Variance-Reduced Accelerated First-order Methods: Central Limit Theorems and Confidence Statements

Jinlong Lei and Uday V. Shanbhag

Abstract

In this paper, we consider a strongly convex stochastic optimization problem and propose three classes of variable sample-size stochastic first-order methods: (i) the standard stochastic gradient descent method; (ii) its accelerated variant; and (iii) the stochastic heavy ball method. In each scheme, the exact gradients are approximated by averaging across an increasing batch size of sampled gradients. We prove that when the sample-size increases geometrically, the generated estimates converge in mean to the optimal solution at a geometric rate for schemes (i) – (iii). Based on this result, we provide central limit statements whereby it is shown that the rescaled estimation errors converge in distribution to a normal distribution with the associated covariance matrix depending on the Hessian matrix, covariance of the gradient noise, and the steplength. If the sample-size increases at a polynomial rate, we show that the estimation errors decay at a corresponding polynomial rate and establish the corresponding central limit theorems (CLTs). Under some conditions, we discuss how both the algorithms and the associated limit theorems may be extended to constrained and nonsmooth regimes. Finally, we provide an avenue to construct confidence regions for the optimal solution based on the established CLTs and test the theoretical findings on a stochastic parameter estimation problem.

1 Introduction

In this paper, we consider the strongly convex optimization problem (1):

\[
\min_{x \in \mathbb{R}^m} f(x) \triangleq \mathbb{E}[F(x, \xi)],
\]

where \( \xi : \Omega \rightarrow \mathbb{R}^d \) is a random variable defined on the probability space \((\Omega, \mathcal{F}, \mathbb{P})\), \( F : \mathbb{R}^m \times \mathbb{R}^d \rightarrow \mathbb{R} \), and the expectation is taken over the distribution of the random vector. Stochastic optimization problems have been extensively studied given wide applicability in almost all areas of science and engineering, ranging from communication and queueing systems to finance (cf. [4, 54]). However, in most situations, this

\[\text{Email: leijinlong@tongji.edu.cn (Jinlong Lei), udaybag@psu.edu (Uday V. Shanbhag). The work has been partly supported by NSF grant 1538605 and 1246887 (CAREER).}
\]

\[\text{†Lei is the Department of Control Science and Engineering, Tongji University; She is also with the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China. Shanbhag is with the Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802, USA}\]
expectation and its derivative are unavailable in closed form requiring the development of sampling approaches. Sample average approximation (SAA) [54] and stochastic approximation (SA) represent two commonly used approaches for contending with stochastic programs. Our focus is on SA schemes, first considered by Robbins and Monro [50] for seeking roots of a regression function with noisy observations. The standard SA algorithm \( x_{k+1} = x_k - \alpha_k \nabla F(x_k, \xi_k) \), also known as the stochastic gradient descent (SGD) algorithm, updates the estimate \( x_{k+1} \) based on a sampled gradient \( \nabla F(x_k, \xi_k) \). The convergence analysis usually requires suitable properties on the gradient map (such as Lipschitzian requirements) and the steplength sequence (such as non-summable but square summable). The almost sure convergence of \( x_k \) to \( x^* \), the unique optimal solution of (1), was established in [5, 8, 12] on the basis of the Robbins-Siegmund theorem [51] while ODE techniques were employed for claiming similar statements in [7, 24, 31]. In addition, in [11], the authors developed a statistical diagnostic test to detect the phase transition, during which the iterative procedure converges towards a region of interest, in the context of SGD with constant learning rate. Tight bounds on the rate of convergence can be obtained by establishing the asymptotic distribution for the iterates (cf. [7, 12, 19, 24, 31]). An instructive review of results up to around 2010 is provided by [46], while asymptotic normality of the suitably scaled iterates for SA with decreasing step-sizes has been proven in [19, 24, 31]. To be specific, the rescaled error process \( \sqrt{\alpha_k} (x_k - x^*) \) asymptotically converges in distribution to a normal distribution with zero mean and with covariance depending on the Hessian matrix, the covariance of the gradient noise, and the steplength. Prior works on CLTs of standard stochastic approximation for smooth convex optimization can be traced to the seminal averaging paper by Polyak and Juditsky [49]. The asymptotic normality was further investigated in [12] for SA with expanding truncations, an avenue that does not require Lipschitz continuity of the gradients. The sequence of iterates generated by the constant steplength SA scheme is shown to be a homogeneous Markov chain with a unique stationary distribution; see [31, Chapter 9], [38, Chapter 17], and [18].

CLTs for SA schemes are significant from the standpoint of algorithm design as well as inference.

(i) Algorithm design. Indeed, the optimal selection of the steplength depends on the Hessian matrix at \( x^* \), see e.g., [12, Chapter 3.4]. Since the Hessian at \( x^* \) is unavailable, the optimal value of \( \alpha_k \) can only be estimated, which has led to the development of adaptive SA methods (cf. [55, 61, 62]). Motivated by the heavy dependence of SA schemes on steplength choices, [64] developed a self-tuned rule that adapts the steplength sequence to problem parameters. In a similar vein, there has been an effort to develop optimal constant steplengths. Specifically, it was shown in [39] that with suitably selected constant step-sizes, the expected function values at the averaged iterate converge to the optimum with rate \( O(1/\sqrt{k}) \) in merely convex case and rate \( O(1/k) \) in strongly convex case, matching the lower bounds [40]. It is further shown in [37] that the constant SGD simulates a Markov chain with a stationary distribution, which might be used to adjust the tuning parameters of constant SGD so as to improve the convergence rate, e.g., SGD for fitting the generalized linear models [59] and approximating the Bayesian posterior inference [37].

