To appear in *Optimization Methods & Software*
Vol. 00, No. 00, Month 20XX, 1–35

Non–Asymptotic Convergence Analysis of Inexact Gradient Methods for Machine Learning Without Strong Convexity

Anthony Man–Cho So[a] and Zirui Zhou[b*]

[a]*Department of Systems Engineering and Engineering Management, and, by courtesy, CUHK–BGI Innovation Institute of Trans–omics, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong;* [b]*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong*

(*Received December 2015; Revised October 2016, February 2017*)

Many recent applications in machine learning and data fitting call for the algorithmic solution of structured smooth convex optimization problems. Although the gradient descent method is a natural choice for this task, it requires exact gradient computations and hence can be inefficient when the problem size is large or the gradient is difficult to evaluate. Therefore, there has been much interest in inexact gradient methods (IGMs), in which an efficiently computable approximate gradient is used to perform the update in each iteration. Currently, non–asymptotic linear convergence results for IGMs are typically established under the assumption that the objective function is strongly convex, which is not satisfied in many applications of interest; while linear convergence results that do not require the strong convexity assumption are usually asymptotic in nature. In this paper, we combine the best of these two types of results by developing a framework for analyzing the non–asymptotic convergence rates of IGMs when they are applied to a class of structured convex optimization problems that includes least squares regression and logistic regression. We then demonstrate the power of our framework by proving, in a unified manner, new linear convergence results for three recently proposed algorithms—the incremental gradient method with increasing sample size [7, 10], the stochastic variance–reduced gradient (SVRG) method [14], and the incremental aggregated gradient (IAG) method [5]. We believe that our techniques will find further applications in the non–asymptotic convergence analysis of other first–order methods.

Keywords: Non-asymptotic convergence rate, global error bound, inexact gradient method, least squares regression, logistic regression.

1. Introduction

Motivated by various applications in machine learning and data fitting, there has been much interest in the design and analysis of fast algorithms for solving large–scale structured convex optimization problems recently. A case in point is the problem of empirical risk minimization, in which one is given a set of input–output samples of a system, and the goal is to minimize the discrepancy between the observed output and the output predicted by certain parametrized model of the system. Such a problem can be formulated

*Corresponding author. Email: zrzhou@se.cuhk.edu.hk

as

$$\min_{x\in\mathbb{R}^d} \; f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where $x \in \mathbb{R}^d$ is the parameter vector, $f_i : \mathbb{R}^d \to \mathbb{R}$ is a convex function that measures the error or loss between the observed and predicted output of the i–th sample, and n is the total number of available samples. When every f_i is smooth, a simple and natural approach for solving Problem (1) is to use gradient descent. However, this requires the computation of the full gradient $\nabla f = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i$ in every iteration and hence can be expensive when M is large or some of the ∇f_i's are difficult to evaluate. Nevertheless, it is possible to circumvent such difficulty by exploiting the finite–sum structure of ∇f. One strategy is to use a subset of the summands that make up the full gradient ∇f to update the solution in each iteration. This leads to the class of incremental gradient methods, whose update formulae take the form

$$x^{k+1} = x^k - \frac{\alpha_k}{|I_k|}\sum_{i\in I_k} \nabla f_i(x^k). \tag{2}$$

Here, $\alpha_k > 0$ is the step size in the k–th iteration, and I_k is a (possibly random) subset of $\mathscr{N} = \{1, 2, \ldots, n\}$ chosen according to some pre–specified rules (see [4] and the references therein for some common choices of $\{I_k\}_{k\geq 0}$). Note that the k–th iteration only requires the $|I_k|$ gradient values $\{\nabla f_i(x^k)\}_{i\in I_k}$. Hence, an iteration of an incremental gradient method will generally be more efficient than that of gradient descent. However, in order to guarantee convergence, incremental gradient methods of the form (2) typically require diminishing step sizes, which results in the slow (sublinear) convergence of these methods [4]. On the other hand, gradient descent with a constant step size can achieve fast (linear) convergence in various settings (see, e.g., [24]). Thus, a problem of fundamental interest is to design methods that can enjoy both the low per–iteration complexity of incremental gradient methods and the fast convergence of gradient descent.

To approach the above problem, it is useful to consider incremental gradient methods of the form (2) under the framework of inexact gradient methods (IGMs). These methods aim at minimizing an arbitrary smooth function f by computing iterates $\{x^k\}_{k\geq 0}$ according to the formula

$$x^{k+1} = x^k - \alpha_k \left(\nabla f(x^k) + e^{k+1}\right), \tag{3}$$

where $G_k = \nabla f(x^k) + e^{k+1} \in \mathbb{R}^d$ is an approximation of the gradient ∇f at x^k, and $e^{k+1} = G_k - \nabla f(x^k) \in \mathbb{R}^d$ is the (possibly random) approximation error. It is easy to see that the update formula (2) is a special case of (3), with

$$e^{k+1} = \frac{1}{|I_k|}\sum_{i\in I_k} \nabla f_i(x^k) - \nabla f(x^k).$$

In fact, many other methods also fall under the IGM framework. For details, we refer the reader to [3, 4, 21, 27] and the discussions therein.

The rationale behind the update (3) is that an approximate gradient can often be computed very efficiently. Thus, IGMs could have significant computational gain in each iteration. However, the convergence rates of such methods depend crucially on the choice

of step sizes $\{\alpha_k\}_{k\geq 0}$ and the magnitude of the error vectors $\{e^k\}_{k\geq 1}$. Many recent works on the convergence analysis of IGMs have focused on the case where the step sizes are constant and developed conditions under which a *non–asymptotic* linear rate of convergence can be achieved. For instance, Blatt et al. [5] proposed an IGM, called the incremental aggregated gradient (IAG) method, for solving Problem (1) and showed that it converges linearly when f is a strongly convex quadratic function. Recently, Gürbüzbalaban et al. [12] refined this result by showing that the linear rate of convergence can still be attained when f is a smooth strongly convex function with Lipschitz continuous gradient. There has also been significant interest in stochastic IGMs for solving Problem (1) lately. These include the stochastic average gradient (SAG) method developed by Le Roux et al. [16], the stochastic variance–reduced gradient (SVRG) method developed by Johnson and Zhang [14], and the SAGA method developed by Defazio et al. [8], just to name a few. All these methods have been shown to converge linearly in expectation when f is strongly convex. It is interesting to note that the above results do not require diminishing step sizes or diminishing error norms. On another front, Byrd et al. [7] established the linear convergence of a certain instantiation of the incremental gradient method (2) when f is strongly convex and has a bounded Hessian. For the general IGM (3) with constant step sizes, Friedlander and Schmidt [10] (see also [27]) showed that it converges (sub)linearly if f is strongly convex and the squared error norms $\{\|e^k\|_2^2\}_{k\geq 1}$ decrease (sub)linearly to zero. It should be noted that all the aforementioned linear convergence results apply only to problems with a strongly convex objective. As such, they do not cover several important applications such as least squares regression and logistic regression. Although many works have studied the non–asymptotic convergence rates of IGMs when the objective function is not strongly convex, the best known rate is only sublinear (see, e.g., [1, 2, 6, 23, 25, 28]).

In another direction, there have been some early works that establish the *asymptotic* linear convergence of IGMs without requiring the objective function to be strongly convex. For instance, Luo and Tseng [20, 21] showed that if the error norms satisfy $\|e^{k+1}\|_2 = O(\|x^k - x^{k+1}\|_2)$, then the IGM (3) with step sizes bounded away from zero has an asymptotic linear rate of convergence when applied to certain structured convex optimization problems. In particular, this result applies to least squares regression and logistic regression. However, it should be noted that the condition on the error norms as stated above is rather strong, for it implies that the objective values of the iterates are strictly decreasing. Subsequently, Li [18] showed that the asymptotic linear convergence result of Luo and Tseng still holds under the weaker condition that the error norms decrease linearly to zero. This shows that even with large gradient approximation errors in the early iterations—which typically yields computational savings but may lead to an increase in the objective value in some iterations—the IGM can still converge quickly.

Motivated by the above discussion, our main contribution in this paper is twofold. First, we develop a new framework for analyzing the non–asymptotic convergence rate of the IGM (3) with step sizes bounded away from zero when it is applied to a class of structured convex optimization problems (which includes least squares regression and logistic regression). One notable feature of our proposed framework is that it can handle IGMs that generate iterates with non–monotonic objective values. Second, we show that our framework leads to a unified non–asymptotic linear convergence analysis of several recently proposed algorithms—namely, the incremental gradient method with increasing sample size [7, 10], the SVRG method [14], and the IAG method [5]—even when strong convexity is absent. Our linear convergence results extend those in [18, 20, 21, 30] in that they hold *non–asymptotically*, and those in [5, 7, 10, 12, 14, 27] in that they cover cases where the objective function is *not necessarily strongly convex*. A key step in our

approach is to develop a global version of the error bound in [20]. Such a global error bound provides a way to measure the progress of the IGM (3) in *every* iteration and not just those iterations that are close to convergence. This, together with the powerful convergence analysis framework developed by Luo and Tseng [21], allows us to establish the desired non–asymptotic convergence rate results. We remark that Wang and Lin [31] and Ma et al. [22] have recently exploited properties similar to the global error bound developed in this paper to study the non–asymptotic convergence rates of feasible descent methods. However, our work differs from theirs in two important aspects. First, the analyses in [22, 31] require the objective values of the iterates to be strictly decreasing either deterministically or in expectation, while our analysis does not have such a requirement. This allows us to analyze several recently proposed first–order methods that generate iterates with non–monotonic objective values, such as the SVRG method [14] and the IAG method [5]. Second, in the context of risk minimization, the analyses in [22, 31] apply only to the class of globally strongly convex loss functions, which precludes many commonly used loss functions such as the logistic loss $u \mapsto \log(1+\exp(-u))$. By contrast, our analysis only requires the loss function to be strongly convex on compact subsets, and hence it applies to a much wider class of loss functions.

2. Preliminaries

2.1 *Basic Setup and Observations*

In this paper, we focus on the following unconstrained convex optimization problem:

$$\min_{x\in\mathbb{R}^d} \ \left\{ f(x) = g(Ex) + q^T x \right\}, \tag{4}$$

where $E \in \mathbb{R}^{m\times d}$ is a linear operator, $q \in \mathbb{R}^d$ is a vector, and $g : \mathbb{R}^m \to \mathbb{R}$ is a function satisfying the following assumptions:

Assumption 1

(a) The function g is continuously differentiable on $\mathbb{R}^m$ and its gradient ∇g is Lipschitz continuous with parameter $L_g > 0$ on $\mathbb{R}^m$; i.e.,

$$\|\nabla g(u) - \nabla g(v)\|_2 \le L_g \|u - v\|_2 \quad \text{for } u, v \in \mathbb{R}^m.$$

(b) The function g is strictly convex on $\mathbb{R}^m$; i.e., g is strongly convex on any compact subset of $\mathbb{R}^m$.

The above setup is motivated by the empirical risk minimization problem (1). Indeed, in many applications, the prediction error of the i–sample f_i can be expressed as $f_i(x) = \ell(b_i, a_i^T x)$, where $(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$ is the i–th input–output sample, and $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function. Thus, the objective function f in Problem (1) can be put into the form $f(x) = g(Ex)$, where E is an $n \times d$ matrix whose i–th row is a_i^T, and $g : \mathbb{R}^n \to \mathbb{R}$ is given by

$$g(y) = \frac{1}{n} \sum_{i=1}^{n} \ell(b_i, y_i).$$

For instance, by taking ℓ to be the square loss $\ell(u, v) = (u - v)^2$, we obtain the least

squares regression problem

$$\min_{x\in\mathbb{R}^d} \; f(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(a_i^T x - b_i\right)^2}_{\ell(b_i, a_i^T x)}. \tag{5}$$

On the other hand, by using the logistic loss $\ell(u,v) = \log(1+\exp(-uv))$, we arrive at the logistic regression problem

$$\min_{x\in\mathbb{R}^d} \; f(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\log\left(1+\exp(-b_i a_i^T x)\right)}_{\ell(b_i, a_i^T x)}. \tag{6}$$

In both examples, it is easy to verify that the corresponding g satisfies Assumption 1.

Going back to Problem (4), we note that the strict convexity of g on $\mathbb{R}^m$ does not necessarily imply the strict convexity of f on $\mathbb{R}^d$, as E may not have full column rank. Now, by Assumption 1(a), it is easy to verify that ∇f is Lipschitz continuous with parameter $L = L_g\|E\|^2$ on $\mathbb{R}^d$, where $\|E\| = \sup_{\|x\|_2=1}\|Ex\|_2$ is the spectral norm of E. Finally, let $\mathcal{X}$ denote the set of optimal solutions to Problem (4). We make the following assumption concerning $\mathcal{X}$:

ASSUMPTION 2 *The optimal solution set $\mathcal{X}$ is non–empty.*

Assumption 2 implies that the optimal value $f_{\min}$ of Problem (4) is finite and bounded from below. This, together with Assumption 1(b), yields the following simple but useful result:

PROPOSITION 1 *The map $x \mapsto Ex$ is invariant over the optimal solution set $\mathcal{X}$; i.e., there exists a $t^* \in \mathbb{R}^m$ such that $Ex^* = t^*$ for all $x^* \in \mathcal{X}$.*

Proposition 1 is well known; see, e.g., [20, 29]. For the sake of completeness, let us include its proof here.

