \right), \text{ where } x_{k,1} := x_k \triangleq x_{k-1,m+1}. \tag{10}$$

Let us define the error term $e_k$ as follows:

$$e_k \triangleq \sum_{i=1}^{m} \{ A_i^T (A_i x_k - b_i) - A_i^T (A_i x_{k,i} - b_i) \} + \eta_k \sum_{i=1}^{m} \left( \nabla f_i(x_k) - \nabla f_i(x_{k,i}) \right). \tag{11}$$

From equations (10) and (11), we have:

$$x_{k+1} = x_k - \gamma_k \left( A^T (Ax_k - b) + \eta_k \nabla f(x_k) - e_k \right), \tag{12}$$

From Assumption 2.1, the twice differentiability of $f_i$ and $\frac{1}{2} \|A_i x - b_i\|^2$, we have:

$$\nabla f(x_k) = S_k^f (x_k - x_{\eta_k}^\star), \quad A^T (Ax_k - b) = S_k (x_k - x_{\eta_k}^\star),$$

where: $S_k^f \triangleq \int_0^{\eta_k} \nabla^2 f(x_{\tau k}^\star + \tau (x_k - x_{\eta_k}^\star)) d\tau$ and $S_k \triangleq A^T A$. Letting $H_k \triangleq S_k + \eta_k S_k^f$,

$$A^T (Ax_k - b) + \eta_k \nabla f(x_k) = H_k (x_k - x_{\eta_k}^\star),$$

Substituting in (12),

$$x_{k+1} - x_{\eta_k}^\star = (I_n - \gamma_k H_k) (x_k - x_{\eta_k}^\star) + \gamma_k e_k.$$

Squaring adding and using the triangle inequality,

$$\|x_{k+1} - x_{\eta_k}^\star\| \leq \| (I_n - \gamma_k H_k) \| \|x_k - x_{\eta_k}^\star\| + \gamma_k \|e_k\|. \tag{13}$$
Now let us first bound term 1. From the definition of $e_k$ and triangle inequality, we have:

$$
\|e_k\| \leq \sum_{i=1}^{m} \|A^T_i (A_i x_k - b_i) - A^T_i (A_i x_{k,i} - b_i)\| + \gamma_k \sum_{i=1}^{m} \|\nabla f_i (x_k) - \nabla f_i (x_{k,i})\|.
$$

From Remark 2.2 using the Lipschitz smoothness,

$$
\|e_k\| \leq \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \|x_{k,1} - x_{k,i}\|, \text{ where } L_i \triangleq \lambda_{\text{max}}(A^T_i A_i).
$$

Note that from Algorithm 1, we used $x_k = x_{k,1}$. Next, adding and subtracting $x_{k,j}$, we have:

$$
\|e_k\| = \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \|x_{k,1} - x_{k,j} + x_{k,j} - x_{k,i}\|, \text{ where } j \in \{2, 3, \ldots, i - 1\},
$$

and $x_{k,j} := 0$ for $i \leq 2$,

$$
\|e_k\| \leq \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) (\|x_{k,1} - x_{k,j}\| + \|x_{k,j} - x_{k,i}\|),
$$

$$
\|e_k\| = \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \sum_{j=1}^{i-1} \left(\|x_{k,j} - x_{k,j+1}\|\right). \quad (14)
$$

Substituting for term 2, from Algorithm 1, we have:

$$
\|e_k\| \leq \gamma_k \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \sum_{j=1}^{i-1} \|A^T_j (A_j x_{k,j} - b_j) + \eta_k \nabla f_j (x_{k,j})\|,
$$

using the triangle inequality, we have:

$$
\|e_k\| \leq \gamma_k \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \sum_{j=1}^{i-1} \|A^T_j (A_j x_{k,j} - b_j)\| + \gamma_k \eta_k \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \sum_{j=1}^{i-1} \|\nabla f_j (x_{k,j})\|.
$$

From Remark 2.1 bounding $\|A^T_j (A_j x_{k,j} - b_j)\|$ and $\|\nabla f_j (x_{k,j})\|$ in the above equation,

$$
\|e_k\| \leq \gamma_k \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \sum_{j=1}^{i-1} C + \gamma_k \eta_k \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) \sum_{j=1}^{i-1} C_f,
$$

$$
= \gamma_k C \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) (i - 1) + \gamma_k \eta_k C_f \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}) (i - 1),
$$

taking into account the nonnegativity of $L_i + \eta_k L_{f_i}$,

$$
\|e_k\| \leq m \gamma_k (C + \eta_k C_f) \sum_{i=1}^{m} (L_i + \eta_k L_{f_i}).
$$