(ii) Confidence statements. Furthermore, CLTs for SA schemes might allow for the possibility of constructing confidence regions for the optimal solution (e.g. [3, 13, 26]). In particular, Hsieh and Glynn [26] designed an approach to rigorously characterize confidence regions without explicitly estimating the covariance of the limiting normal distribution, while Chen et al. [13] proposed a plug-in estimator and a batch-means estima-
tor for the asymptotic covariance of the average iterate from SGD. In addition, Su and Zhu [56] advocated performing SGD updates for a while and then splitting the single thread into several threads; notably, they proceeded to construct \( t \)-based confidence interval that can attain asymptotically exact coverage probability. The monograph [31] has extensively investigated CLTs for SA schemes under both constant and decreasing steplengths. While in the context of SAA schemes, there has been some recent work on developing confidence statements for stochastic optimization [54] and stochastic variational inequality problems [32, 36].

Unfortunately, SA schemes with diminishing steps cannot recover the deterministic convergence rates seen in exact gradient methods while constant steplength SA schemes are only characterized by convergence guarantees to a neighborhood of the optimal solution. Variance-reduction schemes employing an increasing batch-size of sampled gradients (instead of the unavailable true gradient) appear to have been first alluded to in [2, 17, 25, 44] and analyzed in smooth and strongly convex [10, 20, 52, 53], smooth convex regimes [22], nonsmooth (but smoothable) convex [27], nonconvex [33], and game-theoretic [34] regimes. Notably, the linear rates in mean-squared error were derived for strongly convex smooth [53] and a subclass of nonsmooth objectives [27], while a rate \( O(1/k^2) \) and \( O(1/k) \) was obtained for expected sub-optimality in convex smooth [22, 28] and nonsmooth [27], respectively. In each instance, the schemes achieve the corresponding optimal deterministic rates of convergence under suitable growth in sample-size sequences while in almost all cases, the optimal sample-complexity bounds were obtained. An excellent discussion relating sampling rates and the canonical rate (i.e. the Monte-Carlo rate) in stochastic gradient-based schemes has been provided by Pasupathy et al. [45].

**Gaps and motivation.** This paper is motivated by the following gaps. (i) Despite a surge of interest in variance-reduced schemes (with and without acceleration), no limit theorems are available for claiming asymptotic normality of the limiting scaled sequence in either unaccelerated or Nesterov accelerated regimes coupled with variance reduction. In particular, we have little understanding regarding whether such avenues have detrimental impacts in terms of the limiting behavior. Such statements are also unavailable for the related heavy-ball method. (ii) The geometric growth in sample-size required for achieving linear rates of convergence may prove onerous in some settings. Are CLTs available in settings where the sample-size grows at a polynomial rate? (iii) Finally, given a CLT, there is little by way of availability of rigorous confidence statements, barring the work by Hsieh and Glynn [26]. Can such statements be developed for the proposed class of variance-reduced first-order methods?

**Justification and relationship to other variance-reduced schemes.**

(i) **Terminology and applicability.** The moniker “variance-reduced” has been loosely used in stochastic optimization to capture approaches that admit deterministic convergence rates. For instance, for minimizing finite-sum expected-residual minimization problems in machine learning, techniques such as SVRG [29] and SAGA [15] achieve deterministic rates of convergence. In general probability spaces, using increasing batch-sizes of gradients represent progressively accurate approximations of the true gradient, as opposed to noisy sampled variants employed in single sample schemes [20]. The resulting schemes, often referred to as mini-batch SA schemes, may achieve deterministic rates of convergence under suitable growth rates in
sample-sizes.

(ii) **Weaker assumptions, and stronger statements.** The proposed variance-reduced framework has several key benefits often unavailable in single-sample regimes: i) Under suitable assumptions, such schemes often achieve optimal deterministic rates in terms of iteration complexity (e.g. in constrained regimes, complexity of an iteration is essentially that of computing a projection onto a convex set) while achieving near-optimal sample complexity. This assumes profound importance when iterations are expensive as in high-dimensional and nonlinear settings; ii) In addition, the benefits of acceleration are less clear in stochastic single-sample regimes, absent such approaches (to the best of our knowledge). While the present treatment is in an unconstrained regime, we also discuss how extensions to constrained and nonsmooth regimes may be developed.

(iii) **Sampling requirements.** Variance-reduced schemes have obvious benefits when sampling is relatively cheap compared to computational effort of an iteration. While such schemes are often characterized by near-optimal sample-complexity, the proposed schemes are near optimal, one might question how to contend with batch-sizes (denoted by \( N_k \)) tending to \(+\infty\). This issue is somewhat of a red herring since most SA schemes are meant to provide \( \epsilon \)-approximations. For instance, if \( \epsilon = 1e-3 \), then such a scheme requires approximately \( \mathcal{O}(\log(\frac{1}{\epsilon})) \) or \( \mathcal{O}(\log(10^3)) \) steps in strongly convex regimes. Since \( N_k \approx \lceil \rho^{-k} \rceil \) where \( \rho < 1 \), it can be observed the sampling burden even towards the final iterates is not significant. With the ubiquity of multi-core architecture, such requirements are not terribly onerous.