Proof. Let $x^*, y^* \in \mathcal{X}$ be arbitrary. By the convexity of $\mathcal{X}$, we have $(x^*+y^*)/2 \in \mathcal{X}$. Hence, the convexity of f and optimality of x^*, y^* imply that $f((x^*+y^*)/2) = (f(x^*) + f(y^*))/2$, or equivalently, $g((Ex^*+Ey^*)/2) = (g(Ex^*)+g(Ey^*))/2$. Since g is strictly convex on $\mathbb{R}^m$, we conclude that $Ex^* = Ey^*$, as desired. □

2.2 *Inexact Gradient Methods*

One approach for solving Problem (4) is to use inexact gradient methods (IGMs), which compute iterates according to the formula (3). Our goal is to establish the convergence rate of the IGM (3) under Assumptions 1 and 2 and various assumptions on the rate at which the error sequence $\{e^k\}_{k\geq 1}$ tends to zero. We allow for the possibility that $e^1, e^2, \ldots$ are random, in which case the iterates $x^1, x^2, \ldots$ will also be random. To simplify the exposition, we assume that the step sizes $\{\alpha_k\}_{k\geq 0}$ in (3) are constant and equal to some $\alpha > 0$. However, it should be noted that our analysis can also be applied to the case where the step sizes $\{\alpha_k\}_{k\geq 0}$ satisfy $\liminf_{k\geq 0}\alpha_k > 0$.

The first step of our convergence analysis is to understand the behavior of the (possibly random) sequence $\{f(x^k)\}_{k\geq 0}$. It is well–known that when there is no error (i.e., $e^k = \mathbf{0}$ for all $k \geq 1$), the IGM (3) will generate a sequence of iterates $\{x^k\}_{k\geq 0}$ whose associated

objective values $\{f(x^k)\}_{k\geq 0}$ are monotonically decreasing [17]. However, this may not be true in the presence of errors. The following proposition provides a bound on the difference of the objective values of two successive iterates in terms of the error size. Its proof is standard and can be found in Appendix A.

PROPOSITION 2 *The sequence $\{x^k\}_{k\geq 0}$ generated by the IGM (3) satisfies*

$$f(x^k) - f(x^{k+1}) \geq \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x^k - x^{k+1}\|_2^2 - \|e^{k+1}\|_2\|x^k - x^{k+1}\|_2$$

for all $k \geq 0$.

An immediate consequence of Proposition 2 is the following result, whose proof can be found in Appendix B.

COROLLARY 1 *(cf. [18]) Suppose that the step size α satisfies $\alpha \in \left(0, \frac{2}{L}\right)$. Then, for all $k \geq 0$,*

(a)

$$\|x^k - x^{k+1}\|_2^2 \leq \frac{2}{\gamma}\left(f(x^k) - f(x^{k+1}) + \frac{1}{2\gamma}\|e^{k+1}\|_2^2\right),$$

(b)

$$0 \leq f(x^{k+1}) - f_{\min} \leq f(x^k) - f_{\min} + \frac{1}{4\gamma}\|e^{k+1}\|_2^2,$$

where $\gamma = \frac{1}{\alpha} - \frac{L}{2} > 0$.

Although the error sequence $\{e^k\}_{k\geq 1}$ can be random, it should be noted that the inequalities in both Proposition 2 and Corollary 1 hold for every realization of $\{e^k\}_{k\geq 1}$.

3. Error Bound Condition

Since we are interested in analyzing the convergence rate of the IGM (3), we need a measure to quantify its progress towards optimality. One natural candidate would be the distance to the optimal solution set $\mathcal{X}$. Indeed, since $\mathcal{X}$ is non–empty, convex, and closed (the closedness of $\mathcal{X}$ follows from the continuity of g), every $x \in \mathbb{R}^d$ has a unique projection $\bar{x} \in \mathcal{X}$ onto $\mathcal{X}$, and hence the measure $x \mapsto \text{dist}(x, \mathcal{X})$, where

$$\text{dist}(x, \mathcal{X}) = \min_{y\in\mathcal{X}} \|x - y\|_2,$$

is well defined. Despite its intuitive appeal, the measure $\text{dist}(\cdot, \mathcal{X})$ has one major disadvantage; namely, it is not easy to compute. An alternative would be to consider the norm of the gradient $x \mapsto \|\nabla f(x)\|_2$, which is motivated by the fact that the optimality condition of (4) is $\nabla f(x) = \mathbf{0}$. However, since $\|\nabla f(\cdot)\|_2$ is only a surrogate of $\text{dist}(\cdot, \mathcal{X})$, we need to establish a relationship between them. Towards that end, consider the set

$$\mathcal{S}_B = \{y \in \mathbb{R}^m : \|y - t^*\|_2 \leq B\},$$

where $B > 0$ is arbitrary. We then have the following theorem:

THEOREM 1 *Suppose that both Assumptions 1 and 2 hold for Problem (4). Suppose further that g is strongly convex on $\mathcal{S}_B$ for some $B > 0$; i.e.,*

$$g(y) - g(z) \geq (y - z)^T \nabla g(z) + \frac{\sigma_B}{2}\|y - z\|_2^2 \quad \textit{for all } y, z \in \mathcal{S}_B.$$

Then, there exists a constant $\tau_B \geq \frac{1}{L}$ such that

$$\operatorname{dist}(x, \mathcal{X}) \leq \tau_B \|\nabla f(x)\|_2 \tag{7}$$

for all $x \in \mathbb{R}^d$ satisfying $Ex \in \mathcal{S}_B$.

Condition (7) is a so–called *error bound* for Problem (4). The proof of Theorem 1 relies on the following proposition, whose proof can be found in Appendix C.

PROPOSITION 3 *There exist an $\omega > 0$ such that for any $x \in \mathbb{R}^d$, there exists an $x^* \in \mathcal{X}$ satisfying*

$$\|x - x^*\|_2 \leq \omega \left(\|\nabla f(x)\|_2 + \|Ex - t^*\|_2\right). \tag{8}$$

Proof of Theorem 1. The argument is similar to that in [20]. Let $x \in \mathbb{R}^d$ be such that $Ex \in \mathcal{S}_B$. The strong convexity of g on $\mathcal{S}_B$ implies that

$$\begin{aligned}
\frac{\sigma_B}{2}\|Ex - t^*\|_2^2 &\leq g(Ex) - g(t^*) - (Ex - t^*)^T \nabla g(t^*), \\
\frac{\sigma_B}{2}\|Ex - t^*\|_2^2 &\leq g(t^*) - g(Ex) - (t^* - Ex)^T \nabla g(Ex).
\end{aligned}$$

Adding the above two inequalities and using the fact that $\nabla f(x^*) = \mathbf{0}$ yield

$$\begin{aligned}
\sigma_B \|Ex - t^*\|_2^2 &\leq (Ex - t^*)^T (\nabla g(Ex) - \nabla g(t^*)) \\
&= (x - x^*)^T (\nabla f(x) - \nabla f(x^*)) \\
&\leq \|x - x^*\|_2 \|\nabla f(x)\|_2.
\end{aligned} \tag{9}$$

In addition, by Proposition 3, there exists an $x^* \in \mathcal{X}$ such that (8) holds. Hence, using (8), (9), and the identity $(a + b)^2 \leq 2(a^2 + b^2)$, which is valid for all $a, b \in \mathbb{R}$, we compute

$$\begin{aligned}
\|x - x^*\|_2^2 &\leq \omega^2 \left(\|\nabla f(x)\|_2 + \|Ex - t^*\|_2\right)^2 \\
&\leq 2\omega^2 \left(\|\nabla f(x)\|_2^2 + \|Ex - t^*\|_2^2\right) \\
&\leq 2\omega^2 \left[\|\nabla f(x)\|_2 \left(\|\nabla f(x)\|_2 + \frac{1}{\sigma_B}\|x - x^*\|_2\right)\right] \\
&\leq 2\omega^2 \left[\|\nabla f(x)\|_2 \left(\left(1 + \frac{\omega}{\sigma_B}\right)\|\nabla f(x)\|_2 + \frac{\omega}{\sigma_B}\|Ex - t^*\|_2\right)\right].
\end{aligned} \tag{10}$$

Since $\|Ex - t^*\|_2 = \|Ex - Ex^*\|_2 \le \|E\| \cdot \|x - x^*\|_2$, it follows from (10) that

$$\|Ex - t^*\|_2^2 \le 2\|E\|^2\omega^2 \left(1 + \frac{\omega}{\sigma_B}\right) \Big[\|\nabla f(x)\|_2(\|\nabla f(x)\|_2 + \|Ex - t^*\|_2)\Big].$$

Let $\gamma_B = 2\|E\|^2\omega^2 \left(1 + \frac{\omega}{\sigma_B}\right)$. Then, the above inequality is of the form $U^2 \le \gamma_B (V(U + V))$ with $U, V \ge 0$. This implies that $U \le \bar{\gamma}_B V$, where $\bar{\gamma}_B = \left(\gamma_B + \sqrt{\gamma_B^2 + 4\gamma_B}\right)/2$. Hence, we obtain $\|Ex - t^*\|_2 \le \bar{\gamma}_B \|\nabla f(x)\|_2$. This, together with Proposition 3, yields (7) with $\tau_B = \omega(1 + \bar{\gamma}_B)$.

Now, for any $x \in \mathbb{R}^d$ satisfying $Ex \in \mathcal{S}_B$, let $\bar{x}$ be the projection of x onto $\mathcal{X}$. By the optimality of $\bar{x}$, we have $\nabla f(\bar{x}) = \mathbf{0}$. It follows from the Lipschitz continuity of ∇f and (7) that

$$\mathrm{dist}(x, \mathcal{X}) \le \tau_B \|\nabla f(x)\|_2 = \tau_B \|\nabla f(x) - \nabla f(\bar{x})\|_2 \le \tau_B L \|x - \bar{x}\|_2 = \tau_B L \cdot \mathrm{dist}(x, \mathcal{X}),$$

which implies that $\tau_B \ge \frac{1}{L}$. □

The error bound (7) has a close relationship with the so–called quadratic growth condition on f. Such a property will be useful in our subsequent analysis.

PROPOSITION 4 *Consider the setting of Theorem 1 and let $\tau_B \ge \frac{1}{L}$ be the constant that satisfies* (7)*. Then,*

$$f(x) - f_{\min} \ge \frac{1}{2\tau_B^2 L} \mathrm{dist}(x, \mathcal{X})^2$$

for all $x \in \mathbb{R}^d$ satisfying $Ex \in \mathcal{S}_B$.

Proof. Since ∇f is Lipschitz continuous, we have

$$f(y) \ge f(z) + (y - z)^T \nabla f(z) + \frac{1}{2L}\|\nabla f(y) - \nabla f(z)\|_2^2 \quad \text{for all } y, z \in \mathbb{R}^d; \tag{11}$$

see, e.g., [24, Theorem 2.1.5]. Now, let $x \in \mathbb{R}^d$ be such that $Ex \in \mathcal{S}_B$, and let $\bar{x}$ be the projection of x onto $\mathcal{X}$. By taking $y = x$ and $z = \bar{x}$ in (11), we have

$$f(x) - f_{\min} \ge \frac{1}{2L}\|\nabla f(x)\|_2^2. \tag{12}$$

This, together with the error bound (7), implies the required result. □

4. Convergence Analysis of the IGM

Armed with Theorem 1 and Proposition 4, we are now ready to analyze the convergence rate of the IGM (3) in the following two scenarios:

(S1) The function g is strongly convex on $\mathbb{R}^m$.
(S2) The function g is strongly convex on $\mathcal{S}_B$ for all $B \in (0, \infty)$, and the (possibly random) error sequence $\{e^k\}_{k \ge 1}$ satisfies $\sum_{k=1}^{\infty} \|e^k\|_2^2 \le \Gamma$ for some $\Gamma \in (0, \infty)$.

It is easy to verify that the function g corresponding to the least squares regression problem (5) and the logistic regression problem (6) satisfies the assumption in (S1) and (S2), respectively.

The key step of the analysis is the following theorem, which establishes a recurrence relation between $f(x^{k+1})-f_{\min}$ and $f(x^k)-f_{\min}$ and forms the basis of all our subsequent results.

THEOREM 2 *Suppose that both Assumptions 1 and 2 hold for Problem* (4)*. Furthermore, suppose that either (S1) or (S2) holds. Let* $\{x^k\}_{k\geq 0}$ *be the sequence generated by the IGM* (3) *with step size* $\alpha_k = \alpha$ *for all* $k \geq 0$*, where* $\alpha \in \left(0, \frac{1}{L}\right)$*. Then, there exists a constant* $\tau \geq \frac{1}{L}$ *such that*

$$\operatorname{dist}(x^k, \mathcal{X}) \leq \tau \|\nabla f(x^k)\|_2 \tag{13}$$

for all $k \geq 0$*. Moreover, we have*

$$f(x^{k+1}) - f_{\min} \leq \mu(f(x^k) - f_{\min}) + \delta\|e^{k+1}\|_2^2 \tag{14}$$

for all $k \geq 0$*, where* $\mu = 1 - \frac{\alpha}{L\tau^2} \in (0,1)$ *and* $\delta = \frac{\alpha}{2}$*.*

Proof. Let us first verify that in both scenarios (S1) and (S2), there exists a $B > 0$ such that $Ex^k \in \mathcal{S}_B$ for all $k \geq 0$, and that g is strongly convex on $\mathcal{S}_B$. In scenario (S1), we can simply set $B = \infty$ to get $\mathcal{S}_B = \mathbb{R}^m$. In scenario (S2), observe that Corollary 1(b) implies

$$0 \leq f(x^k) - f_{\min} \leq f(x^0) - f_{\min} + \sum_{j=1}^{k} \|e^j\|_2^2 \leq f(x^0) - f_{\min} + \Gamma$$

for all $k \geq 0$. Hence, by [29, Fact 4.1], the sequence $\{Ex^k\}_{k\geq 0}$ is bounded. Consequently, there exists a $B \in (0, \infty)$, which does not depend on the realization of $\{x^k\}_{k\geq 0}$, such that $Ex^k \in \mathcal{S}_B$, and g is strongly convex on $\mathcal{S}_B$.