Consider, $L \triangleq \sum_{i=1}^{m} L_{f_i}$, $L_f \triangleq \sum_{i=1}^{m} L_{f_i}$, and substitute for term 3, we have:

$$
\|e_k\| \leq m \gamma_k (C + \eta_k C_f) (L + \eta_k L_f).
$$

Substituting this bound for term 1 in equation (13), we have:

$$
\|x_{k+1} - x^*_{yk}\| \leq \left\|\sum_{i=1}^{m} (I_n - \eta_k H_k) (x_k - x^*_{yk})\right\| + m (\gamma_k)^2 (C + \eta_k C_f) (L + \eta_k L_f). \quad (15)
$$

Now first we bound term 4. We have, $H_k \triangleq S_k + \eta_k S_k^T$ where:

$$
S_k^T \triangleq \int_0^1 \nabla^2 f (x^*_{yk} + \tau (x_k - x^*_{yk})) \, d\tau \quad \text{and} \quad S_k \triangleq A^T A.
$$
From the definition of $H_k$, there exists some $\tau' \in [0, 1]$ such that:

$$H_k = A^T A + \int_0^1 \eta_k \nabla^2 f (x^*_\eta_k + \tau (x_k - x^*_\eta_k)) \, d\tau$$

$$= A^T A + \eta_k \nabla^2 f (x^*_\eta_k + \tau' (x_k - x^*_\eta_k)),$$

where $x_k, x^*_\eta_k \in \mathcal{X}$. Also note that from the convexity of $\mathcal{X}$, $x^*_\eta_k + \tau' (x_k - x^*_\eta_k) \in \mathcal{X}$.

Now from Remark 2.2, $L_i \triangleq \max_{z \in \mathcal{X}} \| \nabla^2 f_i (z) \|$ and $L_i \triangleq \lambda_{\max} (A_i^T A_i)$. Consider the following:

$$\sum_{i=1}^m L_i + \eta_k \sum_{i=1}^m L_i = \sum_{i=1}^m (\lambda_{\max} (A_i^T A_i)) + \eta_k \sum_{i=1}^m \max_{z \in \mathcal{X}} \| \nabla^2 f_i (z) \|$$

$$\geq \sum_{i=1}^m (\lambda_{\max} (A_i^T A_i)) + \eta_k \sum_{i=1}^m \| \nabla^2 f_i (x) \|,$$

using the triangle inequality and taking into account the continuity of derivatives,

$$\sum_{i=1}^m L_i + \eta_k \sum_{i=1}^m L_i \geq \sum_{i=1}^m (\lambda_{\max} (A_i^T A_i)) + \eta_k \left\| \nabla^2 \sum_{i=1}^m f_i (x) \right\| \quad \text{for any } x \in \mathcal{X},$$

$$= \sum_{i=1}^m \| A_i^T A_i \| + \eta_k \left\| \nabla^2 \sum_{i=1}^m f_i (x) \right\| \quad \text{for any } x \in \mathcal{X},$$

note that the second equality is obtained just from the definition of spectral norm. Now note that $A_i^T A_i$ is a symmetric matrix and we can further write above equation as,

$$\sum_{i=1}^m L_i + \eta_k \sum_{i=1}^m L_i \geq \| A \|^2 + \eta_k \left\| \nabla^2 f (x) \right\| \quad \text{for any } x \in \mathcal{X},$$

using the definitions of $L_f$ and $L$, we have:

$$L + \eta_k L_f \geq \| A \|^2 + \eta_k \left\| \nabla^2 f (x) \right\| \quad \text{for any } x \in \mathcal{X}. \quad (17)$$

Consider equation (16). By squaring adding on the both sides and bounding using the triangle inequality,

$$\| H_k \| \leq \| A^T A \| + \eta_k \left\| \nabla^2 f (x^*_\eta_k + \tau' (x_k - x^*_\eta_k)) \right\|,$$

from spectral norm definition and noting $z \triangleq x^*_\eta_k + \tau' (x_k - x^*_\eta_k) \in \mathcal{X}$, we have:

$$\| H_k \| \leq \| A \|^2 + \eta_k \left\| \nabla^2 f (z) \right\|. \quad (18)$$

From equations (17) and (18), for $z \in \mathcal{X}$, we have:

$$L + \eta_k L_f \geq \| H_k \|. \quad (19)$$

From Assumption 2.2 the Hessian $(\nabla^2 f (z) \text{ for all } z \in \mathcal{X})$ is symmetric. Exploiting the definition of spectral norm, we can say that,

$$H_k \preceq (L + \eta_k L_f) I_n. \quad (20)$$

Consider again equation (16), we have for $\tau' \in [0, 1]$ the following:

$$H_k = A^T A + \eta_k \nabla^2 f (x^*_\eta_k + \tau' (x_k - x^*_\eta_k)),$$

where $x_k, x^*_\eta_k \in \mathcal{X}$, and from the convexity of $\mathcal{X}$, we have: $x^*_\eta_k + \tau' (x_k - x^*_\eta_k) \triangleq z \in \mathcal{X}$. Now consider problem (P) again. From the definition of convexity and Assumption 2.2, for any $x \in \mathcal{X},$

$$\left\| \nabla^2 f (x) \right\| \geq \mu_f, \quad \lambda_{\min} (A^T A) \geq 0.$$