**Outline and contributions.** To address these gaps, we present CLTs and confidence statements for first-order stochastic variance-reduced algorithms for resolving (1), including the classical SGD [50], the stochastic variants of the Nesterov’s accelerated method [42], and the heavy ball method [47]. We provide statements when batch-sizes increases at either a geometric or a polynomial rate. Our main contributions are summarized next.

(I) **CLTs for variable sample-size gradient methods.** In Section 2.1, we recall the variance-reduced (VR) stochastic gradient algorithm with a constant steplength (see Algorithm 1), where the gradient is estimated by the average of an increasing batch of sampled gradients. In Section 3, when the batch-size increases at a geometric rate, the mean-squared error diminishes at a geometric rate (see Proposition 1) and provide a preliminary Lemma 6 for establishing a CLT for a noised-corrupted linear recursion. Based on this Lemma and the linear approximation of the gradient function at \( x^* \), we may derive a CLT (see Theorem 1) in this setting. We proceed to show that the covariance of the limiting normal distribution depends on the Hessian matrix at the solution, the condition number, the covariance of gradient noise, etc and the steplength. Additionally, we show in Section 4 that when the batch-size is increased at a polynomial rate, the sequence of iterates converge at a corresponding polynomial rate (see Proposition 4). Then based on the CLT shown in Lemma 9 for the time-varying linear recursion, the CLT for Algorithm 1 with polynomially increasing batch-size is established in Theorem 4.

(II) **CLTs for VR-accelerated gradient method.** In Section 2.2, we consider a VR-accelerated gradient algorithm with constant steplengths (see Algorithm 2) and by leveraging the geometric rate of convergence (Proposition 2), we establish amongst the first CLTs in accelerated regimes (Theorem 2) when the batch-size increases at a geometric rate. It is well-known that the accelerated variant has better constants in the iteration
complexity in deterministic regimes and this is also seen in stochastic settings. Akin to earlier, when the batch-size increases at a polynomial rate, a polynomial convergence rate and the corresponding CLT are in Proposition 5 and Theorem 5.

III. CLTs for VR-heavy-ball schemes. In Section 2.3, we design a VR-heavy ball scheme with constant steplengths (see Algorithm 3). In contrast with Algorithms 1 and 2, the heavy ball method imposes a twice-differentiability assumption on \( f \). When the batch-size increases geometrically (or polynomially), a geometric (or polynomial) rate of convergence is shown in Proposition 3 (or Proposition 6), while a CLT is established in Theorem 3 (or Theorem 6).

IV. Confidence statements. In Section 5, inspired by [26], we provide rigorous confidence regions for the optimal solution and function value. Then in Section 6, we implement some simulations on a parameter estimation problem in the stochastic environment to validate the theoretical findings.

Notation. Let \( I_m \) denote the identity matrix of dimension \( m \) and \( 0_m \in \mathbb{R}^{m \times m} \) denote the matrix with all entries equal zero. Let \{\( X_k \)\} be a sequence of random variables. \( X_k \xrightarrow{d}{k \to \infty} N(0, S) \) denotes that \( X_k \) converges in distribution to a normal distribution \( N(0, S) \) with mean 0 and covariance \( S \), and \( X_k \xrightarrow{P}{k \to \infty} X \) denotes that \( X_k \) converges in probability to \( X \). Following the definitions of stochastic \( o_P(\cdot) \) and \( O_P(\cdot) \) symbols in [60], the notation \( e_k \) is \( o_P(1) \) implies that a sequence of random vectors \{\( e_k \)\} converges to zero in probability. Similarly, the expression \( e_k \) is \( O_P(1) \) implies that a sequence \{\( e_k \)\} is bounded in probability.

2 First-Order Variable Sample-size Stochastic Algorithms.

Since the exact gradient \( \nabla f(x) \) is expectation-valued and unavailable in a closed form, we assume there exists a stochastic first-order oracle such that for any \( x, \xi \), a sampled gradient \( \nabla F(x, \xi) \) is returned, assumed to be an unbiased estimator of \( \nabla f(x) \). In this section, we present three first-order stochastic algorithms to find the optimal solution to (1). Throughout the paper, time is slotted at \( k = 0, 1, 2, \ldots \) and an iterate at time \( k \) is denoted by \( x_k \in \mathbb{R}^m \).

2.1 Gradient Descent Method.

We recall a variable sample-size stochastic gradient algorithm (Algorithm 1) to solve (1), where at iteration \( k \), the unavailable exact gradient \( \nabla f(x_k) \) is estimated via the average of an increasing batch-size of sampled gradients.
Algorithm 1 Variance reduced SGD

Given an arbitrary initial value \( x_0 \in \mathbb{R}^m \), a positive constant \( \alpha > 0 \), and a positive integer sequence \( \{N_k\}_{k \geq 0} \). Then iterate the following equation for \( k \geq 0 \).

\[
x_{k+1} = x_k - \alpha \frac{\sum_{j=1}^{N_k} \nabla F(x_k, \xi_{j,k})}{N_k},
\]

where \( \alpha > 0 \) is the constant steplength, \( N_k \) is the number of sampled gradients used at time \( k \), and \( \xi_{j,k}, j = 1, \cdots, N_k \), denote the independent and identically distributed (i.i.d.) realizations of \( \xi \).

\[
\text{Algorithm 1:}
\]

If the gradient observation noise \( w_{k,N_k} \) is defined as

\[
w_{k,N_k} \triangleq \frac{\sum_{j=1}^{N_k} \nabla F(x_k, \xi_{j,k})}{N_k} - \nabla f(x_k),
\]

then the update (2) can be rewritten as follows:

\[
x_{k+1} = x_k - \alpha (\nabla f(x_k) + w_{k,N_k}).
\]

Define \( F_k \triangleq \sigma \{x_0, \xi_{j,t}, 1 \leq j \leq N_t, 0 \leq t \leq k-1\} \). Then \( x_k \) is adapted to \( F_k \) by Algorithm 1. We impose the following conditions on the objective function, the conditional expectation and the second moments of the sampled gradients produced by the stochastic first-order oracle.