The above argument implies that Theorem 1 applies to both scenarios (S1) and (S2). Hence, there exists a $\tau > 0$, which does not depend on the realization of $\{e^k\}_{k\geq 1}$, such that

$$\operatorname{dist}(x^k, \mathcal{X}) \leq \tau \|\nabla f(x^k)\|_2$$

for all $k \geq 0$. Since ∇f is Lipschitz continuous, we have

$$f(y) \leq f(z) + (y-z)^T \nabla f(z) + \frac{L}{2}\|y-z\|_2^2 \quad \text{for all } y, z \in \mathbb{R}^d; \tag{15}$$

see, e.g., [24, Theorem 2.1.5]. Hence, by (15) and the iteration formula $x^{k+1} = x^k -$

$\alpha(\nabla f(x^k) + e^{k+1})$, we obtain

$$\begin{aligned} f(x^{k+1}) - f_{\min} &= f(x^k - \alpha(\nabla f(x^k) + e^{k+1})) - f_{\min} \\ &\leq f(x^k) - \alpha(\nabla f(x^k) + e^{k+1})^T \nabla f(x^k) + \frac{\alpha^2 L}{2}\|\nabla f(x^k) + e^{k+1}\|_2^2 - f_{\min} \\ &= f(x^k) - f_{\min} - \alpha\left(1 - \frac{\alpha L}{2}\right)\|\nabla f(x^k)\|_2^2 + \frac{\alpha^2 L}{2}\|e^{k+1}\|_2^2 \\ &\quad + (\alpha^2 L - \alpha)\nabla f(x^k)^T e^{k+1}. \end{aligned} \tag{16}$$

Since $\alpha \leq \frac{1}{L}$, we have $\alpha^2 L - \alpha < 0$. It follows that

$$\begin{aligned} (\alpha^2 L - \alpha)\nabla f(x^k)^T e^{k+1} &\leq |\alpha^2 L - \alpha| \cdot \|\nabla f(x^k)\|_2 \cdot \|e^{k+1}\|_2 \\ &= \alpha(1 - \alpha L) \cdot \|\nabla f(x^k)\|_2 \cdot \|e^{k+1}\|_2. \end{aligned}$$

Using the identity $ab \leq (a^2 + b^2)/2$, which is valid for all $a, b \in \mathbb{R}$, the above leads to

$$(\alpha^2 L - \alpha)\nabla f(x^k)^T e^{k+1} \leq \frac{\alpha(1 - \alpha L)}{2}\|\nabla f(x^k)\|_2^2 + \frac{\alpha(1 - \alpha L)}{2}\|e^{k+1}\|_2^2.$$

Upon substituting this into (16), we obtain

$$f(x^{k+1}) - f_{\min} \leq f(x^k) - f_{\min} - \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2 + \frac{\alpha}{2}\|e^{k+1}\|_2^2. \tag{17}$$

Now, by taking $y = x^k$ and $x = \bar{x}^k$ (the projection of x^k onto $\mathcal{X}$) in (15), we have

$$f(x^k) - f_{\min} \leq \frac{L}{2}\mathrm{dist}(x^k, \mathcal{X})^2.$$

This, together with (13), implies that

$$\|\nabla f(x^k)\|_2^2 \geq \frac{1}{\tau^2}\mathrm{dist}(x^k, \mathcal{X})^2 \geq \frac{2}{\tau^2 L}(f(x^k) - f_{\min}). \tag{18}$$

Upon combining (17) and (18), we obtain

$$f(x^{k+1}) - f_{\min} \leq f(x^k) - f_{\min} - \frac{\alpha}{\tau^2 L}(f(x^k) - f_{\min}) + \frac{\alpha}{2}\|e^{k+1}\|_2^2,$$

which implies the desired result. □

Theorem 2 suggests that the convergence behavior of the IGM (3) can be deduced from the behavior of the squared norms of the error vectors $\{e^k\}_{k\geq 1}$. As a simple application of Theorem 2, observe that by unrolling the inequality (14), we obtain

$$f(x^k) - f_{\min} \leq \mu^k \left(f(x^0) - f_{\min}\right) + \delta \sum_{j=1}^{k} \mu^{k-j}\|e^j\|_2^2. \tag{19}$$

This motivates the following results, whose proof can be found in Appendix D:

COROLLARY 2 *Consider the setting of Theorem 2.*

(a) (Sublinear Convergence) Suppose that for some $\rho > 0$, we have $\|e^k\|_2^2 \leq B_k = O\left(\frac{1}{k^{1+\rho}}\right)$ for all $k \geq 1$. Then, the sequence of iterates $\{x^k\}_{k\geq 0}$ satisfies

$$f(x^{k+1}) - f_{\min} \leq O\left(\frac{1}{(k+1)^{1+\rho}}\right) \quad \text{and} \quad \operatorname{dist}(x^k, \mathcal{X}) \leq O\left(\frac{1}{(k+1)^{(1+\rho)/2}}\right)$$

for all $k \geq 0$. In particular, the sequence $\{f(x^k)\}_{k\geq 0}$ (resp. $\{x^k\}_{k\geq 0}$) converges at least sublinearly to $f_{\min}$ (resp. an element in $\mathcal{X}$).

(b) (Linear Convergence) Suppose that for some $\rho \in (0,1)$, we have $\|e^k\|_2^2 \leq B_k = O(\rho^k)$ for all $k \geq 1$. Then, there exists a $c \in (0,1)$ such that the sequence of iterates $\{x^k\}_{k\geq 0}$ satisfies

$$f(x^{k+1}) - f_{\min} \leq O(c^{2(k+1)}) \quad \text{and} \quad \operatorname{dist}(x^k, \mathcal{X}) \leq O(c^k)$$

for all $k \geq 0$. In particular, the sequence $\{f(x^k)\}_{k\geq 0}$ (resp. $\{x^k\}_{k\geq 0}$) converges at least linearly to $f_{\min}$ (resp. an element in $\mathcal{X}$).

In the context of inexact gradient methods, Corollary 2 extends the results of Schmidt et al. [27] and Friedlander and Schmidt [10] in two ways. First, it shows that when applied to the structured convex optimization problem (4), the IGM (3) can achieve an $O\left(\frac{1}{k^2}\right)$ convergence rate for the sequence $\{f(x^k) - f_{\min}\}_{k\geq 0}$ even when the error norms $\{\|e^k\|_2\}_{k\geq 1}$ decrease at an $O\left(\frac{1}{k}\right)$ rate. This should be contrasted with the case of a general convex optimization problem, for which the IGM (3) is only known to achieve an $O\left(\frac{\log^2 k}{k}\right)$ convergence rate for the sequence $\{\min_{0\leq j\leq k} f(x^j) - f_{\min}\}_{k\geq 0}$ [27, Proposition 1]. Second, our analysis shows that even when the objective function f is not strongly convex, it is possible to establish a sublinear (resp. linear) convergence rate for the sequence of iterates $\{x^k\}_{k\geq 0}$, provided that the error norms $\{\|e^k\|_2\}_{k\geq 1}$ decrease to zero at a sublinear (resp. linear) rate.

REMARKS. Since the relationship in Theorem 2 holds for every realization of the error sequence $\{e^k\}_{k\geq 1}$, it also holds in expectation. Thus, we can derive bounds on the expected convergence rates of $\{f(x^k)\}_{k\geq 0}$ and $\{x^k\}_{k\geq 0}$ whenever bounds on $\{\mathbb{E}\left[\|e^k\|_2^2\right]\}_{k\geq 1}$ are available. As an illustration, we have the following extension of Corollary 2:

COROLLARY 3 *Consider the setting of Theorem 2.*

(a) (Expected Sublinear Convergence) Suppose that for some $\rho > 0$, we have $\mathbb{E}\left[\|e^k\|_2^2\right] \leq B_k = O\left(\frac{1}{k^{1+\rho}}\right)$ for all $k \geq 1$. Then, the sequence of iterates $\{x^k\}_{k\geq 0}$ satisfies

$$\mathbb{E}\left[f(x^{k+1}) - f_{\min}\right] \leq O\left(\frac{1}{(k+1)^{1+\rho}}\right) \quad \text{and} \quad \mathbb{E}\left[\operatorname{dist}(x^k, \mathcal{X})\right] \leq O\left(\frac{1}{(k+1)^{(1+\rho)/2}}\right)$$

for all $k \geq 0$.

(b) (Expected Linear Convergence) Suppose that for some $\rho \in (0,1)$, we have $\mathbb{E}\left[\|e^k\|_2^2\right] \leq B_k = O(\rho^k)$ for all $k \geq 1$. Then, there exists a $c \in (0,1)$ such that the sequence of iterates $\{x^k\}_{k\geq 0}$ satisfies

$$\mathbb{E}\left[f(x^{k+1}) - f_{\min}\right] \leq O(c^{2(k+1)}) \quad \text{and} \quad \mathbb{E}\left[\operatorname{dist}(x^k, \mathcal{X})\right] \leq O(c^k)$$

for all $k \geq 0$.

The proof of Corollary 3 can be found in Appendix E.

5. Applications to Least Squares Regression and Logistic Regression

We now demonstrate the power of the analysis framework developed in previous sections by applying it to three recently proposed first–order methods—namely, the incremental gradient method with increasing sample size [10], the stochastic variance–reduced gradient method [14], and the incremental aggregated gradient method [30]—and establishing, in a unified manner and for the first time, that these methods converge linearly when applied to the least squares regression problem and/or the logistic regression problem. To set the stage for our analysis, let us collect some useful properties of the least squares regression and logistic regression problems.

Recall that the least squares regression problem is

$$\min_{x\in\mathbb{R}^d} \quad f(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(a_i^T x - b_i\right)^2}_{f_i(x)}, \tag{LSR}$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are the given data, while the logistic regression problem is

$$\min_{x\in\mathbb{R}^d} \quad f(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\log\left(1+\exp(-b_i a_i^T x)\right)}_{f_i(x)}, \tag{LR}$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, 1\}$ are the given data. As discussed in Section 2.1, both (LSR) and (LR) satisfy Assumptions 1 and 2. Moreover, as remarked in Section 4, the objective functions of (LSR) and (LR) satisfy the requirements in scenarios (S1) and (S2), respectively. The following proposition provides an estimate of the Lipschitz constant of ∇f in (LSR) and (LR).

PROPOSITION 5 *Let* $L > 0$ *be a constant such that* $\max_{1\leq i\leq n} \|a_i\|_2^2 \leq L$. *Then, for either* (LSR) *or* (LR), ∇f *is Lipschitz continuous with parameter at most* L.

Proof. Note that for both (LSR) and (LR), f is twice continuously differentiable on $\mathbb{R}^d$. Hence, the Lipschitz constant of ∇f is bounded above by $\sup_{x\in\mathbb{R}^d} \|\mathcal{H}_f(x)\|$, where $\mathcal{H}_f(x)$ is the Hessian of f at x and $\|\cdot\|$ is the operator norm. For (LSR), we have

$$\mathcal{H}_f(x) = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T.$$

Since $\|a_i a_i^T\| = \|a_i\|_2^2 \leq L$, it follows that $\sup_{x\in\mathbb{R}^d} \|\mathcal{H}_f(x)\| \leq L$.

On the other hand, for (LR), we have

$$\mathcal{H}_f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{b_i^2 \exp(-b_i a_i^T x)}{(1+\exp(-b_i a_i^T x))^2} \cdot a_i a_i^T.$$

Since $b_i \in \{-1, 1\}$, we have $b_i^2 = 1$. Moreover, the fact that $\exp(-b_i a_i^T x)$ is positive implies

$$\frac{\exp(-b_i a_i^T x)}{(1+\exp(-b_i a_i^T x))^2} \leq \frac{1}{4}.$$

Therefore, we obtain $\|\mathcal{H}_f(x)\| \leq \frac{1}{4}\left\|\frac{1}{n}\sum_{i=1}^n a_i a_i^T\right\| \leq L$ for all $x \in \mathbb{R}^d$, as desired. □

5.1 *Incremental Gradient Method with Increasing Sample Size*

The incremental gradient method is specifically designed for solving optimization problems with a finite–sum structure such as Problem (1) (we refer the readers to the excellent survey [4]). The high–level description of the method is given in Algorithm 1.

Algorithm 1 Incremental Gradient Method with Increasing Sample Size

Input: initial sample set $I_0 \subset \mathscr{N} = \{1, 2, \ldots, n\}$, step size $\alpha > 0$, initial point $x^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
calculate the search direction by

$$G^k = \frac{1}{|I_k|}\sum_{i \in I_k} \nabla f_i(x^k)$$

update x^k by the formula

$$x^{k+1} = x^k - \alpha G^k$$

choose $I_{k+1} \subset \mathscr{N}$ according to some pre–specified rule
end for

By viewing Algorithm 1 as an inexact gradient method of the form (3), it can be easily verified that the error vector e^{k+1} in the k–th iteration is given by

$$e^{k+1} = G_k - \nabla f(x^k) = \frac{n - |I_k|}{n|I_k|}\sum_{i \in I_k} \nabla f_i(x^k) - \frac{1}{n}\sum_{i \in \mathscr{N} \setminus I_k} \nabla f_i(x^k). \tag{20}$$

If we form I_k by sampling a fixed number of elements from $\mathscr{N}$ uniformly without replacement and the sampling is done independent of $I_0, I_1, \ldots, I_{k-1}$ for all $k \geq 0$, then we also have

$$\mathbb{E}\left[\|e^{k+1}\|_2^2 \,\middle|\, \mathscr{F}_k\right] = \left(\frac{n - |I_k|}{n|I_k|}\right)\left(\frac{1}{n-1}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f(x^k)\|_2^2\right) \tag{21}$$

for all $k \geq 0$, where $\mathscr{F}_k$ is the σ–algebra generated by $e^1, e^2, \ldots, e^k$ with $\mathscr{F}_0 = \emptyset$; see [10, Section 3.2].