Now, from equation (21), we can write the following for any $z \triangleq x^*_\eta_k + \tau' (x_k - x^*_\eta_k) \in \mathcal{X}$, using the matrix norm,

$$\| H_k \| = \| A^T A + \eta_k \nabla^2 f (z) \| \geq \| A^T A \| + \eta_k \left\| \nabla^2 f (z) \right\| \geq \mu_f \eta_k. \quad (22)$$
Again, noting that Hessian is a symmetric matrix. Exploiting the definition of spectral norm, we have:

\[ \mu_f \eta_k I_n \preceq H_k \preceq (L + \eta_k L_f) I_n. \] (23)

Now using the positivity of the strong convexity parameter \( \mu_f \), bounding term 4 in equation (15),

\[ \|x_{k+1} - x^*_{\eta_k}\| \leq (1 - \gamma_k \eta_k \mu_f) \|x_k - x^*_{\eta_k}\| + m (\gamma_k)^2 (C + \eta_k C_f) (L + \eta_k L_f). \] (24)

From triangle inequality, term 5 can be written as,

\[ \|x_{k+1} - x^*_{\eta_k}\| \leq (1 - \gamma_k \mu_f \eta_k) \|x_k - x^*_{\eta_k-1}\| + (1 - \gamma_k \mu_f \eta_k) \|x^*_{\eta_k-1} - x^*_{\eta_k}\| \\
+ m (\gamma_k)^2 (C + \eta_k C_f) (L + \eta_k L_f), \]

\[ \leq (1 - \gamma_k \mu_f \eta_k) \|x_k - x^*_{\eta_k-1}\| + \|x^*_{\eta_k-1} - x^*_{\eta_k}\| \\
+ m (\gamma_k)^2 (C + \eta_k C_f) (L + \eta_k L_f). \]

Bounding term 6 in the above equation using Lemma [3.1] we have:

\[ \|x_{k+1} - x^*_{\eta_k}\| \leq (1 - \gamma_k \mu_f \eta_k) \|x_k - x^*_{\eta_k-1}\| + \frac{C_F}{\mu_f} \left| 1 - \frac{\eta_{k-1}}{\eta_k} \right| \\
+ m (\gamma_k)^2 (C + \eta_k C_f) (L + \eta_k L_f), \]

\[ \leq (1 - \gamma_k \mu_f \eta_k) \|x_k - x^*_{\eta_k-1}\| + \frac{C_F}{\mu_f} \left| 1 - \frac{\eta_{k-1}}{\eta_k} \right| \\
+ m (\gamma_k)^2 (C + \eta_k C_f) (L + \eta_k L_f), \]

from Assumption [3.1] (b), we have:

\[ \|x_{k+1} - x^*_{\eta_k}\| \leq (1 - \gamma_k \mu_f \eta_k) \|x_k - x^*_{\eta_k-1}\| \\
+ \gamma_k^2 \left( B_1 \frac{C_F}{\mu_f} + m (C + \eta_1 C_f) (L + \eta_1 L_f) \right). \]

Let \( B_2 \triangleq B_1 \frac{C_F}{\mu_f} + m (C + \eta_1 C_f) (L + \eta_1 L_f). \) We have for \( k \geq k_1: \)

\[ \|x_{k+1} - x^*_{\eta_k}\| \leq (1 - \gamma_k \mu_f \eta_k) \|x_k - x^*_{\eta_k-1}\| + B_2 \gamma_k^2. \] (25)

Now, let us define the sequences: \( v_{k+1} \triangleq \|x_{k+1} - x^*_{\eta_k}\|, \alpha_k \triangleq \gamma_k \mu_f \eta_k, \beta_k \triangleq B_2 \gamma_k^2, \) Equation 25 now becomes:

\[ v_{k+1} \leq (1 - \alpha_k) v_k + \beta_k. \] (26)

Now, let us consider the following conditions on the sequences \( \alpha_k, \beta_k, \) and \( v_k : \)

\[ \alpha_k \in (0, 1), \beta_k \geq 0, \text{ and } \frac{\beta_{k-1}}{\alpha_{k-1}} \leq \frac{\beta_k}{\alpha_k} (1 + \rho \eta \alpha_k), \]

where \( \rho \in (0, 1), \) then using the induction, we can, easily show for \( k \geq 1: \)

\[ v_{k+1} \leq \frac{\beta_k}{\alpha_k} \tau, \text{ where } \tau \triangleq \max \left\{ \frac{v_2 \alpha_1}{\beta_1}, \frac{1}{1 - \rho} \right\}. \] (27)