Assumption 1 (i) \( f(\cdot) \) is continuously differentiable in \( x \in \mathbb{R}^m \) with Lipschitz continuous gradient, i.e., there exists a constant \( L > 0 \) such that \( \|\nabla f(x) - \nabla f(x')\| \leq L \|x - x'\| \) for any \( x, x' \in \mathbb{R}^m \). (ii) \( f(\cdot) \) is \( \eta \)-strongly convex, i.e., \( (\nabla f(x) - \nabla f(x'))^T (x - x') \geq \eta \|x - x'\|^2 \) for any \( x, x' \in \mathbb{R}^m \).

(iii) There exists a constant \( \nu > 0 \) such that for any \( k \geq 0 \) and \( j = 1, \cdots, N_k \), \( \mathbb{E}[\nabla F(x_k, \xi_{j,k})|F_k] = \nabla f(x_k) \) almost surely and \( \mathbb{E}[\|\nabla F(x_k, \xi_{j,k}) - \nabla f(x_k)\|^2|F_k] \leq \nu^2 \) almost surely.

Since \( x_k \) is adapted to \( F_k \) and the samples \( \xi_{j,k}, j = 1, \cdots, N_k \), are independent, we obtain from Assumption 1(iii) that for any \( k \geq 0 \),

\[
\mathbb{E}[w_{k,N_k}|F_k] = \frac{\sum_{j=1}^{N_k} \mathbb{E}[\nabla F(x_k, \xi_{j,k}) - \nabla f(x_k)|F_k]}{N_k} = 0, \quad a.s.,
\]

\[
\mathbb{E}[\|w_{k,N_k}\|^2|F_k] = \frac{\sum_{j=1}^{N_k} \mathbb{E}[\|\nabla F(x_k, \xi_{j,k}) - \nabla f(x_k)\|^2|F_k]}{N_k^2} \leq \frac{\nu^2}{N_k}, \quad a.s.
\]

Since \( f(\cdot) \) is strongly convex, it has a unique optimal solution denoted by \( x^* \). Then by the optimality condition, \( \nabla f(x^*) = 0 \). We now introduce an inequality [41, Eqn. (2.1.24)] on \( f(\cdot) \) satisfying Assumptions 1(i) and 1(ii):

\[
(x - y)^T (\nabla f(x) - \nabla f(y)) \geq \frac{\eta L \|x - y\|^2}{\eta + L} + \frac{\|\nabla f(x) - \nabla f(y)\|^2}{\eta + L}, \quad \forall x, y \in \mathbb{R}^m.
\]

We then establish a recursion on the mean-squared estimation error, which can be proved by making a simple modification to the proof of [41, Theorem 2.1.15].
Lemma 1 Let Algorithm 1 be applied to (1). Suppose Assumption 1 holds and \( \alpha \in (0, \frac{2}{\eta+L}] \). Then
\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\alpha \eta L}{\eta+L}\right) \mathbb{E}[\|x_k - x^*\|^2] + \frac{\alpha^2 \nu^2}{N_k}, \quad \forall k \geq 0. \tag{7}
\]

Proof. From (4) it follows that
\[
\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha (x_k - x^*)^T (\nabla f(x_k) + w_{k,N_k})
+ \alpha^2 \left(\|\nabla f(x_k)\|^2 + 2\nabla f(x_k)^T w_{k,N_k} + \|w_{k,N_k}\|^2\right). \tag{8}
\]
Since \( \nabla f(x^*) = 0 \), by using (6), we obtain that
\[
(x_k - x^*)^T \nabla f(x_k) \geq \frac{\eta L}{\eta + L} \|x_k - x^*\|^2 + \frac{1}{\eta + L} \|\nabla f(x_k)\|^2. \tag{9}
\]
Since \( x_k \) is adapted to \( F_k \), by taking expectations conditioned on \( F_k \) on both sides of (8), and using (5) from Assumption 1(iii), we obtain that
\[
\mathbb{E}[\|x_{k+1} - x^*\|^2 | F_k] \leq \|x_k - x^*\|^2 - 2\alpha (x_k - x^*)^T \nabla f(x_k) + \alpha^2 \|\nabla f(x_k)\|^2 + \frac{\alpha^2 \nu^2}{N_k}
\leq \left(1 - \frac{2\alpha \eta L}{\eta + L}\right) \|x_k - x^*\|^2 - \alpha \left(\frac{2}{\eta + L} - \alpha\right) \|\nabla f(x_k)\|^2 + \frac{\alpha^2 \nu^2}{N_k}, \quad a.s. \tag{9}
\]
Then by \( \alpha \in (0, \frac{2}{\eta+L}] \) and taking unconditional expectations, we achieve (7). \( \square \)