In order to apply the machinery developed in Section 4 to understand the convergence behavior of Algorithm 1, we need to bound the squared error norms $\{\|e^k\|_2^2\}_{k \geq 1}$. Towards that end, let us study the least squares regression and the logistic regression problems separately.

5.1.1 Least Squares Regression

Consider first the case where we apply Algorithm 1 to solve the least squares regression problem (LSR). The following proposition provides bounds on the error norms $\{\|e^k\|_2\}_{k\geq 1}$ under two different rules of choosing the sets $\{I_k\}_{k\geq 0}$. Its proof can be found in Appendix F.

PROPOSITION 6 *The following hold:*

(a) (Generic) *Suppose that the sets $\{I_k\}_{k\geq 0}$ satisfy $\frac{n}{2} \leq |I_k| \leq n$ for all $k \geq 0$. Then, we have*

$$\|e^{k+1}\|_2^2 \leq 8L\frac{n-|I_k|}{n} f(x^k) \tag{22}$$

for all $k \geq 0$, where $L > 0$ is given in Proposition 5.

(b) (Uniform Sampling without Replacement) *Suppose that the sets $\{I_k\}_{k\geq 0}$ are formed by sampling a fixed number of elements from $\mathcal{N}$ uniformly without replacement. Then, we have*

$$\mathbb{E}\left[\|e^{k+1}\|_2^2\right] \leq 16L\frac{n-|I_k|}{(n-1)|I_k|}\mathbb{E}\left[f(x^k)\right] \tag{23}$$

for all $k \geq 0$, where $L > 0$ is given in Proposition 5.

Since the premises of Theorem 2 are satisfied by the least squares regression problem (LSR), under the assumption that the sets $\{I_k\}_{k\geq 0}$ satisfy $\frac{n}{2} \leq |I_0| \leq |I_1| \leq \cdots$, the inequalities (14) and (22) together yield

$$\begin{aligned} f(x^{k+1}) - f_{\min} &\leq \mu\left(f(x^k) - f_{\min}\right) + 8\delta L\frac{n-|I_k|}{n}\left[\left(f(x^k) - f_{\min}\right) + f_{\min}\right] \\ &= \left(\mu + 8\delta L\frac{n-|I_k|}{n}\right)\left(f(x^k) - f_{\min}\right) + 8\delta L f_{\min}\frac{n-|I_k|}{n} \\ &\leq \bar{\mu}\left(f(x^k) - f_{\min}\right) + \bar{\delta}E_{k+1}, \end{aligned} \tag{24}$$

where

$$\bar{\mu} = \mu + 8\delta L\frac{n-|I_0|}{n}, \quad \bar{\delta} = 8\delta L f_{\min}, \quad E_{k+1} = \frac{n-|I_k|}{n} \quad \text{for } k = 0, 1, \ldots.$$

Hence, if $\bar{\mu} \in (0,1)$ (which can be guaranteed when $|I_0|$ is sufficiently large), then Corollary 2 implies that as long as $\{E_k\}_{k\geq 1}$ decreases (sub)linearly to zero, the sequence $\{f(x^k)\}_{k\geq 0}$ (resp. $\{x^k\}_{k\geq 0}$) converges at least (sub)linearly to $f_{\min}$ (resp. an element in $\mathcal{X}$).

REMARK. The assumptions $|I_0| \geq \frac{n}{2}$ and $\bar{\mu} \in (0,1)$ are only made for the sake of simplicity and can be dropped altogether. Indeed, as long as $\{E_k\}_{k\geq 1}$ decreases to zero, there will be an index $K \geq 1$ such that $|I_k| \geq \frac{n}{2}$ and $\mu + 8\delta L E_{k+1} \in (0,1)$ for all $k \geq K$. Hence, one can still derive the desired convergence rate results using arguments similar to the proof of Corollary 2.

On the other hand, under the assumption that the sets $\{I_k\}_{k\geq 0}$ are obtained by sampling uniformly from $\mathscr{N}$ without replacement and satisfy $|I_0| \leq |I_1| \leq \cdots$, the inequalities (14) and (23) imply

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) - f_{\min}\right] &\leq \left(\mu + 16\delta L\frac{n-|I_k|}{(n-1)|I_k|}\right)\mathbb{E}\left[f(x^k) - f_{\min}\right] + 16\delta L f_{\min}\frac{n-|I_k|}{(n-1)|I_k|} \\ &\leq \tilde{\mu}\mathbb{E}\left[f(x^k) - f_{\min}\right] + \tilde{\delta}\tilde{E}_{k+1},\end{aligned}$$

where

$$\tilde{\mu} = \mu + 16\delta L\frac{n-|I_0|}{(n-1)|I_0|}, \quad \tilde{\delta} = 16\delta L f_{\min}, \quad \tilde{E}_{k+1} = \frac{n-|I_k|}{(n-1)|I_k|} \quad \text{for } k = 0, 1, \ldots.$$

In particular, if $\tilde{\mu} \in (0,1)$ (which again can be guaranteed when $|I_0|$ is sufficiently large), then by Corollary 3, we see that the expected rate at which $\{f(x^k)\}_{k\geq 0}$ (resp. $\{x^k\}_{k\geq 0}$) converges to $f_{\min}$ (resp. an element in $\mathcal{X}$) is (sub)linear, provided that $\{\tilde{E}_k\}_{k\geq 1}$ decreases (sub)linearly to zero.

5.1.2 Logistic Regression

Next, let us consider the case where Algorithm 1 is used to solve the logistic regression problem (LR). Recall that for (LR), we have $f_i(x) = \log(1 + \exp(-b_i a_i^T x))$ for $i = 1, 2, \ldots, n$. To bound the error norms $\{\|e^k\|_2\}_{k\geq 1}$, we first compute

$$\nabla f_i(x) = \frac{-b_i \exp(-b_i a_i^T x)}{1 + \exp(-b_i a_i^T x)} a_i \quad \text{for } i = 1, 2, \ldots, n.$$

Upon noting that $\max_{1\leq i\leq n} \|a_i\|_2^2 \leq L$ and $b_i \in \{-1, 1\}$, we obtain

$$\|e^{k+1}\|_2^2 \leq \left(\frac{n-|I_k|}{n|I_k|}\sum_{i\in I_k}\|\nabla f_i(x^k)\|_2 + \frac{1}{n}\sum_{i\in\mathscr{N}\setminus I_k}\|\nabla f_i(x^k)\|_2\right)^2 \leq 4L^2 E_{k+1}^2,$$

where, as before, $E_{k+1} = \frac{n-|I_k|}{n}$. Since the premises of Theorem 2 are satisfied by the logistic regression problem (LR), the above inequality, together with (14), implies that

$$f(x^{k+1}) - f_{\min} \leq \mu\left(f(x^k) - f_{\min}\right) + 4\delta L E_{k+1}^2.$$

In particular, if $\{E_k^2\}_{k\geq 1}$ decreases (sub)linearly to zero, then by Corollary 2, we conclude that $\{f(x^k)\}_{k\geq 0}$ (resp. $\{x^k\}_{k\geq 0}$) converges at least (sub)linearly to $f_{\min}$ (resp. an element in $\mathcal{X}$).

If the sets $\{I_k\}_{k\geq 0}$ are obtained via uniform sampling from $\mathscr{N}$ without replacement,

then using (21) and $\max_{1\leq i\leq n}\|a_i\|_2^2 \leq L$, we have

$$\begin{aligned}\mathbb{E}\left[\|e^{k+1}\|_2^2 \,|\, \mathscr{F}_k\right] &\leq \left(\frac{n-|I_k|}{n|I_k|}\right)\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\|\nabla f_i(x^k)\|_2 + \|\nabla f(x^k)\|_2\right)^2\right] \\ &\leq 4L^2\frac{n-|I_k|}{(n-1)|I_k|}.\end{aligned}$$

This implies that for all $k \geq 0$,

$$\mathbb{E}\left[f(x^{k+1}) - f_{\min}\right] \leq \mu\mathbb{E}\left[f(x^k) - f_{\min}\right] + 4\delta L^2\tilde{E}_{k+1},$$

where, as before, $\tilde{E}_{k+1} = \frac{n-|I_k|}{(n-1)|I_k|}$. Hence, by Corollary 3, the expected rate at which $\{f(x^k)\}_{k\geq 0}$ (resp. $\{x^k\}_{k\geq 0}$) converges to $f_{\min}$ (resp. an element in $\mathcal{X}$) is (sub)linear, provided that $\{\tilde{E}_k\}_{k\geq 1}$ decreases (sub)linearly to zero.

5.2 *Stochastic Variance–Reduced Gradient Method for Least Squares Regression*

The stochastic variance–reduced gradient (SVRG) method is recently proposed by Johnson and Zhang [14] for solving data fitting problems with a finite–sum structure. The high–level description of the SVRG method is given in Algorithm 2. Unlike the standard stochastic gradient descent (SGD) method, the SVRG method has an additional epoch index s. In each epoch, the SVRG method performs a fixed number of SGD–type updates with a correction term $\mu - \nabla f_{i_k}(\omega)$ (see (25)). Such a correction term aims to reduce the variance of the approximate gradient ∇f_{i_k}. Under the assumption that f is strongly convex and some other standard assumptions, it is proven in [14] (cf. [26]) that by choosing the step size α and update frequency l properly, the sequence $\{\omega^s\}_{s\geq 0}$ generated by the SVRG method will converge to the optimal solution ω^* linearly in expectation; i.e., $\mathbb{E}\left[\|\omega^{s+1} - \omega^*\|_2\right] \leq \rho\mathbb{E}\left[\|\omega^s - \omega^*\|_2\right]$ for some constant $\rho \in (0,1)$.

Algorithm 2 Stochastic Variance–Reduced Gradient Method

Input: number of iterations in each epoch $l \geq 1$, step size $\alpha > 0$, initial point $\omega^0 \in \mathbb{R}^d$
for $s = 1, 2, \ldots$ **do**
  $\omega = \omega^{s-1}$
  $\mu = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\omega)$
  $x^0 = \omega$
  for $k = 0, 1, \ldots, l-1$ **do**
    sample an index $i_k \in \mathscr{N}$ uniformly at random
    update x^k by the formula

$$x^{k+1} = x^k - \alpha(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega) + \mu) \tag{25}$$

  end for
  sample an index $k \in \{0, 1, \ldots, l-1\}$ uniformly at random and set $\omega^s = x^k$
end for

In this section, we study the convergence performance of the SVRG method for solving

the least squares regression problem (LSR). Note that the objective function of (LSR) is not strongly convex in general. Thus, existing analyses of the SVRG method, such as those in [14, 26], do not apply here. However, using the machinery developed in Section 4, we can still show that the iterates generated by the SVRG method will converge linearly in expectation. We remark that a similar result has been established independently by Gong and Ye [11]. However, our proof highlights the fact that the SVRG method can be treated as an IGM.

To begin our analysis, consider an arbitrary epoch $s \geq 1$. For simplicity, let ω and ω^+ denote the input and output of this epoch, respectively. Also, let $\{x^k\}_{0\leq k\leq l}$ be the sequence generated by update formula (25) in this epoch. Thus, we have $x^0 = \omega$, and ω^+ is selected from $\{x^0, x^1, \ldots, x^{l-1}\}$ uniformly at random. Upon viewing the update formula (25) as an instance of (3), we obtain

$$\begin{aligned} e^{k+1} &= \nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega) + \mu - \nabla f(x^k) \\ &= \nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega) + \frac{1}{n}\sum_{i=1}^{n}\Big[\nabla f_i(\omega) - \nabla f_i(x^k)\Big] \end{aligned} \tag{26}$$

for $k = 0, 1, \ldots, l-1$, where i_k is a random variable satisfying $\Pr(i_k = i) = \frac{1}{n}$ for $i = 1, 2, \ldots, n$. Driven by the randomness of i_k, the error vectors $\{e^k\}_{1\leq k\leq l}$ are also random. Let $\mathscr{F}_k$ be the σ–algebra generated by $e^1, e^2, \ldots, e^k$ with $\mathscr{F}_0 = \emptyset$. By utilizing the structure of e^k, we can prove the following bounds:

PROPOSITION 7 *Suppose that the SVRG method is used to solve the least squares regression problem* (LSR). *Then, for $k = 0, 1, \ldots, l-1$, we have*

$$\begin{aligned} \mathbb{E}\left[f(x^{k+1}) - f_{\min} \,\middle|\, \mathscr{F}_k\right] &\leq \left[1 - \frac{2\alpha}{\tau^2 L}\left(1 - \frac{\alpha L}{2}\right)\right](f(x^k) - f_{\min}) \\ &\quad + \frac{\alpha^2 L}{2}\mathbb{E}\left[\|e^{k+1}\|_2^2 \,\middle|\, \mathscr{F}_k\right], \end{aligned} \tag{27}$$

$$\mathbb{E}\left[\|e^{k+1}\|_2^2 \,\middle|\, \mathscr{F}_k\right] \leq 4L(f(x^k) - f_{\min} + f(\omega) - f_{\min}). \tag{28}$$

The proof of Proposition 7 can be found in Appendix G. Armed with Proposition 7, we can now apply the machinery developed in Section 4 to establish the global linear convergence of the SVRG method for solving (LSR).