First verify the base case holds for \( k = 1. \) We have: \( \tau \geq \frac{v_2 \alpha_1}{\beta_1}, \) i.e., \( v_2 \leq \frac{\beta_1 \tau}{\alpha_1}. \)

Now consider for any integer \( k \geq 1, \) the following holds:

\[ v_k \leq \frac{\beta_{k-1}}{\alpha_{k-1}} \tau, \] (28)
we next show that this holds for $k+1$. From (26) and (28), $v_{k+1} \leq (1 - \alpha_k) \frac{\beta_k}{\alpha_k} \tau + \beta_k$. Substituting $\frac{\beta_k}{\alpha_k} \leq \frac{\beta_k}{\alpha_k} (1 + \rho \alpha_k)$, we have:

$$v_{k+1} \leq \tau (1 - \alpha_k) (1 + \rho \alpha_k) \frac{\beta_k}{\alpha_k} + \beta_k = \tau (1 - \alpha_k + \rho \alpha_k - \rho \alpha_k^2) \frac{\beta_k}{\alpha_k} + \beta_k.$$  

$$= \tau \frac{\beta_k}{\alpha_k} - (1 - \rho) \beta_k \tau + \frac{1}{\tau} \rho \beta_k \beta_k + \beta_k,$$

from the nonpositivity of term 7, we have $v_{k+1} \leq \tau \frac{\beta_k}{\alpha_k} + (1 - (1 - \rho)) \beta_k$,

now from the $\tau \geq \frac{1}{1 - \rho}$, we have: term 8 $\leq 0$. Bounding term 8, we have: $v_{k+1} \leq \tau \frac{\beta_k}{\alpha_k}$

In order to use the above result for bounding sequence $v_{k+1} \triangleq \|x_{k+1} - x^*\|$, we first need to verify all the conditions hold given in equation (27). We will do that next.

From Assumption 3.1 (a), we have $\gamma_k [\eta_k \mu_f] < 1$, from this, $0 < \eta_k < 1$.

Next, consider sequence definition for $\beta_k \triangleq B_2 \gamma_k^2$. From the definition of $\gamma_k$, $\eta_k$, positivity of $B_1$, and nonnegativity of the rest of the parameters (Lipschitz constants and bound on the norms of gradient), we have $B_2 \geq 0$.

Next, we show that third condition in equation (27) holds. From Assumption 3.1 (c), we have:

$$\gamma_k^{k-1} \eta_k \leq \gamma_k (1 + \rho \gamma_k \eta_k \mu_f),$$

equation (29) can also be written as:

$$\frac{\gamma_k^{k-1}}{\gamma_k \eta_k^{k-1}} \leq \frac{\gamma_k}{\gamma_k \eta_k} (1 + \rho \gamma_k \eta_k \mu_f).$$

Multiplying both sides by a nonnegative scalar $B_2$ and dividing $\mu_f$, we have:

$$B_2 \gamma_k^{k-1} \eta_k \leq B_2 \gamma_k^2 \eta_k \mu_f (1 + \rho \gamma_k \eta_k \mu_f).$$

Now, from the definitions of $\beta_k$ and $\alpha_k$,

$$\frac{\beta_k}{\alpha_k} \leq \beta_k (1 + \rho \gamma_k \eta_k \mu_f).$$

All the conditions of equation (27) are satisfied. Therefore, the inequality in equation (26) holds. We have: $v_{k+1} \leq \frac{\beta_k}{\alpha_k} \tau$, where $\tau \triangleq \max \left\{ \frac{\gamma_k}{\eta_k}, \frac{1}{1 - \rho} \right\}$. Substituting for the definitions of $\beta_k$, $\alpha_k$, and $v_k$, we get the result. □

Next, we will use the convergence and the recursive error bound for the Tikhonov trajectory (proved in Lemma 3.1) and the preceding result to derive an upper bound on the error of the objective function $f$ and a bound on the infeasibility of iterates.

**Proposition 3.2** (Suboptimality and infeasibility error bounds). Consider problem (P). Let Assumptions 2.1, 2.2, and 3.1 hold. Let $L_f$, $L_i$ denote the Lipschitz parameters given in Remark 2.2 and $\{x_k\}$ be generated by Algorithm 1. Then, there exists a scalar $M_f > 0$ such that we have:

(a) $f(x_{k+1}) - f(x^*) \leq \frac{B_2^{2}, \tau^2 (L + n \mu_f) \gamma_k}{2 \mu_f \eta_k}$,

(b) $\frac{1}{2} \sum_{i=1}^{m} \| A_i x_{k+1} - b_i \|^2 \leq \frac{B_2^{2}, \tau^2 (L + n \mu_f) \gamma_k}{2 \mu_f \eta_k}$ + $2 M_f \eta_k$,

where $\tau$, $\rho$, and $B_2$ are given in Proposition 3.1.