2.2 Accelerated Gradient Method.

Nesterov’s accelerated gradient descent method generates a sequence that converges to the solution at a rate \( \mathcal{O}(q^k) \) where \( q \triangleq 1 - \sqrt{\eta/L} \) for \( \eta \)-strongly convex and \( L \)-smooth functions [41], and with a rate \( \mathcal{O}(1/k^2) \) for merely convex functions [42]. Nesterov proved that this is the best possible rate for any first-order method. As such, we combine the Nesterov’s accelerated method with Algorithm 1 and propose an accelerated variable sample-size stochastic gradient descent algorithm (Algorithm 2) so as to improve the rate of convergence. Such a scheme has been employed for smooth strongly convex [27], smooth convex [22], and nonsmooth (but smoothable) convex [27] stochastic optimization problems with associated rates of convergence given by \( \mathcal{O}(q^k) \), \( \mathcal{O}(1/k^2) \), and \( \mathcal{O}(1/k) \), respectively. The present paper takes a crucial step towards developing CLTs and confidence regions for the optimal solution.

Algorithm 2 Variance-reduced accel. SGD

Given arbitrary initial values \( x_0 = y_0 \in \mathbb{R}^m \), positive constants \( \alpha, \beta \), and a positive integer sequence \( \{N_k\}_{k \geq 0} \). Then iterate the following equations for \( k \geq 0 \).

\[
y_{k+1} = x_k - \alpha (\nabla f(x_k) + w_{k,N_k}), \tag{10a}
\]
\[
x_{k+1} = y_{k+1} + \frac{\beta}{2} (y_{k+1} - y_k), \tag{10b}
\]
where \( w_{k,N_k} \) is defined as in (3), \( \alpha > 0 \) and \( \beta > 0 \) are constant steplengths.
Next, we establish a bound on the expected sub-optimality gap of the iterates generated by Algorithm 2. Its proof is similar to that in [41]; hence it is omitted here but included in the appendix. This is an important preliminary result to be used in the rate analysis of Algorithm 2.

Lemma 2 Let Algorithm 2 be applied to the problem (1). Suppose Assumption 1 holds and $\alpha \in (0, \frac{1}{L})$. Set $\beta = \frac{1-\gamma}{1+\gamma}$ with $\gamma = \sqrt{\alpha \eta}$. Then for any $k \geq 1$,

$$
\mathbb{E}[f(y_k)] - f^* \leq \frac{(\eta+L)(1-\gamma)^k}{2} \mathbb{E}[\|x_0 - x^*\|^2] + \nu^2 \left( \alpha + \frac{(1-\gamma)^2}{2\nu} \right) \sum_{i=0}^{k-1} \frac{(1-\gamma)^i}{N_{k-1-i}}.
$$

(11)

2.3 The Heavy Ball Method.

The classical heavy ball method of Polyak [47] takes the form $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1})$ with a steplength $\alpha > 0$ and a momentum parameter $\beta > 0$. It was shown in [48] that for smooth strongly convex, the heavy ball method generates sequences convergent to the optimal solution at a faster rate than the gradient descent method. Stochastic variants of the heavy ball method have been employed widely in practice (cf. [30, 57, 58] for applications to machine learning). Recent efforts [21, 63] have analyzed the rate in stochastic settings and [63] proved a sublinear rate $O(1/\sqrt{k})$ for general Lipschitz continuous convex objectives with bounded variance, while a rate $O(1/k^\beta)$ with $\beta \in (0, 1)$ was provided in [21] for the case of the quadratic strongly convex functions. Convergence properties remain unaddressed in variance-reduced regimes; we consider a variable sample-size variant of the stochastic heavy ball method assuming constant steplengths, in contrast with the diminishing steplengths utilized in [21, 63], and prove the global linear convergence. In the developed Algorithm 3, we add a momentum term $\beta(x_k - x_{k-1})$ to the variable sample-size stochastic gradient step (2) and obtain the update (12).

Algorithm 3 Variance-reduced heavy-ball SGD

Given an arbitrary initial value $x_0 \in \mathbb{R}^m$, two positive constants $\alpha, \beta > 0$, and a positive integer sequence $\{N_k\}_{k \geq 0}$. Set $x_{-1} = x_0$. Then iterate the following equation for $k \geq 0$.

$$
x_{k+1} = x_k - \alpha (\nabla f(x_k) + w_{k,N_k}) + \beta (x_k - x_{k-1}),
$$

(12)

where $w_{k,N_k}$ is defined as in (3), $\alpha > 0$ and $\beta > 0$ are constant steplengths.

In Section 3, we derive a geometric rate and asymptotic normality statements for Algorithm 3 under geometrically increasing batch-sizes, whereas the corresponding results of Algorithm 3 under polynomially increasing batch-sizes will be established in Section 4. In the following, we give a simple recursion on the expected mean-squared error of the generated sequence $\{x_k\}$. The proof can be found in Appendix A.

Lemma 3 Suppose that Assumption 1 holds and $f(\cdot)$ is twice continuously differentiable. Consider Algorithm 3, where $\alpha \in (0, 4/L)$ and $\beta \triangleq \max\{1 - \sqrt{\alpha \eta}, 1 - \sqrt{6\alpha L}\} < 1$. Then

$$
\mathbb{E} \left[ \left\| \frac{x_{k+1} - x^*}{x_k - x^*} \right\|^2 \right] \leq \beta \mathbb{E} \left[ \left\| \frac{x_k - x^*}{x_{k-1} - x^*} \right\|^2 \right] + \frac{\alpha^2 \nu^2}{N_k}, \forall k \geq 0.
$$

(13)
2.4 Pathway for addressing constrained and nonsmooth regimes.