THEOREM 3 *Suppose that the SVRG method is used to solve the least squares regression problem* (LSR). *If the step size α and update frequency l satisfy*

$$\rho = \frac{1 + 2l\alpha^2 L^2}{l\left[\frac{2\alpha}{\tau^2 L}\left(1 - \frac{\alpha L}{2}\right) - 2\alpha^2 L^2\right]} < 1,$$

where $\tau \geq \frac{1}{L}$ is the constant in Theorem 2, then the sequence of iterates $\{\omega^s\}_{s\geq 0}$ from each epoch satisfies

$$\mathbb{E}\left[f(\omega^s) - f_{\min}\right] = O(\rho^s) \quad \textit{and} \quad \mathbb{E}\left[\operatorname{dist}(\omega^s, \mathcal{X})\right] = O(\rho^{s/2}).$$

Proof. We first show that by choosing α and l properly, $\rho < 1$ can be achieved. From Proposition 5, we know that L is the Lipschitz constant of ∇f. Hence, by Theorem 2,

we have $\tau \geq \frac{1}{L}$. We claim that as long as $\alpha < \frac{1}{3\tau^2 L^3}$, we have

$$\frac{2\alpha}{\tau^2 L}\left(1 - \frac{\alpha L}{2}\right) - 2\alpha^2 L^2 > 3\alpha^2 L^2. \tag{29}$$

Indeed, simple calculations show that (29) is equivalent to $\alpha L(5L^2\tau^2 + 1) < 2$. Since $L\tau \geq 1$, this is satisfied if $\alpha < \frac{1}{3\tau^2 L^3}$. In light of (29), it is immediate from the definition of ρ that for any fixed α satisfying $\alpha < \frac{1}{3\tau^2 L^3}$, we have $\rho < \frac{1+2l\alpha^2 L^2}{3l\alpha^2 L^2}$. Hence, by choosing l large enough, $\rho < 1$ can be achieved.

Now, consider an arbitrary epoch $s \geq 0$. For simplicity, let ω denote ω^s and ω^+ denote ω^{s+1}. Also, let $\{x^k\}_{0\leq k\leq l}$ be the sequence generated by the update formula (25) in this epoch. By (28) and (27), we have

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) - f_{\min} \,\middle|\, \mathscr{F}_k\right] \;\leq\; & \left[1 - \frac{2\alpha}{\tau^2 L}\left(1 - \frac{\alpha L}{2}\right)\right](f(x^k) - f_{\min}) \\ & + 2\alpha^2 L^2 (f(x^k) - f_{\min} + f(\omega) - f_{\min}).\end{aligned}$$

Taking expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) - f_{\min}\right] \;\leq\; & \left[1 - \frac{2\alpha}{\tau^2 L}\left(1 - \frac{\alpha L}{2}\right) + 2\alpha^2 L^2\right]\mathbb{E}\left[f(x^k) - f_{\min}\right] \\ & + 2\alpha^2 L^2 \left(f(\omega) - f_{\min}\right).\end{aligned}$$

Summing the above inequality over $k = 0, 1, \ldots, l-1$ and using the fact that $x^0 = \omega$ and $f(x^l) - f_{\min} \geq 0$, we have

$$\left[\frac{2\alpha}{\kappa^2 L}\left(1 - \frac{\alpha L}{2}\right) - 2\alpha^2 L^2\right]\sum_{k=0}^{l-1}\mathbb{E}\left[f(x^k) - f_{\min}\right] \leq 2l\alpha^2 L^2 (f(\omega) - f_{\min}).$$

By definition of ρ, the above can be simplified as

$$\frac{1}{l}\sum_{k=0}^{l-1}\mathbb{E}\left[f(x^k) - f_{\min}\right] \leq \rho(f(\omega) - f_{\min}). \tag{30}$$

Since ω^+ is selected from $\{x^0, x^1, \ldots, x^{l-1}\}$ uniformly at random, by taking expectation with respect to $x^0, x^1, \ldots, x^{l-1}$, we have

$$\mathbb{E}\left[f(\omega^+)\right] = \frac{1}{l}\sum_{k=0}^{l-1} f(x^k).$$

Taking another expectation with respect to $e^1, e^2, \ldots, e^l$ on both sides of the above yields

$$\mathbb{E}\left[f(\omega^+)\right] = \frac{1}{l}\sum_{k=0}^{l-1}\mathbb{E}\left[f(x^k)\right].$$

This, together with (30), implies that $\mathbb{E}[f(\omega^+) - f_{\min}] \leq \rho(f(\omega) - f_{\min})$, where the expectation is taken with respect to $e^1, e^2, \ldots, e^l$ and the randomness in selecting ω^+. Since the above inequality holds for an arbitrary epoch, we conclude that

$$\mathbb{E}[f(\omega^s) - f_{\min}] = O(\rho^s).$$

The convergence result for $\text{dist}(\omega^s, \mathcal{X})$ now follows from Proposition 4. □

Before we close this sub–section, let us remark that the techniques developed above can be extended to establish similar linear convergence results for other variants of the SVRG method, such as the prox–SVRG method in [32] and the mini–batch variant of the SVRG method in [15]. However, in view of the current length of the paper, we shall not pursue these directions here.

5.3 *Incremental Aggregated Gradient Method for Least Squares Regression*

Similar to the incremental gradient method and the SVRG method, the incremental aggregated gradient (IAG) method is also specifically designed for solving data fitting problems with the finite–sum structure. Algorithm 3 is a high–level description of the IAG method. The sequence of index vectors $\{\pi^k\}_{k\geq 0}$ is arbitrary but satisfies the delay bound (31), which ensures that each component ∇f_i in the summands (32) is updated at least once within every $K+1$ consecutive iterations of IAG. For example, the original IAG method proposed in [5] is a realization of Algorithm 3, where $\pi_i^0 = 0$ for $i = 1, \ldots, n$ and

$$\pi_i^k = \begin{cases} k & \text{if } i = (k-1 \bmod n) + 1, \\ \pi_i^{k-1} & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$ and $k = 1, 2, \ldots$. Hence, the components $\{\nabla f_i\}_{1\leq i\leq n}$ in the summation (32) are updated cyclically and the delay bound K is thus $n-1$. The generalized version of the IAG method in Algorithm 3 is developed in [30]. By introducing a uniform delay bound K, it allows us to model the distributed computation setting, where the center assigns the task of computing $\nabla f_1, \ldots, \nabla f_n$ to n processors and update x in an asynchronous manner.

In [30], it has been shown that the IAG method is globally convergent. Moreover, by assuming a local error bound property, the rate at which the sequence of iterates $\{x^k\}_{k\geq 0}$ generated by the IAG method converges to the optimal solution set $\mathcal{X}$ is asymptotically linear. Recently, it is shown in [12] that the IAG method enjoys a non–asymptotic linear convergence rate when the objective function f is strongly convex. In this section, we will show that when applying the IAG method to solve the least squares regression problem (LSR), the non–asymptotic linear convergence rate can be achieved without assuming strong convexity.

To begin, we note that the update formula (33) can be expressed as an instance of (3) with the error vector e^{k+1} given by

$$e^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \left[\nabla f_i(x^{\pi_i^k}) - \nabla f_i(x^k) \right], \tag{34}$$

where $k = 0, 1, \ldots$. Using the structure of e^{k+1} and the delay bound K, we have the

Algorithm 3 Incremental Aggregated Gradient Method

Input: step size $\alpha > 0$, delay bound $K \in \mathbb{N}_+$, initial point $x^0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

choose the index vector $\pi^k \in \mathbb{Z}^n$ such that each entry π_i^k satisfies

$$(k - K)_+ \leq \pi_i^k \leq k \quad \text{for } i = 1, 2, \ldots, n \tag{31}$$

calculate the search direction by

$$G^k = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^{\pi_i^k}) \tag{32}$$

update x^k by the formula

$$x^{k+1} = x^k - \alpha G^k \tag{33}$$

end for

following bound on $\|e^{k+1}\|_2^2$, whose proof can be found in Appendix H.

PROPOSITION 8 *Suppose that the IAG method is used to solve the least squares regression problem* (LSR). *Then, for $k = 0, 1, \ldots$, we have*

$$\|e^{k+1}\|_2^2 \leq 8nL \cdot \max_{(k-K)_+ \leq l \leq k} (f(x^l) - f_{\min}).$$

Equipped with Proposition 8, we are now ready to establish the global linear convergence of the IAG method for solving (LSR).

THEOREM 4 *Suppose that the IAG method is used to solve the least squares regression problem* (LSR). *If the step size α is chosen such that*

$$\rho = \left(1 - \frac{\alpha}{L\tau^2} + 10n\alpha^3 L^3 K^2\right)^{\frac{1}{2K+1}} < 1,$$

where $\tau \geq \frac{1}{L}$ is the constant in Theorem 2, then

$$f(x^k) - f_{\min} \leq O(\rho^k) \quad \textit{and} \quad \operatorname{dist}(x^k, \mathcal{X}) \leq O(\rho^{k/2}).$$

Proof. Since the premises of Theorem 2 hold for the least squares regression problem (LSR), we have $\tau \geq \frac{1}{L}$. Using (34) and the fact that ∇f_i is Lipschitz continuous with parameter L for $i = 1, \ldots, n$, we have

$$\|e^{k+1}\|_2 \leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^{\pi_i^k}) - \nabla f_i(x^k)\|_2 \leq \frac{L}{n} \sum_{i=1}^{n} \|x^{\pi_i^k} - x^k\|_2.$$

This implies that

$$\|e^{k+1}\|_2 \leq \frac{L}{n}\sum_{i=1}^{n}\sum_{j=(k-K)_+}^{k-1}\|x^{j+1}-x^j\|_2 \tag{35}$$

$$\leq L\sum_{j=(k-K)_+}^{k-1}\|x^{j+1}-x^j\|_2$$

$$= \alpha L\sum_{j=(k-K)_+}^{k-1}\|G^j\|_2 \tag{36}$$

$$\leq \alpha L\sum_{j=(k-K)_+}^{k-1}\left(\|\nabla f(x^j)\|_2+\|e^{j+1}\|_2\right), \tag{37}$$

where (35) follows from the triangle inequality and the fact that $(k-K)_+ \leq \pi_i^k \leq k$, (36) follows from (33), and (37) follows from (32), (34), and the triangle inequality. Hence,

$$\|e^{k+1}\|_2^2 \leq 2\alpha^2L^2K\sum_{j=(k-K)_+}^{k-1}\left(\|\nabla f(x^j)\|_2^2+\|e^{j+1}\|_2^2\right). \tag{38}$$

Since ∇f is Lipschitz continuous with parameter L, we see from (12) that

$$\|\nabla f(x^j)\|_2^2 \leq 2L(f(x^j)-f_{\min}) \quad \text{for } j=(k-K)_+,\ldots,k-1.$$

Together with (38) and Proposition 8, this yields

$$\begin{aligned}\|e^{k+1}\|_2^2 &\leq 2\alpha^2L^2K\sum_{j=(k-K)_+}^{k-1}\left(2L(f(x^j)-f_{\min})+8nL\cdot\max_{(j-K)_+\leq l\leq j}(f(x^l)-f_{\min})\right)\\ &\leq 2\alpha^2L^2K\sum_{j=(k-K)_+}^{k-1}\left(10nL\cdot\max_{(j-K)_+\leq l\leq j}(f(x^l)-f_{\min})\right)\\ &\leq 20n\alpha^2L^3K^2\cdot\max_{(k-2K)_+\leq l\leq k}(f(x^l)-f_{\min}).\end{aligned}$$

Upon combining this with (14), we obtain

$$f(x^{k+1})-f_{\min} \leq \left(1-\frac{\alpha}{L\tau^2}\right)(f(x^k)-f_{\min})+10n\alpha^3L^3K^2\cdot\max_{(k-2K)_+\leq l\leq k}(f(x^l)-f_{\min}). \tag{39}$$

To solve the above recurrence, we need the following result:

FACT 1 *([9, Lemma 3]) Let $\{V_k\}_{k\geq 0}$ be a sequence of non-negative real numbers satisfying*

$$V_{k+1} \leq pV_k+q\max_{(k-T)_+\leq j\leq k}V_j$$

for some constants $p, q \geq 0$ *and integer* $T \geq 0$. *Suppose that* $p + q < 1$. *Then, by letting* $\rho = (p+q)^{\frac{1}{T+1}} < 1$, *we have*

$$V_k \leq \rho^k V_0 \quad \textit{for } k = 0, 1, \ldots.$$

Now, observe that (39) corresponds to

$$V_k = f(x^k) - f_{\min}, \quad p = 1 - \frac{\alpha}{L\tau^2}, \quad q = 10n\alpha^3 L^3 K^2, \quad T = 2K$$

in Fact 1. Hence, if we choose α such that $p+q < 1$, or equivalently, $\alpha < \frac{1}{\sqrt{10n}K\tau L^2}$, then by Fact 1, we have

$$f(x^k) - f_{\min} \leq \left(1 - \frac{\alpha}{L\tau^2} + 10n\alpha^3 L^3 K^2\right)^{\frac{1}{2K+1}} (f(x^0) - f_{\min}).$$

Finally, using Proposition 4, the convergence result for $\mathrm{dist}(x^k, \mathcal{X})$ follows. □

6. Numerical Illustrations

In this section, we perform numerical experiments on the three instantiations of the IGM (3) in Section 5 to support our theoretical developments. Specifically, we use both synthetic and real datasets to test (i) the incremental gradient method with increasing sample size (IGM–ISS) for solving the least squares regression problem (LSR) and the logistic regression problem (LR), (ii) the stochastic variance–reduced gradient (SVRG) method for solving the least squares regression problem (LSR), and (iii) the incremental aggregated gradient (IAG) method for solving the least squares regression problem (LSR).

6.1 *On Identifying Linear Convergence*

To numerically investigate the rate at which a sequence $\{w^k\}_{k\geq 0}$ converges to w^*, it is common to plot the sequence $\{\log(\|w^k - w^*\|_2)\}_{k\geq 0}$ with respect to the index k. If the resulting curve roughly follows a straight line with negative slope, then it indicates that the convergence rate is at least linear. Based on this discussion, we shall present the following two types of figures for determining the convergence rates of IGMs:

(i) We plot the sequence $\{\log(f(x^k) - f_{\min})\}_{k\geq 0}$ with respect to the number of iterations $k \geq 0$. Since it is in general difficult to determine $f_{\min}$ exactly, we approximate it by solving the optimization problem to a high accuracy. Such a plot will allow us to deduce the convergence rate of the function value.