**Proof.** (a) Consider problem (P). From the Lipschitz continuity of the gradients of $f$, for $i \in \{1, 2, \ldots, m\}$, discussed in Remark 2.2, and defining $L_f \triangleq \sum_{i=1}^{m} L_{f_i}$, we have:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L_f}{2} \| y - x \|^2$$

for any $x, y \in \mathcal{X}$.
Similarly, defining \( L \triangleq \sum_{i=1}^{m} \lambda_{\max}(A_i^T A_i) \), for the feasibility metric, \( \frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 \), we have:

\[
\frac{1}{2} \sum_{i=1}^{m} \|A_i y - b_i\|^2 \leq \frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 + \sum_{i=1}^{m} (A_i^T (A_i x - b_i))^T (y - x) + \frac{L}{2} \|y - x\|^2
\]

for any \( x, y \in \mathcal{X} \). \( \text{(31)} \)

From equations (30) and (31), and the nonnegativity of \( \eta_k \), we have:

\[
\frac{1}{2} \sum_{i=1}^{m} \|A_i y - b_i\|^2 + \eta_k f(y) \leq \frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 + \eta_k f(x) + \frac{L + \eta_k L_f}{2} \|y - x\|^2
\]

\[
+ \sum_{i=1}^{m} (A_i^T (A_i x - b_i)^T + \eta_k f(x)) (y - x) \quad \text{for any } x, y \in \mathcal{X}.
\]

Now, consider the function \( f_{\eta_k} \) in Definition 1. We can write the above equation as:

\[
f_{\eta_k}(y) \leq f_{\eta_k}(x) + \nabla f_{\eta_k}(x)^T (y - x) + \frac{L + \eta_k L_f}{2} \|y - x\|^2 \quad \text{for any } x, y \in \mathcal{X}.
\]

Substituting \( x := x_{\eta_k}^* \) and \( y := x \) in the preceding relation and taking into account that, we have

\[
\nabla f_{\eta_k}(x_{\eta_k}) = 0,
\]

we have:

\[
f_{\eta_k}(x) - f_{\eta_k}(x_{\eta_k}) \leq \frac{L + \eta_k L_f}{2} \|x - x_{\eta_k}^*\|^2 \quad \text{for any } x \in \mathcal{X}.
\]

Now, we know \( x_{\eta_k}^* \) is the unique solution to problem \((P_{\eta_k})\), therefore, \( f_{\eta_k}(x_{\eta_k}) = f_{\eta_k}(x^*) \), where \( x^* \) is the optimal solution of problem \((P)\). Substituting this in the above equation, we have:

\[
f_{\eta_k}(x) - f_{\eta_k}(x^*) \leq \frac{L + \eta_k L_f}{2} \|x - x_{\eta_k}^*\|^2 \quad \text{for any } x \in \mathcal{X}.
\]

From the definition of \( f_{\eta_k}(\cdot) \) and using the nonincreasing property of \( \eta_k \), we have:

\[
\frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 - \frac{1}{2} \sum_{i=1}^{m} \|A_i x^* - b_i\|^2 + \eta_k (f(x) - f(x^*)) \leq \frac{L + \eta_k L_f}{2} \|x - x_{\eta_k}^*\|^2,
\]

\[
\leq \frac{L + \eta_k L_f}{2} \|x - x_{\eta_k}^*\|^2 \quad \text{for any } x \in \mathcal{X}.
\]

Note that term 1 is always nonnegative. From the optimality of \( x^* \) for problem \((P)\), it must be a feasible point to problem \((P_{\eta_k})\). Therefore, term 2 is zero. Thus, from equation (32), we obtain:

\[
f(x) - f(x^*) \leq \frac{L + \eta_k L_f}{2 \eta_k} \|x - x_{\eta_k}^*\|^2 \quad \text{for any } x \in \mathcal{X}.
\]

From Assumption 2.2 we have \( x_{k+1} \in \mathcal{X} \). Therefore, we can write:

\[
f(x_{k+1}) - f(x^*) \leq \frac{L + \eta_k L_f}{2 \eta_k} \|x_{k+1} - x_{\eta_k}^*\|^2 \leq \frac{B_2^2 r^2 (L + \eta_k L_f) \gamma_k^2}{2 \mu_f^2 \eta_k^3},
\]

where the in last inequality, we used the bound given in Proposition 3.1 Therefore, the desired relation holds.