We believe that the presented avenues hold promise for contending with constrained and nonsmooth regimes.

(i) Constrained problems. Consider the constrained problem

\[ \min_{x \in \mathcal{X}} f(x) \equiv \mathbb{E}[F(x, \xi)], \tag{14} \]

where \( \mathcal{X} \) is a closed and convex with a nonempty interior. Algorithm 1 can be extended to a constrained regime with an additional projection onto the convex set \( \mathcal{X} \). Under suitable conditions, Lemma 1 holds and \( x_k \) converges almost surely to the optimal solution \( x^* \). For the case when the optimal solution \( x^* \) lies in the interior of \( \mathcal{X} \), the sequence \( \{x_k\}_{k \geq K} \) will lie in the interior of the constrained set in an a.s. sense for sufficiently large \( K \). Then by proceeding in a similar fashion as in Section 3, CLTs may be developed for constrained problems. If however, solutions are on the boundary of the set, a possible resolution may lie in computing CLTs of approximate minimizers of an unconstrained reformulation via penalization and barrier methods [43].

(ii) Nonsmooth problems. When the \( f \) is not necessarily smooth, then one avenue lies in employing smoothing approaches. Under some conditions [1], one can construct an \( (a, b) \)-smoothed convex approximation of \( f \), denoted by \( f_\eta \), where \( f_\eta(x) \leq f(x) \leq f_\eta(x) + \eta b \) and \( \|\nabla f_\eta(x) - \nabla f_\eta(y)\| \leq \frac{\eta}{a}\|x - y\| \) for all \( x, y \). One may then integrate iterative smoothing within the variance-reduced accelerated gradient scheme as follows.

\[
\begin{align*}
y_{k+1} &= x_k - \alpha_k(\nabla f_\eta_k(x_k) + w_{k,N_k}), \\
x_{k+1} &= y_{k+1} + \beta_k(y_{k+1} - y_k). \tag{15a}
\end{align*}
\]

Under suitable assumptions on \( \alpha_k, \beta_k, \eta_k, N_k \), the sequence \( \{y_k\} \) converges to a unique solution of the original problem. Under an assumption that \( f \) is smooth in a neighborhood of \( x^* \), one can again develop CLTs in this setting. If however, \( f \) is not necessarily smooth at \( x^* \), then one avenue might lie in developing \( C^2 \) smoothing-based approximations and provide CLTs for \( \epsilon \)-solutions.

A comprehensive examination of (i) and (ii) is beyond the scope of this paper but we believe the smooth and unconstrained analysis in this work is a crucial building block.

3 Central Limit Theorems under Geometrically Increasing Batch-size.

In this section, we establish CLTs for Algorithms 1, 2, and 3 when the number of sampled gradients, denoted by \( N_k \), increases at a geometric rate.

3.1 Rate and Oracle Complexities.

Based on Lemmas 1- 3, we can establish the geometric rate of convergence along with the iteration and oracle complexity of the iterates generated by Algorithms 1-3. Related results regarding linear convergence for stochastic gradient methods can be found in [10, 20, 28, 45, 53] while a linear rate of the accelerated variants has been provided in [27, 52]. The proofs of Propositions 1-3 are similar to those in our prior work [33, Theorem 4.2 and Corollary 4.7].
Proposition 1 (Rate and Oracle Complexity for Algorithm 1) Let Assumption 1 hold and $\alpha \in (0, \frac{2}{\eta + L}]$. Consider Algorithm 1, where $N_k \triangleq \lceil \rho_1^{-(k+1)} \rceil$ for some $\rho_1 \in (q, 1)$ with $q \triangleq 1 - \frac{2\alpha\eta L}{\eta + L}$. Then

$$
\mathbb{E}[\|x_k - x^*\|^2] \leq \rho_1^k \left( \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\alpha^2 \nu^2}{1 - q/\rho_1} \right), \quad \forall k \geq 1.
$$

(16)

The iteration and oracle complexity for computing an $\epsilon$-solution, defined as $\mathbb{E}[\|x_k - x^*\|^2] \leq \epsilon$, are respectively bounded by $O(\kappa \ln(1/\epsilon))$ and $O(\kappa/\epsilon)$, where $\kappa \triangleq \frac{L}{\eta}$ denotes the condition number.

Proof. By substituting $N_k \triangleq \lceil \rho_1^{-(k+1)} \rceil$ into (7), using $q = 1 - \frac{2\alpha\eta L}{\eta + L}$ and $\rho_1 \in (q, 1)$, one obtains

$$
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q\mathbb{E}[\|x_k - x^*\|^2] + \alpha^2 \nu^2 \rho_1^{k+1} \leq q^{k+1}\mathbb{E}[\|x_0 - x^*\|^2] + \alpha^2 \nu^2 \sum_{t=0}^{k} q^t \rho_1^{k+1-t}.
$$

Hence, we derive (16). Suppose we set $\alpha \triangleq \frac{2}{\eta + L}$ and $\rho_1 \triangleq \left( \frac{\kappa}{\kappa + 1} \right)^2 > q = \left( \frac{\kappa - 1}{\kappa + 1} \right)^2$. From (16) it follows that $\mathbb{E}[\|x_k - x^*\|^2] \leq \epsilon$ for any $k \geq K(\epsilon)$, where

$$
K(\epsilon) \triangleq \frac{\ln \left( \frac{\mathbb{E}[\|x_0 - x^*\|^2] + \alpha^2 \nu^2}{1 - q/\rho_1} + \ln(1/\epsilon) \right)}{2 \ln(1 + 1\epsilon)}. \quad (17)
$$