(ii) We plot the sequence $\{\log(\|\nabla f(x^k)\|_2)\}_{k\geq 0}$ with respect to the number of iterations $k \geq 0$. This is for identifying the convergence rate of $\{\mathrm{dist}(x^k, \mathcal{X})\}_{k\geq 0}$. Due to the error bound property (13) and the Lipschitz continuity of $\nabla f(x)$, the rate at which $\{\mathrm{dist}(x^k, \mathcal{X})\}_{k\geq 0}$ converges to 0 is identical with the rate at which $\{\|\nabla f(x^k)\|_2\}_{k\geq 0}$ converges to 0.

Since the focus of this work is on analyzing the convergence rates of IGMs, we shall not report the computational times of the tested methods.

6.2 *Datasets with Non–Strongly Convex Objectives*

The details of the tested datasets are listed in Table 1. The two datasets madelon and cpusmall can be downloaded from the LIBSVM datasets[1], and syntic is synthetic. We note that the matrix A in Table 1 is defined as $A = [a_1, a_2, \ldots, a_n]^T \in \mathbb{R}^{n\times d}$, where the a_i's correspond to the vectors in (LSR) and (LR). Given the information about the column rank of A (see the last column of Table 1), it is clear that the instances of (LSR) and (LR) defined by these datasets are not strongly convex.

Name	Problem	n	d	Column rank of A
madelon	logistic regression	2000	1000	≤ 500
syntic	least squares regression	2000	3000	≤ 2000
cpusmall	least squares regression	8192	20	≤ 12

Table 1. Details of the tested data sets.

6.3 *Convergence Performance*

We are now ready to present the results of our experiments. In all the experiments, we estimate the Lipschitz constant by letting $L = \max_{1\leq i\leq n} \|a_i\|_2^2$ (see Proposition 5).

6.3.1 IGM–ISS

From the discussion in Section 5.1, we see that for both the least squares regression problem (LSR) and the logistic regression problem (LR), if the sets $\{I_k\}_{k\geq 0}$ are obtained via uniform sampling from $\mathscr{N}$ without replacement, then as long as the sequence $\{\tilde{E}_k\}_{k\geq 1}$ defined by $\tilde{E}_{k+1} = \frac{n-|I_k|}{(n-1)|I_k|}$ converges to zero linearly, the resulting algorithm will have a linear convergence rate. In our implementation, we set

$$|I_{k+1}| = \left\lceil n \times \left(1 + \frac{0.9 \times (n - |I_k|)}{|I_k|}\right)^{-1} \right\rceil. \tag{40}$$

With this strategy of obtaining $\{I_k\}_{k\geq 0}$, it can be verified that the resulting errors $\{\tilde{E}_k\}_{k\geq 1}$ satisfy $\tilde{E}_{k+1} = O(0.9^k)$ for $k = 0, 1, \ldots$ and thus converge to zero at least linearly. Figures 1 and 2 show the convergence performance of IGM–ISS when solving the least squares regression problems cpusmall and syntic and the logistic regression problem madelon, respectively. It can be seen from the figures that the convergence rates of the objective values and the gradient norms are both at least linear.

6.3.2 SVRG and IAG

The theoretical results of Sections 5.2 and 5.3 show that when applied to the least squares regression problem (LSR), both the SVRG and IAG methods can achieve a linear rate of convergence when the parameters are chosen properly. In our implementation of the SVRG method, we set the step size $\alpha = 0.001/L$ and the iteration number of each epoch $l = 2n$. In the implementation of the IAG method, we set the same step size $\alpha = 0.001/L$

[1]Information about these two data sets can be found in `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

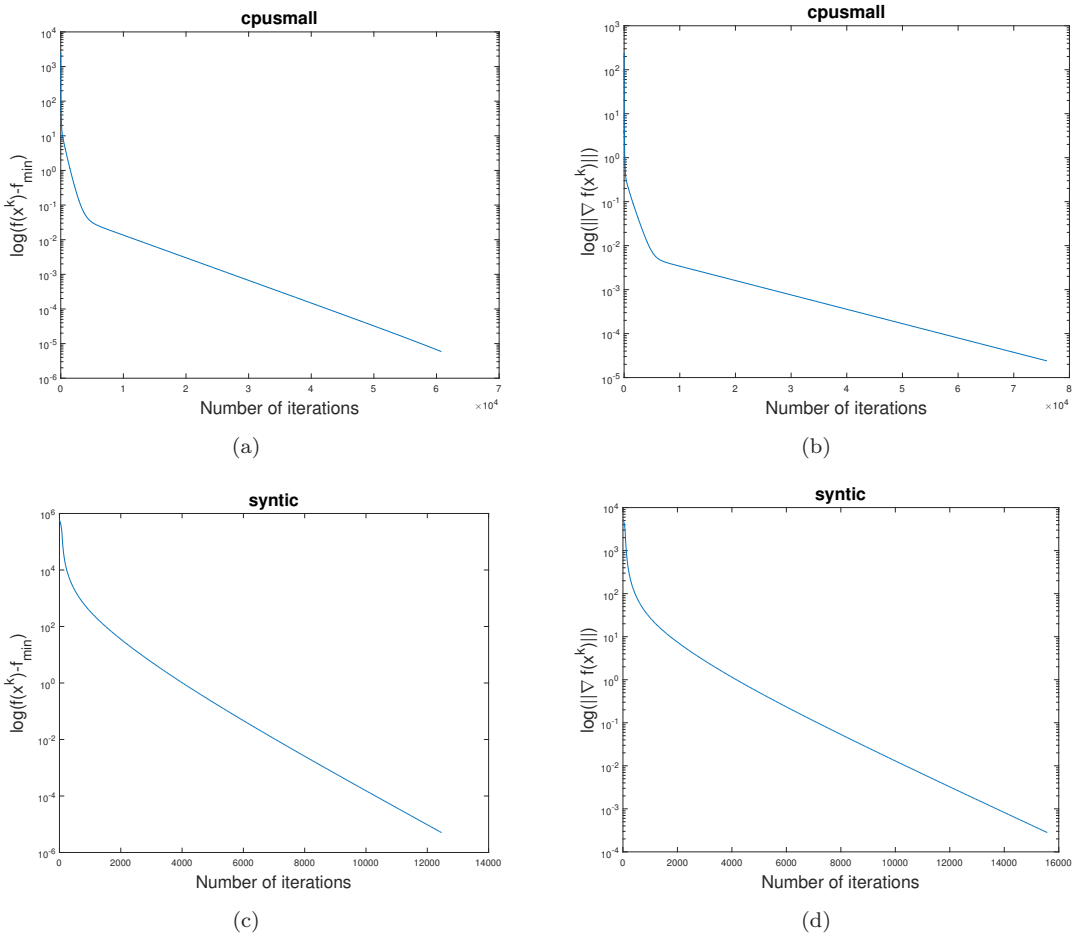

Figure 1. Convergence performance of IGM–ISS when solving the least squares regression problems cpusmall and syntic. Figures 1(a) and 1(c) show the convergence rates of the objective values. Figures 1(b) and 1(d) show the convergence rates of the gradient norms.

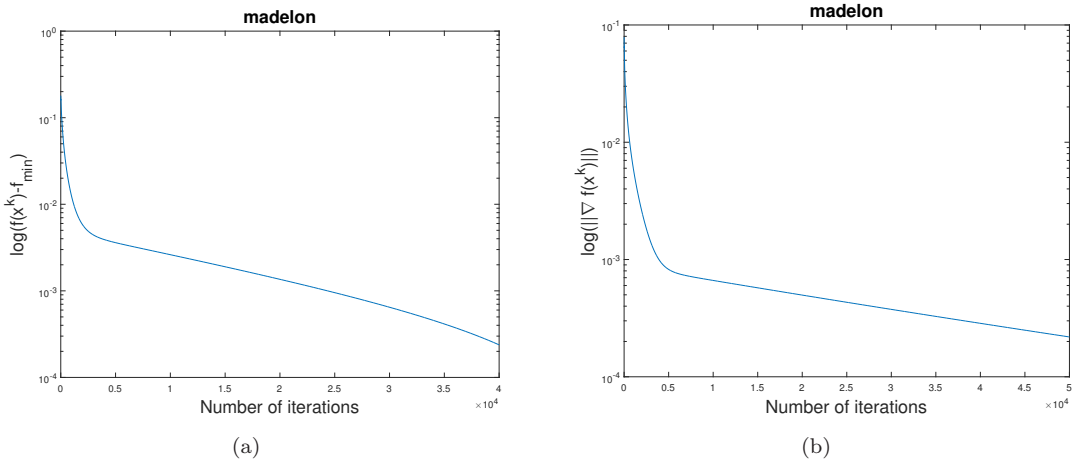

Figure 2. Convergence performance of IGM–ISS when solving the logistic regression problem madelon. Figure 2(a) shows the convergence rate of the objective values. Figure 2(b) shows the convergence rate of the gradient norms.

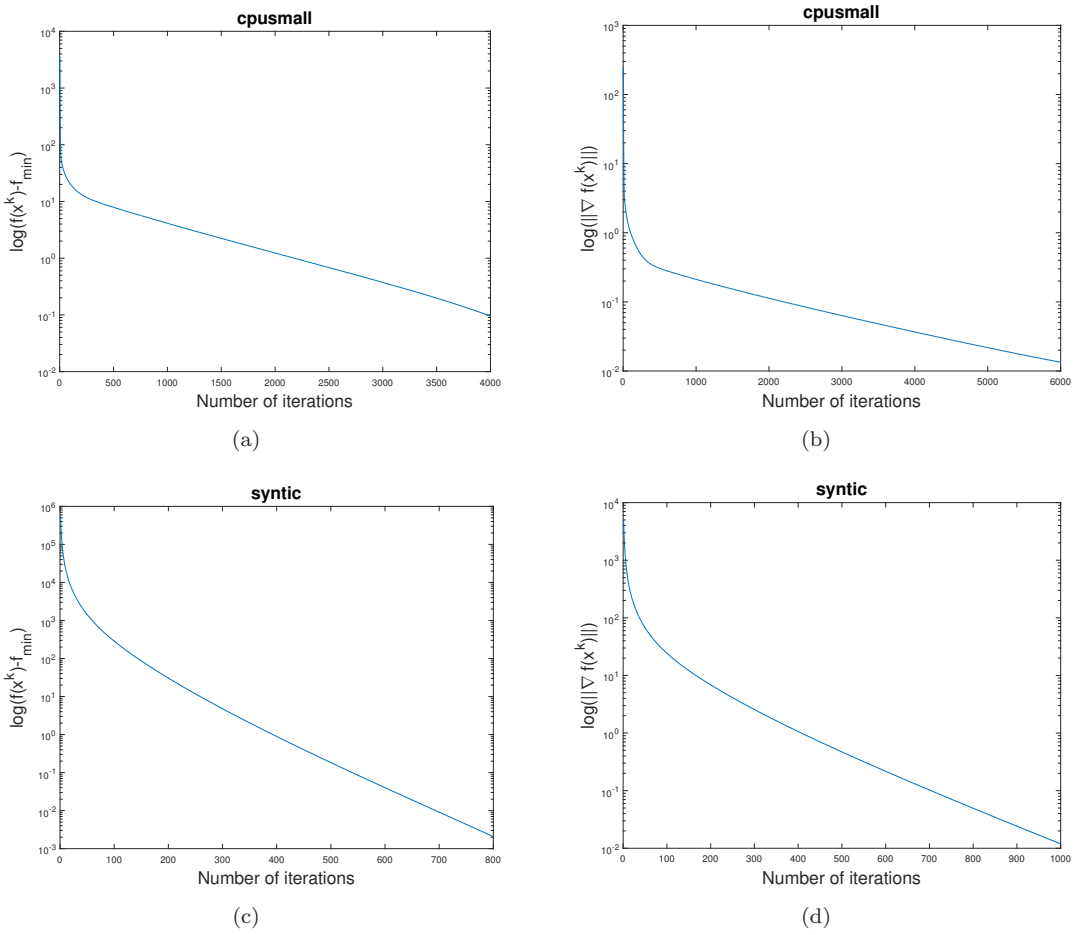

Figure 3. Convergence performance of the SVRG method when solving the least squares regression problems cpusmall and syntic. Figures 3(a) and 3(c) show the convergence rates of the objective values. Figures 3(b) and 3(d) show the convergence rates of the gradient norms.

and the delay bound $K = 10$. Figures 3 and 4 show the convergence performance of these two methods when solving the least squares regression problems cpusmall and syntic. It can be seen from the figures that both the objective values and the gradient norms converge at least linearly.

7. Concluding Remarks

In this paper, we considered a class of structured unconstrained convex optimization problems, in which the objective function is the composition of an affine mapping with a strictly convex function that has certain smoothness and curvature properties. This encapsulates many problems in machine learning and data fitting, such as least squares regression and logistic regression. We showed, in a unified manner, that a host of inexact gradient methods in the literature for solving this class of problems have a global linear rate of convergence. To obtain our results, we developed a so–called global error bound, which, roughly speaking, measures the distance between a point and the optimal set in terms of some easily computable quantities. In general, error bounds are very useful for proving strong convergence rate results for optimization algorithms (see, e.g., [21, 34, 35]).