(b) Using the compactness of the set \( \mathcal{X} \) and continuity of \( f \), there exists a scalar \( M_f > 0 \) such that \( f(x) \leq M_f \) for all \( x \in \mathcal{X} \) implying that \( |f(x) - f(x^*)| \leq 2M_f \). Thus, from (32) we have:

\[
\frac{1}{2} \sum_{i=1}^{m} \|A_i x - b_i\|^2 \leq \frac{L + \eta_k L_f}{2} \|x - x_{\eta_k}^*\|^2 + 2\eta_k M_f \quad \text{for any } x \in \mathcal{X}.
\]

From Assumption 2.2 we know \( x_{k+1} \in \mathcal{X} \). Therefore, we have,

\[
\frac{1}{2} \sum_{i=1}^{m} \|A_i x_{k+1} - b_i\|^2 \leq \frac{L + \eta_k L_f}{2} \|x_{k+1} - x_{\eta_k}^*\|^2 + 2\eta_k M_f.
\]

(33)
Next, bounding term 4 in the preceding relation using Proposition 3.1 we have:

\[ \frac{1}{2} \sum_{i=1}^{m} \| A_i x_{k+1} - b_i \|^2 \leq \frac{B^2 \tau^2 (L + \eta_0 L_f) \gamma_k^2}{2 \mu_f^2 \eta_k^2} + 2\eta_k M_f. \]

Therefore, the desired infeasibility bound holds as well.

Now, as mentioned earlier, following is the example of sequences for the stepsize (\( \gamma_k \)) and regularization parameter (\( \eta_k \)), which satisfy Assumption 3.1.

**Lemma 3.2** (Feasible sequences for Assumption 3.1). Let \( \gamma_k \) and \( \eta_k \) be the sequences generated using \( \gamma_k = \frac{\eta_0}{(1+k)^\alpha} \) and \( \eta_k = \frac{\eta_0}{(1+k)^\beta} \), where \( \gamma_0 \eta_0 < \frac{1}{\mu_f} \) and \( \eta_k \) satisfies Assumption 3.1.

Then, \( \gamma_k \) and \( \eta_k \) satisfy Assumption 3.1.

**Proof.** (a) Assumption 3.1(a) holds because \( \gamma_0, \eta_0, a, b \) are strictly positive and \( \gamma_0 \eta_0 < 1/\mu_f \).

(b) Consider the relation given in Assumption 3.1(b). Substituting for \( \gamma_k \) and \( \eta_k \), we obtain:

\[ \frac{1 - \frac{\eta_k - 1}{\eta_k}}{1 - \left( \frac{k+1}{k} \right)^b} = \left( 1 + \frac{b}{k} \right)^b - 1 = \frac{b(b-1)}{2} + \ldots \right) = \mathcal{O}(k^{-1}), \quad (34) \]

Comparing with (33) and taking to account that \( a < 0.5 \), we conclude that Assumption 3.1(b) holds.

(c) Assumption 3.1(c) can also be written as: \( \frac{n \gamma_k}{\eta_k} - 1 \leq \rho \gamma_k \eta_f \). Substituting the values \( \gamma_k \) and \( \eta_k \), noting that \( a, b > 0 \), we have:

\[ \frac{\eta_k \gamma_k - 1}{\gamma_k \eta_k - 1} = \left( \frac{k+1}{k} \right)^{a-b} - 1 = \left( 1 + \frac{a-b}{k} + \frac{(a-b)(a-b-1)}{2k^2} + \ldots \right) - 1 = \mathcal{O}(k^{-1}). \]

Next, from the values for the \( \gamma_k \) and \( \eta_k \), we have \( \rho \gamma_k \eta_f = \rho \gamma_0 \eta_0 \frac{1}{(1+k)^{\alpha+\beta}} \). Now, from equations (35), Assumption 3.1(c) is satisfied due to the condition \( a + b < 1 \).

The main result of our work is presented in Theorem 3.1 where we derive a sublinear convergence rate for the objective function value of (P) and establish an asymptotic convergence for the infeasibility of the iterates generated by Algorithm 1.

**Theorem 3.1** (Suboptimality and infeasibility rate results). Consider problem (P). Let Assumptions 2.1 and 3.1 hold and \( \{ x_k \} \) be the sequence generated by Algorithm 1. Let \( \gamma_k \) and \( \eta_k \) be the stepsize and regularization parameter sequences generated using \( \gamma_k = \frac{\eta_0}{(1+k)^\alpha} \) and \( \eta_k = \frac{\eta_0}{(1+k)^\beta} \), where \( \alpha := 0.5 - \frac{\gamma_0 \eta_0}{\mu_f} \) and \( b := \frac{\beta}{\gamma_0 \eta_0} \) are strictly positive and \( \gamma_0 \eta_0 < 1/\mu_f \). Then:

(a) The objective function \( f(x_{k+1}) \) converges to the optimal value \( f(x^*) \) with the rate \( \mathcal{O}(k^{-1+\epsilon}) \).