It is noticed that $\ln(1 + 1/\kappa) \approx 1/\kappa$ for large $\kappa$, then the number of iterations required to obtain an $\epsilon$-optimal solution in a mean-squared sense, i.e. $\mathbb{E}[\|x - x^*\|^2] \leq \epsilon$, is $O(\kappa \ln(1/\epsilon))$. Finally, based on the iteration complexity bound (17), we achieve the following oracle complexity bound, measured by the number of sampled gradients, for obtaining an $\epsilon$-optimal solution:

$$
\sum_{k=0}^{K(\epsilon) - 1} N_k \leq \sum_{k=1}^{K(\epsilon)} \rho_1^{-k} \leq \int_{1}^{K(\epsilon) + 1} \rho_1^{-t} dt \leq \frac{\rho_1^{-K(\epsilon) - 1}}{\ln(1/\rho_1)} = \frac{\mathbb{E}[\|x_0 - x^*\|^2] + \alpha^2 \nu^2}{2 \epsilon \rho_1 \ln(1 + 1/\epsilon)} = O\left( \frac{\kappa}{\epsilon} \right). \quad \square
$$

In the following lemma, we show that the result of Proposition 1 holds as well when the noise condition Assumption 1(iii) is replaced by some state-dependent noise condition.

Lemma 4 Consider Algorithm 1, where $N_k \triangleq \lceil \rho_1^{-(k+1)} \rceil$ for some $\rho_1 \in (q, 1)$ with $q \triangleq 1 - \frac{2\alpha\eta L}{\eta + L}$. Let $\alpha \in (0, \frac{2}{\eta + L}]$ and Assumptions 1(i)-(ii) hold. Suppose, in addition, that there exist constants $\nu_1 > 0, \nu_2 > 0$ such that

$$
\mathbb{E}[w_{k,N_k}|\mathcal{F}_k] = 0 \text{ and } \mathbb{E}[\|w_{k,N_k}\|^2|\mathcal{F}_k] \leq \frac{\nu_1^2 + \nu_2^2 \|x_k\|^2}{N_k}, \text{ a.s., } \forall k \geq 0.
$$

(18)

Then there some constant $c > 0$ such that the following holds.

$$
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \rho_1^k \left( \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\alpha^2 c}{1 - q/\rho_1} \right), \quad \forall k \geq 0.
$$

(19)

The iteration and oracle complexity are bounded by $O(\kappa \ln(1/\epsilon))$ and $O(\kappa/\epsilon)$, respectively.
Proof. Since \( x_k \) is adapted to \( \mathcal{F}_k \), by taking expectations conditioned on \( \mathcal{F}_k \) on both sides of (8), using \( \alpha \in (0, \frac{2}{\eta+L}) \), (9) and (18), we obtain the following.

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] \leq \left(1 - \frac{2\alpha \eta L}{\eta + L}\right) \|x_k - x^*\|^2 + \alpha^2 (\nu_1^2 + 2\nu_2^2 \|x_k\|^2), \quad \text{a.s.}
\]

By taking unconditional expectations, using \( N_k \triangleq \left[ \rho_1^{-1(k+1)} \right] \) and \( \|x_k\|^2 \leq 2(\|x_k - x^*\|^2 + \|x^*\|^2) \), we obtain that

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\alpha \eta L}{\eta + L} + 2\alpha^2 \nu_2^2 \rho_1^{-1(k+1)}\right) \mathbb{E}[\|x_k - x^*\|^2] + \alpha^2 \rho_1^{-1(k+1)}(\nu_1^2 + 2\nu_2^2 \|x^*\|^2).
\]

(20)

Next, we show that \( \mathbb{E}[\|x_k - x^*\|^2] \) is uniformly bounded by some constants.

**Case 1:** If \( 2\alpha^2 \nu_2^2 \leq \rho_1 - q \), then \( 1 - \frac{2\alpha \eta L}{\eta + L} + 2\alpha^2 \nu_2^2 \rho_1^{-1(k+1)} \leq q + 2\alpha^2 \nu_2^2 \rho_1 \triangleq \hat{q} < q + 2\alpha^2 \nu_2^2 \leq \rho_1 \).

Similarly to the derivation of (16), we obtain from (20) that for any \( k \geq 0 \),

\[
\mathbb{E}[\|x_k - x^*\|^2] \leq \hat{q}\mathbb{E}[\|x_k - x^*\|^2] + \alpha^2 \rho_1^{-1(k+1)}(\nu_1^2 + 2\nu_2^2 \|x^*\|^2) \leq \rho_1^{-1}(\mathbb{E}[\|x_0 - x^*\|^2] + \frac{\alpha^2(\nu_1^2 + 2\nu_2^2 \|x^*\|^2)}{1-\hat{q}/\rho_1}) \triangleq c_1.
\]