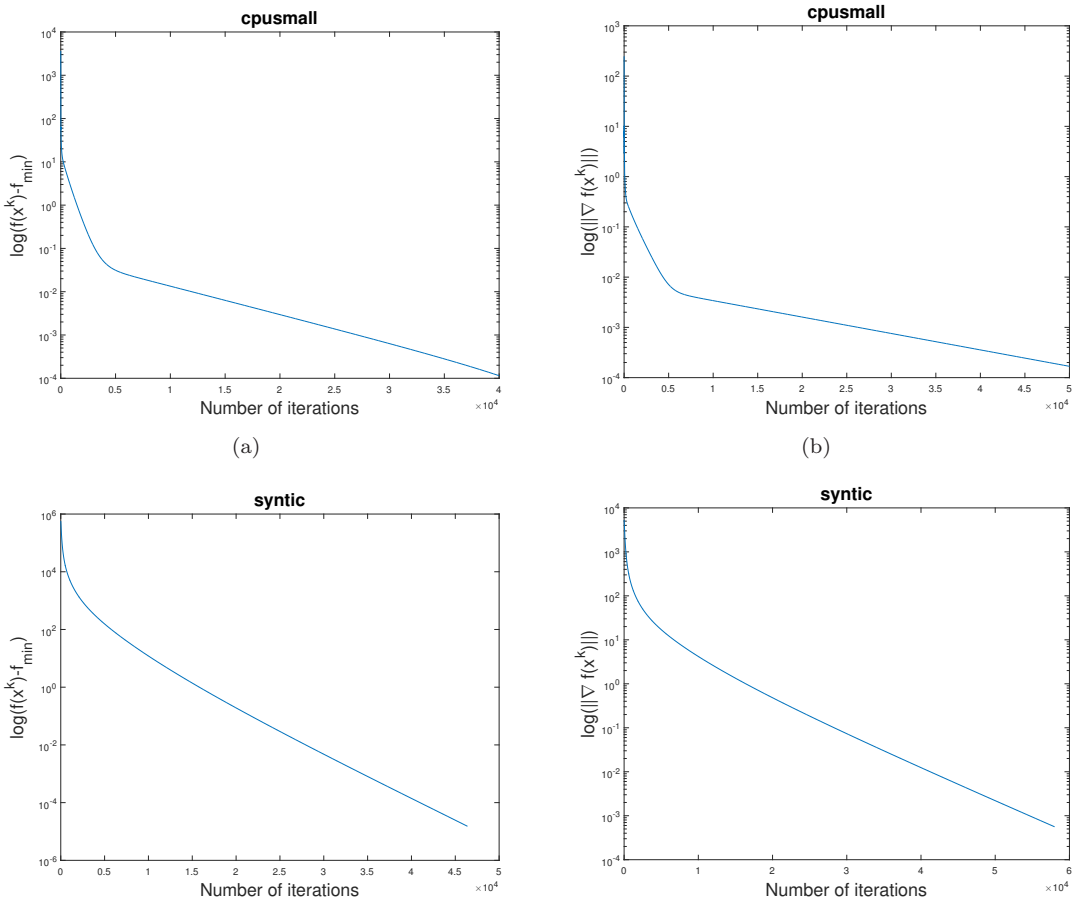

Figure 4. Convergence performance of the IAG method when solving the least squares regression problems cpusmall and syntic. Figures 4(a) and 4(c) show the convergence rates of the objective values. Figures 4(b) and 4(d) show the convergence rates of the gradient norms.

Thus, it would be interesting to see whether such an approach can be used to exploit the structure of optimization problems arising in machine learning and establish the linear convergence of some other first–order methods.

Acknowledgement

This work is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 14206814 and in part by a gift grant from Microsoft Research Asia.

References

[1] A. Agarwal, P.L. Barlett, P. Ravikumar, and M.J. Wainwright, *Information–Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization*, IEEE Transactions on Information Theory 58 (2012), pp. 3235–3249.

[2] F. Bach and E. Moulines, *Non–Asymptotic Analysis of Stochastic Approximation Algorithms for*

Machine Learning, in *Advances in Neural Information Processing Systems 24: Proceedings of the 2011 Conference*, 2011, pp. 451–459.

[3] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, Massachusetts, 1999.

[4] D.P. Bertsekas, *Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization*, in *Optimization for Machine Learning*, S. Sra, S. Nowozin, and S.J. Wright, eds., Neural Information Processing Series, MIT Press, Cambridge, Massachusetts, 2012, pp. 85–119.

[5] 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.

[6] L. Bottou and O. Bousquet, *The Tradeoffs of Large Scale Learning*, in *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 2007, pp. 161–168.

[7] R.H. Byrd, G.M. Chin, J. Nocedal, and Y. Wu, *Sample Size Selection in Optimization Models for Machine Learning*, Mathematical Programming, Series B 134 (2012), pp. 127–155.

[8] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A Fast Incremental Gradient Method with Support for Non–Strongly Convex Composite Objectives*, in *Advances in Neural Information Processing Systems 27: Proceedings of the 2014 Conference*, 2014, pp. 1646–1654.

[9] H.R. Feyzmahdavian, A. Aytekin, and M. Johansson, *A Delayed Proximal Gradient Method with Linear Convergence Rate*, in *Proceedings of the 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[10] M.P. Friedlander and M. Schmidt, *Hybrid Deterministic–Stochastic Methods for Data Fitting*, SIAM Journal on Scientific Computing 34 (2012), pp. A1380–A1405.

[11] P. Gong and J. Ye, *Linear Convergence of Variance–Reduced Stochastic Gradient without Strong Convexity*, 2014. manuscript, available at `http://arxiv.org/abs/1406.1102`.

[12] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo, *Convergence Rate of Incremental Aggregated Gradient Algorithms*, 2015. manuscript, available at `http://arxiv.org/abs/1506.02081`.

[13] A.J. Hoffman, *On Approximate Solutions of Systems of Linear Inequalities*, Journal of Research of the National Bureau of Standards 49 (1952), pp. 263–265.

[14] R. Johnson and T. Zhang, *Accelerating Stochatic Gradient Descent Using Predictive Variance Reduction*, in *Advances in Neural Information Processing Systems 26: Proceedings of the 2013 Conference*, 2013, pp. 315–323.

[15] J. Konečný, J. Liu, P. Richtárik, and M. Takáč, *Mini–Batch Semi–Stochastic Gradient Descent in the Proximal Setting*, IEEE Journal of Selected Topics in Signal Processing 10 (2016), pp. 242–255.

[16] N. Le Roux, M. Schmidt, and F. Bach, *A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets*, in *Advances in Neural Information Processing Systems 25: Proceedings of the 2012 Conference*, 2012, pp. 2672–2680.

[17] E.S. Levitin and B.T. Polyak, *Constrained Minimization Methods*, USSR Computational Mathematics and Mathematical Physics 6 (1966), pp. 1–50.

[18] W. Li, *Remarks on Convergence of the Matrix Splitting Algorithm for the Symmetric Linear Complementarity Problem*, SIAM Journal on Optimization 3 (1993), pp. 155–163.

[19] W. Li, *The Sharp Lipschitz Constants for Feasible and Optimal Solutions of a Perturbed Linear Program*, Linear Algebra and Its Applications 187 (1993), pp. 15–40.

[20] Z.Q. Luo and P. Tseng, *On the Linear Convergence of Descent Methods for Convex Essentially Smooth Minimization*, SIAM Journal on Control and Optimization 30 (1992), pp. 408–425.

[21] Z.Q. Luo and P. Tseng, *Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach*, Annals of Operations Research 46 (1993), pp. 157–178.

[22] C. Ma, R. Tappenden, and M. Takáč, *Linear Convergence of the Randomized Feasible Descent Method under the Weak Strong Convexity Assumption*, Journal of Machine Learning Research 17 (2016), pp. 1–24.

[23] 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.

[24] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Boston, 2004.

[25] A. Rakhlin, O. Shamir, and K. Sridharan, *Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization*, in *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, 2012, pp. 449–456.

[26] S.J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola, *On Variance Reduction in Stochastic Gradient Descent and Its Asynchronous Variants*, in *Advances in Neural Information Processing Systems 28: Proceedings of the 2015 Conference*, 2015, pp. 2647–2655.

[27] M. Schmidt, N. Le Roux, and F. Bach, *Convergence Rates of Inexact Proximal–Gradient Methods for Convex Optimization*, in *Advances in Neural Information Processing Systems 24: Proceedings of*

the 2011 Conference, 2011, pp. 1458–1466.

[28] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *Pegasos: Primal Estimated Sub–Gradient Solver for SVM*, Mathematical Programming, Series B 127 (2011), pp. 3–30.

[29] P. Tseng, *Descent Methods for Convex Essentially Smooth Minimization*, Journal of Optimization Theory and Applications 71 (1991), pp. 425–463.

[30] P. Tseng and S. Yun, *Incrementally Updated Gradient Methods for Constrained and Regularized Optimization*, Journal of Optimization Theory and Applications 160 (2014), pp. 832–853.

[31] P.W. Wang and C.J. Lin, *Iteration Complexity of Feasible Descent Methods for Convex Optimization*, Journal of Machine Learning Research 15 (2014), pp. 1523–1548.

[32] L. Xiao and T. Zhang, *A Proximal Stochastic Gradient Method with Progressive Variance Reduction*, SIAM Journal on Optimization 24 (2014), pp. 2057–2075.

[33] S. Zhang, *Global Error Bounds for Convex Conic Problems*, SIAM Journal on Optimization 10 (2000), pp. 836–851.

[34] Z. Zhou and A.M.C. So, *A Unified Approach to Error Bounds for Structured Convex Optimization Problems*, 2016. accepted for publication in *Mathematical Programming, Series A*.

[35] Z. Zhou, Q. Zhang, and A.M.C. So, *$\ell_{1,p}$–Norm Regularization: Error Bounds and Convergence Rate Analysis of First–Order Methods*, in *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015, pp. 1501–1510.

Appendix

Appendix A. Proof of Proposition 2

Since ∇f is L–Lipschitz continuous, we have

$$f(x^{k+1}) - f(x^k) \leq \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2;$$

see, e.g., [24, Theorem 2.1.5]. Using (3) and the fact that $\alpha_k = \alpha$ for all $k \geq 0$, we obtain

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq \frac{L}{2}\|x^{k+1} - x^k\|_2^2 + \left(\frac{1}{\alpha}(x^k - x^{k+1}) - e^{k+1}\right)^T (x^{k+1} - x^k) \\ &\leq \left(\frac{L}{2} - \frac{1}{\alpha}\right)\|x^{k+1} - x^k\|_2^2 + \|e^{k+1}\|_2\|x^{k+1} - x^k\|_2, \end{aligned}$$

as desired.

Appendix B. Proof of Corollary 1

By Proposition 2, we have

$$f(x^k) - f(x^{k+1}) \geq \gamma\left(\|x^k - x^{k+1}\|_2 - \frac{1}{2\gamma}\|e^{k+1}\|_2\right)^2 - \frac{1}{4\gamma}\|e^{k+1}\|_2^2. \tag{B1}$$

Using the identity $(a+b)^2 \leq 2(a^2+b^2)$, which is valid for all $a, b \in \mathbb{R}$, we have

$$\begin{aligned} \|x^k - x^{k+1}\|_2^2 &= \left(\|x^k - x^{k+1}\|_2 - \frac{1}{2\gamma}\|e^{k+1}\|_2 + \frac{1}{2\gamma}\|e^{k+1}\|_2\right)^2 \\ &\leq 2\left[\left(\|x^k - x^{k+1}\|_2 - \frac{1}{2\gamma}\|e^{k+1}\|_2\right)^2 + \frac{1}{4\gamma^2}\|e^{k+1}\|_2^2\right]. \end{aligned} \tag{B2}$$

Upon combining (B1) and (B2), we obtain (a). Now, observe from (B1) that

$$f(x^k) - f(x^{k+1}) = f(x^k) - f_{\min} + f_{\min} - f(x^{k+1}) \geq -\frac{1}{4\gamma}\|e^{k+1}\|_2^2.$$

Upon rearranging, we obtain (b).

Appendix C. Proof of Proposition 3

We begin with the following result, which is known as the Hoffman error bound:

FACT 2 *(cf. [13]) Let $C \in \mathbb{R}^{m\times n}$ and $s \in \mathbb{R}^m$ be given. Suppose that the linear system*

$$Cu = s \tag{C1}$$

in $u \in \mathbb{R}^n$ is feasible. Then, there exists a $\theta > 0$, which depends only on C, such that for any $x \in \mathbb{R}^n$, there exists an $\bar{x} \in \mathbb{R}^n$ satisfying (C1) and

$$\|x - \bar{x}\|_2 \leq \theta \|Cx - s\|_2.$$

The quantity $\theta > 0$ in Fact 2 is known as the *Hoffman constant*, for which sharp estimates are known. We refer the reader to [19, 33] for details.

To prove Proposition 3, consider the following linear system in $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$:

$$\begin{aligned} v &= u - E^T \nabla g(t^*) - q, \\ Eu &= t^*, \\ u &= v. \end{aligned} \tag{C2}$$

Note that $(\bar{x}, \bar{x}) \in \mathbb{R}^d \times \mathbb{R}^d$ is feasible for (C2) if and only if $\bar{x} \in \mathcal{X}$. Thus, it follows from Assumption 2 that (C2) is feasible. Now, let $z = x - \nabla f(x) = x - E^T \nabla g(Ex) - q$. By Fact 2, there exist a constant $\theta > 0$ and a feasible solution (x^*, z^*) to (C2) such that

$$\|(x, z) - (x^*, z^*)\|_2 \leq \theta \Big[\|Ex - t^*\|_2 + \|\nabla f(x)\|_2 + \big\|E^T \nabla g(Ex) - E^T \nabla g(t^*)\big\|_2 \Big].$$

Since $\|E^T \nabla g(Ex) - E^T \nabla g(t^*)\|_2 \leq L \cdot \|E\| \cdot \|Ex - t^*\|_2$, the desired result follows by setting $\omega = \theta(1 + L\|E\|)$.