(b) The infeasibility term \( \frac{1}{2} \sum_{i=1}^{m} \| A_i x_{k+1} - b_i \|^2 \) converges to zero with the rate \( \mathcal{O}(k^{\epsilon/5}) \).

**Proof.** (a) Consider the bound on \( f \) from Proposition 3.2(a). Substituting the update rules of the stepsize \( \gamma_k \) and the regularization parameter \( \eta_k \), we have:

\[ f(x_{k+1}) - f(x^*) \leq \frac{ \gamma_0^2 B^2 \tau^2 (L + \eta_0 L_f) }{2 \eta_0^2 \mu_f^2 (1+k)^{2a-3\alpha}}. \]

Let us choose the sequence parameters \( a := 0.5 - \frac{\gamma_0 \eta_0}{\mu_f} \) and \( b := \frac{\beta}{\gamma_0 \eta_0} \), where \( \epsilon \) is an arbitrary small number. Substituting them in equation (36), we have:

\[ f(x_{k+1}) - f(x^*) \leq \frac{ \gamma_0^2 B^2 \tau^2 (L + \eta_0 L_f) }{2 \eta_0^2 \mu_f^2 (1+k)^{1-\epsilon}} = \mathcal{O}(k^{-1+\epsilon}) \).

(b) Substituting the update rules of the stepsize \( \gamma_k \) and the regularization parameter \( \eta_k \) in Proposition 3.2(b), we obtain:

\[ \frac{1}{2} \sum_{i=1}^{m} \| A_i x_{k+1} - b_i \|^2 \leq \frac{B^2 \tau^2 (L + \eta_0 L_f)}{2 \eta_0^2 \mu_f^2 (1+k)^{1-3\alpha/5}} + \frac{2M_f}{(1+k)^{\epsilon/5}} = \mathcal{O}(k^{\epsilon/5}). \]
4 Numerical results

In this section, we present the comparison between Algorithm 1 and IG schemes including standard projected IG, IAG, SAG, and SAGA. We consider minimizing the following function,

\[
f(x) := \frac{1}{2} \|x\|^2 + \theta \|x\|_1,
\]

over the constraints compactly written as \(Ax = b\), where \(\theta > 0\) and \(\|\cdot\|_1\) denotes the \(\ell_1\)-norm. Accordingly, the objective function of agent \(i\) is defined as,

\[
f_i(x) \triangleq \sum_{j=\frac{(i-1)n+1}{m}}^{\frac{in}{m}} \frac{\|x_j\|^2}{2} + \frac{\theta \|x\|_1}{m} \quad \text{for all} \quad i \in \{1, \ldots, m\}.
\]

We generate the parameters \(A \in \mathbb{R}^{p \times n}\) and vector \(b \in \mathbb{R}^p\) randomly. In the experiments, we chose \(\gamma_k = \frac{\gamma_0}{(1+k)^\eta} \) and \(\eta_k = \frac{\eta_0}{(1+k)^\gamma} \). We address the nondifferentiability of \(L_1\) norm by employing smoothing via the Moreau envelope. Accordingly, we consider the smoothed local functions \(f_i\), given by:

\[
f_i(x) \triangleq \sum_{j=\frac{(i-1)n+1}{m}}^{\frac{in}{m}} \frac{x_j^2}{2} + \theta \frac{M_{\lambda}\|x\|_1(x)}{m} \quad \text{where} \quad M_{\lambda}\|\cdot\|_1(x) = \begin{cases} \frac{x^2}{2} & \text{for} \quad \|x\|_1 \leq \lambda \\ \|x\|_1 - \frac{\lambda}{2} & \text{otherwise.} \end{cases}
\]

Recall that Algorithm 1 does not require any projection onto the feasible set. However, in other schemes including IG, IAG, SAG, and SAGA, a projection (more generally a proximal step) is needed. For evaluating projection, we use the Python solver \texttt{scipy.sparse.linalg.lsqr} [19,13]. This solver is effective in calculating the least norm solution to the linear system of equations \([22]\) and thus can address the Euclidean projection.

Figure 1 shows the performance of Algorithm 1, projected IG, IAG, SAG, and SAGA for the different choices of dimensionality \(n\) and the total number of linear constraints \(p\). Note that this comparison is done with respect to time. For this experiment, time was fixed to 20 seconds and the performance of each scheme is recorded.
Insights. (i) With increasing the dimension and the number of constraints, the projection evaluations take longer and consequently, the performance of the projected variant of the aforementioned IG schemes is deteriorated. This is the case in particular when $p = 5000$ and $p = 7000$. (ii) We observe that SAGA performs relatively well in addressing the higher dimensional problems. But when the number of the constraints increases, Algorithm 1 outperforms SAGA.