Appendix D. Proof of Corollary 2

(a) By the assumption on $\{B_k\}_{k \geq 1}$, we have

$$\sum_{j=1}^{k} \mu^{k-j} \|e^j\|_2^2 \leq \sum_{j=1}^{k} \mu^{k-j} O\left(\frac{1}{j^{1+\rho}}\right) \tag{D1}$$

for all $k \geq 1$. To bound the quantity on the right–hand side, let us define

$$S_k = \sum_{j=1}^{k} \frac{\mu^{k-j}}{j^{1+\rho}} \quad \text{for } k = 1, 2, \ldots.$$

Let $K \equiv K(\mu, \rho) > 0$ be such that $\mu' = \mu \left(1 + \frac{1}{k}\right)^{1+\rho} < 1$ for all $k \geq K$, and let $C \equiv C(\mu, \rho) \geq (1 - \mu')^{-1}$ be such that $S_k \leq C k^{-(1+\rho)}$ for $k = 1, 2, \ldots, K$. We now show by induction that

$$S_k \leq C k^{-(1+\rho)} \quad \text{for all } k \geq 1. \tag{D2}$$

The statement is trivially true for $k = 1, 2, \ldots, K$. For $k > K$, the inductive hypothesis and our choice of C imply that

$$S_{k+1} = \mu S_k + \frac{1}{(k+1)^{1+\rho}} \leq \left[1 + C\mu \left(1 + \frac{1}{k}\right)^{1+\rho}\right] \frac{1}{(k+1)^{1+\rho}} \leq \frac{C}{(k+1)^{1+\rho}}.$$

This completes the inductive step.

Now, using (19), (D1), (D2), and Proposition 4, we have

$$\begin{aligned} f(x^{k+1}) - f_{\min} &\leq O(\mu^{k+1}) + O(S_{k+1}) \leq O\left(\frac{1}{(k+1)^{1+\rho}}\right), \\ \operatorname{dist}(x^k, \mathcal{X})^2 &\leq O\left(S_{k+1} + \left(\frac{1+\mu}{2}\right)^k\right) = O\left(\frac{1}{(k+1)^{1+\rho}}\right) \end{aligned}$$

for all $k \geq 0$. This completes the proof of (a).

(b) The assumption on $\{B_k\}_{k\geq 1}$ implies that

$$\sum_{j=1}^{k} \mu^{k-j} \|e^j\|_2^2 \leq \sum_{j=1}^{k} \mu^{k-j} O(\rho^j) \leq O(k c_1^k) \leq O(c_2^k)$$

for all $k \geq 1$, where $c_1 = \max\{\mu, \rho\} \in (0,1)$ and $c_2 = \frac{1+c_1}{2} \in (c_1, 1)$. Hence, by (19) and Proposition 4, we have

$$\begin{aligned} f(x^{k+1}) - f_{\min} &\leq O(\mu^{k+1}) + O(c_2^{k+1}) = O(c_2^{k+1}), \\ \operatorname{dist}(x^k, \mathcal{X})^2 &\leq O\left(c_2^{k+1} + \left(\frac{1+\mu}{2}\right)^k\right) \leq O(c_2^k) \end{aligned}$$

for all $k \geq 0$. The desired result then follows by setting $c = \sqrt{c_2} \in (0,1)$.

Appendix E. Proof of Corollary 3

By Theorem 2, we have

$$\mathbb{E}\left[f(x^k) - f_{\min}\right] \leq \mu^k \left(f(x^0) - f_{\min}\right) + \delta \sum_{j=1}^{k} \mu^{k-j} \mathbb{E}\left[\|e^j\|_2^2\right],$$

for all $k \geq 0$. Upon noting

$$\mathbb{E}\left[\operatorname{dist}(x^k, \mathcal{X})\right] \leq \left(\mathbb{E}\left[\operatorname{dist}(x^k, \mathcal{X})^2\right]\right)^{1/2}$$

and using the assumption that $\mathbb{E}\left[\|e^k\|_2^2\right] \leq B_k$, the rest of the proof is essentially the same as that of Corollary 2.

Appendix F. Proof of Proposition 6

(a) Recall that for (LSR), we have $f_i(x) = (a_i^T x - b_i)^2$ for $i = 1, \ldots, n$. Since $\max_{1 \le i \le n} \|a_i\|_2^2 \le L$ by Proposition 5, we use (20) to compute

$$\begin{aligned}
\|e^{k+1}\|_2^2 &\le \left(\frac{n - |I_k|}{n|I_k|} \sum_{i \in I_k} \|\nabla f_i(x^k)\|_2 + \frac{1}{n} \sum_{i \in \mathcal{N} \setminus I_k} \|\nabla f_i(x^k)\|_2 \right)^2 \\
&\le \left(\frac{n - |I_k|}{n} \right)^2 \left[\sqrt{\frac{1}{|I_k|} \sum_{i \in I_k} \|\nabla f_i(x^k)\|_2^2} + \sqrt{\frac{1}{n - |I_k|} \sum_{i \in \mathcal{N} \setminus I_k} \|\nabla f_i(x^k)\|_2^2} \right]^2 \\
&\le 8 \left(\frac{n - |I_k|}{n} \right)^2 \left(\frac{1}{|I_k|} \sum_{i \in I_k} \left(a_i^T x^k - b_i \right)^2 \cdot \|a_i\|_2^2 \right. \\
&\qquad \left. + \frac{1}{n - |I_k|} \sum_{i \in \mathcal{N} \setminus I_k} \left(a_i^T x^k - b_i \right)^2 \cdot \|a_i\|_2^2 \right) \\
&\le 8L \left(\frac{n - |I_k|}{n} \right)^2 \left[\frac{1}{|I_k|} \sum_{i \in I_k} \left(a_i^T x^k - b_i \right)^2 + \frac{1}{n - |I_k|} \sum_{i \in \mathcal{N} \setminus I_k} \left(a_i^T x^k - b_i \right)^2 \right] \quad \text{(F1)}
\end{aligned}$$

where the second inequality follows from the concavity of $x \mapsto \sqrt{x}$ and Jensen's inequality; the third inequality follows from the identity $(a+b)^2 \le 2(a^2 + b^2)$, which is valid for all $a, b \in \mathbb{R}$, and

$$\nabla f_i(x) = 2 \left(a_i^T x - b_i \right) a_i \quad \text{for } i = 1, 2, \ldots, n.$$

Now, observe that for $\frac{n}{2} \le |I_k| \le n$, we have

$$\begin{aligned}
&\frac{1}{|I_k|} \sum_{i \in I_k} \left(a_i^T x^k - b_i \right)^2 + \frac{1}{n - |I_k|} \sum_{i \in \mathcal{N} \setminus I_k} \left(a_i^T x^k - b_i \right)^2 \\
&= \frac{n}{n - |I_k|} f(x^k) + \left(\frac{1}{|I_k|} - \frac{1}{n - |I_k|} \right) \sum_{i \in I_k} \left(a_i^T x^k - b_i \right)^2 \\
&\le \frac{n}{n - |I_k|} f(x^k). \quad \text{(F2)}
\end{aligned}$$

The desired result then follows from (F1) and (F2).

(b) Using (21) and $\max_{1\leq i\leq n}\|a_i\|_2^2 \leq L$, we have

$$\begin{aligned}
\mathbb{E}\left[\|e^{k+1}\|_2^2 \,\middle|\, \mathscr{F}_k\right] &\leq \frac{n-|I_k|}{n|I_k|}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\|\nabla f_i(x^k)\|_2 + \|\nabla f(x^k)\|_2\right)^2\right] \\
&\leq \frac{n-|I_k|}{n|I_k|}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(2\sqrt{L}\left|a_i^T x^k - b_i\right| + \frac{1}{n}\sum_{j=1}^{n}\left\|\nabla f_j(x^k)\right\|_2\right)^2\right] \\
&\leq \frac{n-|I_k|}{n|I_k|}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(2\sqrt{L}\left|a_i^T x^k - b_i\right| + \sqrt{\frac{1}{n}\sum_{j=1}^{n}\|\nabla f_j(x^k)\|_2^2}\right)^2\right] \\
&\leq \frac{n-|I_k|}{n|I_k|}\left[\frac{8L}{n-1}\sum_{i=1}^{n}\left(\left(a_i^T x^k - b_i\right)^2 + \frac{1}{n}\sum_{j=1}^{n}\left(a_j^T x^k - b_j\right)^2\right)\right] \\
&= 16L\frac{n-|I_k|}{(n-1)|I_k|}f(x^k).
\end{aligned}$$

It follows from the tower property of conditional expectation that for all $k \geq 0$,

$$\mathbb{E}\left[\|e^{k+1}\|_2^2\right] \leq 16L\frac{n-|I_k|}{(n-1)|I_k|}\mathbb{E}\left[f(x^k)\right],$$

as desired.

Appendix G. Proof of Proposition 7

Recall that for the least squares regression problem (LSR), the premises of Theorem 2 hold. Moreover, using (26), we have $\mathbb{E}\left[e^{k+1} \mid \mathscr{F}_k\right] = \mathbf{0}$. Thus, by conditioning on $\mathscr{F}_k$ and taking expectation with respect to e^{k+1} on both sides of (16), we obtain

$$\mathbb{E}\left[f(x^{k+1}) - f_{\min} \,\middle|\, \mathscr{F}_k\right] \leq f(x^k) - f_{\min} - \alpha(1-\frac{\alpha L}{2})\|\nabla f(x^k)\|_2^2 + \frac{\alpha^2 L}{2}\mathbb{E}\left[\|e^{k+1}\|_2^2 \,\middle|\, \mathscr{F}_k\right].$$

Combining the above with (18), we obtain (27).

Next, let x^* be an arbitrary point in the optimal solution set $\mathcal{X}$. Since $\max_{1\leq i\leq n}\|a_i\|_2^2 \leq L$ and $f_i(x) = (a_i^T x - b_i)^2$, ∇f_i is Lipschitz continuous with parameter L for $i = 1, \ldots, n$. By (11), we have

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|_2^2 \leq 2L\left[f_i(x) - f_i(x^*) - (x-x^*)^T\nabla f_i(x^*)\right]$$

for all $x \in \mathbb{R}^d$ and $i = 1, \ldots, n$. Summing the above inequality over $i = 1, \ldots, n$ and using the fact that $\nabla f(x^*) = \mathbf{0}$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(x^*)\|_2^2 \leq 2L(f(x) - f_{\min}). \tag{G1}$$

Since i_k is uniformly sampled from $\{1, 2, \ldots, n\}$, it follows that

$$\mathbb{E}\left[\nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\nabla f_i(x^k) - \nabla f_i(\omega)\right],$$

where the expectation is taken with respect to i_k. Substituting the above into (26), we have

$$\begin{aligned}
\mathbb{E}\left[\|e^{k+1}\|_2^2 \,\middle|\, \mathscr{F}_k\right] &= \mathbb{E}\left[\left\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega) - \mathbb{E}\left[\nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega)\right]\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(\omega)\|_2^2\right] \\
&\leq 2\left(\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2\right] + \mathbb{E}\left[\|\nabla f_{i_k}(\omega) - \nabla f_{i_k}(x^*)\|_2^2\right]\right) \\
&\leq 4L\left(f(x^k) - f_{\min} + f(\omega) - f_{\min}\right),
\end{aligned}$$

where the first inequality is due to the fact that for any random variable $v \in \mathbb{R}^n$, we have

$$\mathbb{E}\left[\|v - \mathbb{E}[v]\|_2^2\right] = \mathbb{E}\left[\|v\|_2^2\right] - (\mathbb{E}[v])^2 \leq \mathbb{E}\left[\|v\|_2^2\right],$$

and the third inequality is by (G1). This establishes (28) and completes the proof.

Appendix H. Proof of Proposition 8

Using (34), we bound

$$\begin{aligned}
\|e^{k+1}\|_2^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^{\pi_i^k}) - \nabla f_i(x^k)\|_2^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\nabla f_j(x^{\pi_i^k}) - \nabla f_j(x^k)\|_2^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\nabla f_j(x^{\pi_i^k}) - \nabla f_j(x^*)\|_2^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\nabla f_j(x^k) - \nabla f_j(x^*)\|_2^2, \qquad \text{(H1)}
\end{aligned}$$

where the first and third inequalities are due to the fact that $\left(\sum_{i=1}^m a_i\right)^2 \leq m\sum_{i=1}^m a_i^2$ for any $a_1, \ldots, a_m \in \mathbb{R}$. Since $\max_{1\leq i\leq n}\|a_i\|_2^2 \leq L$ and $f_i(x) = (a_i^T x - b_i)^2$, both ∇f and the ∇f_i's are Lipschitz continuous with parameter L. Hence, by letting x^* to be an

arbitrary point in $\mathcal{X}$ and using the same arguments as those for (G1), we have

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\|\nabla f_j(x^{\pi_i^k})-\nabla f_j(x^*)\|_2^2 &\leq 2L(f(x^{\pi_i^k})-f_{\min}),\\ \frac{1}{n}\sum_{j=1}^{n}\|\nabla f_j(x^k)-\nabla f_j(x^*)\|_2^2 &\leq 2L(f(x^k)-f_{\min}).\end{aligned}$$

Combining the above two inequalities with (H1), we obtain

$$\|e^{k+1}\|_2^2 \leq 4L\sum_{i=1}^{n}(f(x^{\pi_i^k})-f_{\min}+f(x^k)-f_{\min}). \tag{H2}$$

In addition, since the index vectors $\{\pi^k\}_{k\geq 0}$ obey the bounded delay rule (31), we have

$$f(x^{\pi_i^k})-f_{\min}+f(x^k)-f_{\min}\leq 2\max_{(k-K)_+\leq l\leq k}(f(x^l)-f_{\min}) \quad \text{for } i=1,2,\ldots,n.$$

The desired result then follows by substituting the above into (H2).