5 Concluding remarks

We consider the problem of minimizing the finite-sum where each component function is associated with a block of linear constraints. Our work is motivated by the computational challenges in projected incremental gradient schemes when the number of constraints or the dimension of the solution space is large. We develop an iteratively regularized incremental gradient scheme where we employ a novel regularization-based relaxation technique. The proposed algorithm is designed in a way that it does not require a projection step. We derive a sublinear rate of convergence for the objective function value and establish an asymptotic convergence for the infeasibility of the generated iterates. We compare the performance of the proposed scheme with the state-of-the-art incremental gradient schemes including projected IG, IAG, SAG, and SAGA. We observe that the proposed scheme outperforms the projected schemes as the number of constraints or the dimension of the solution space increases. A future direction to our research is extensions to the regimes where the constraints are more complicated due to the presence of nonlinearity.
Acknowledgments and Disclosure of Funding

Farzad Yousefian gratefully acknowledges the support of the National Science Foundation through CAREER grant ECCS-1944500.

References

[1] M. Amini and F. Yousefian. An iterative regularized incremental projected subgradient method for a class of bilevel optimization problems. American Control Conference (ACC), Philadelphia, PA, USA, pages 4069–4074, 2019.

[2] N. S. Aybat and E. Y. Hamedani. A primal-dual method for conic constrained distributed optimization problems. Advances in Neural Information Processing Systems, pages 5049–5057, 2016.

[3] A. Beck. First-Order Methods in Optimization. MOS-SIAM Series on Optimization, Philadelphia, PA, 2017.

[4] D. P. Bertsekas. Incremental least squares methods and the extended Kalman filter. SIAM Journal on Optimization, 6(3):807–822, 1996.

[5] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. arXiv:1507.01030v2, 2011.

[6] D. P. Bertsekas. Incremental aggregated proximal and augmented Lagrangian algorithms. arXiv:1509.09257, 2015.

[7] D. P. Bertsekas. Nonlinear Programming. Athena Scientific, Bellmont, MA, third edition, 2016.

[8] D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant stepsize. SIAM Journal on Optimization, 18(1):29–51, 2007.

[9] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Advances in Neural Information Processing Systems, pages 1646–1654, 2014.

[10] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. SIAM Journal on Optimization, 27(2):1035–1048, 2017.

[11] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Convergence rate of incremental gradient and incremental Newton methods. SIAM Journal on Optimization, 29(4):2542–2565, 2019.

[12] M. Gürbüzbalaban, A. Ozdaglar, and P.A. Parrilo. A globally convergent incremental Newton method. Mathematical Programming, 151:283–313, 2015.

[13] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm for general convex-concave saddle point problems. arXiv:1805.01401, 2019.

[14] H. Kaushik and F. Yousefian. A randomized block coordinate iterative regularized subgradient method for high-dimensional ill-posed convex optimization. American Control Conference (ACC), Philadelphia, PA, USA, pages 3420–3425, 2019.

[15] A. Nedić. Subgradient Methods for Convex Minimization. PhD thesis, MIT, 2002.

[16] A. Nedić and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. SIAM Journal on Optimization, 12(1):109–138, 2001.

[17] A. Nedić and T. Tatarenko. Convergence rate of a penalty method for strongly convex problems with linear constraints. arXiv:2004.13417v1, 2020.

[18] C. C. Paige and M. A. Saunders. Algorithm 583 LSQR: Sparse linear equations and least squares problems. ACM Transactions on Mathematical Software, 8(2):195–209, 1982.

[19] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Transactions on Mathematical Software, 8(1):43–71, 1982.
[20] M. Rabbat and R. D. Nowak. Distributed optimization in sensor networks. The International Conference on Information Processing in Sensor Networks, pages 20–27, 2004.

[21] N. L. Roux, M. Schmidt, and F.R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. Advances in Neural Information Processing Systems, pages 2663–2671, 2012.

[22] M. A. Saunders. Solution of sparse rectangular systems using LSQR and CRAIG. BIT Numerical Mathematics, 35:588–604, 1995.

[23] N. D. Vanli, M. Gürbüzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. SIAM Journal on Optimization, 28(2):1282–1300, 2018.

[24] F. Yousefian, A. Nedić, and U. V. Shanbhag. On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. Mathematical Programming, 165(1):391–431, 2017.

[25] F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic and deterministic quasi-Newton methods for nonstrongly convex optimization: Asymptotic convergence and rate analysis. SIAM Journal on Optimization, 30(2):1144–1172, 2020.