**Case 2:** If \( 2\alpha^2 \nu_2^2 > \rho_1 - q \), we define \( \tilde{k} \triangleq \left\lceil \frac{\ln(\rho_1 - q/\hat{q})}{\ln(\rho_1^{-1})} \right\rceil \). Then for any \( k \geq \tilde{k} \), \( \rho_1^{-1(k+1)} > \frac{2\alpha^2 \nu_2^2}{\rho_1 - q} \) and hence \( 2\alpha^2 \nu_2^2 \rho_1^{-1(k+1)} < \rho_1 - q \). Note from (20) that for any \( k \leq \tilde{k} \),

\[
\mathbb{E}[\|x_k - x^*\|^2] \leq (1 + 2\alpha^2 \nu_2^2 \rho_1^{-1}) \mathbb{E}[\|x_{k-1} - x^*\|^2] + \alpha^2(\nu_1^2 + 2\nu_2^2 \|x^*\|^2) \leq (1 + 2\alpha^2 \nu_2^2 \rho_1^{-1} \tilde{k})(\mathbb{E}[\|x_0 - x^*\|^2] + \alpha^2(\nu_1^2 + 2\nu_2^2 \|x^*\|^2))(1 + 2\alpha^2 \nu_2^2 \rho_1^{-1} \tilde{k})^{-1} \triangleq c_2.
\]

By defining \( \hat{q} \triangleq q + 2\alpha^2 \nu_2^2 \rho_1^{-1} \tilde{k} \), we have \( 1 - \frac{2\alpha \eta L}{\eta + L} + 2\alpha^2 \nu_2^2 \rho_1^{-1} \tilde{k} \leq \hat{q} < \rho_1 \) for any \( k \geq \tilde{k} \). Then it follows from (20) that for any \( k \geq \tilde{k} \),

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \hat{q}\mathbb{E}[\|x_k - x^*\|^2] + \alpha^2 \rho_1^{-1(k+1)}(\nu_1^2 + 2\nu_2^2 \|x^*\|^2) \leq \hat{q}^{k-\tilde{k}} \mathbb{E}[\|x_k - x^*\|^2] + \alpha^2(\nu_1^2 + 2\nu_2^2 \|x^*\|^2) \sum_{t=0}^{k-\tilde{k}} \rho_1^{-1(k+1-t)} \hat{q}^t \leq c_2 + \frac{\alpha^2(\nu_1^2 + 2\nu_2^2 \|x^*\|^2)}{1-\hat{q}/\rho_1}.
\]

Thus, for the case \( 2\alpha^2 \nu_2^2 > \rho_1 - q \), we have that \( \mathbb{E}[\|x_k - x^*\|^2] \leq c_2 + \frac{\alpha^2(\nu_1^2 + 2\nu_2^2 \|x^*\|^2)}{1-\hat{q}/\rho_1} \) for any \( k \geq 1 \).

By combining \( 2\alpha^2 \nu_2^2 \leq \rho_1 - q \) (**Case 1**) and \( 2\alpha^2 \nu_2^2 > \rho_1 - q \) (**Case 2**), we conclude that there exists some constant \( c_3 > 0 \) such that \( \mathbb{E}[\|x_k - x^*\|^2] \leq c_3 \) for any \( k \geq 1 \). This combines with (20) produces

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\alpha \eta L}{\eta + L}\right) \mathbb{E}[\|x_k - x^*\|^2] + \alpha^2 \rho_1^{-1(k+1)}(\nu_1^2 + 2\nu_2^2 \|x^*\|^2 + 2\nu_2^2 c_3) \triangleq c_4.
\]

The rest of the proof is the same as that of Proposition 1.

Next, we provide similar statements for Algorithms 2 and 3, for which the proofs are provided in the supplementary material for completeness.
Proposition 2 (Rate and Oracle Complexity for Algorithm 2) Let Assumption 1 hold. Consider Algorithm 2, where \( \alpha \in (0,1) \), \( \gamma = \sqrt{\alpha \eta} \), \( \beta = \frac{1-\gamma}{1+\gamma} \), and \( N_k \triangleq \lceil \rho_2^{-k+1} \rceil \) with \( \rho_2 \in (1-\gamma,1) \). Then for any \( k \geq 0 \),

\[
\mathbb{E}[f(y_k)] - f^* \leq \rho_2^k \left( \frac{n+L}{2} \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\rho_2 \nu^2}{\rho_2 - (1-\gamma)^2} \left( \alpha + \frac{(1-\gamma)\gamma}{2\eta} \right) \right),
\]

(21)

and \( \mathbb{E}[\|x_k - x^*\|^2] \leq c \rho_2^k \) for some constant \( c > 0 \). In addition, the iteration and oracle complexity for obtaining an \( \epsilon \)-solution are \( O(\sqrt{\kappa \ln(1/\epsilon)}) \) and \( O(\sqrt{\kappa/\epsilon}) \), respectively.

Proposition 3 (Rate and Oracle Complexity for Algorithm 3) Suppose that Assumption 1 holds and that \( f(\cdot) \) is twice continuously differentiable. Consider Algorithm 3, where \( \alpha = \frac{4}{(\sqrt{\eta} + \sqrt{L})^2} \), \( \beta = \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right) \), and \( N_k \triangleq \lceil \rho_3^{-(k+1)} \rceil \) with \( \rho_3 \in (\beta,1) \). Then

\[
\mathbb{E} \left[ \left\| \left( \frac{x_{k+1} - x^*}{x_k - x^*} \right) \right\|^2 \right] \leq \left( 2 \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\alpha^2 \nu^2}{1-\beta/\rho_3} \right) \rho_3^{k+1}.
\]

(22)

In addition, the iteration and oracle complexity required for obtaining an \( \epsilon \)-solution in a mean-squared sense are respectively \( O(\sqrt{\kappa \ln(1/\epsilon)}) \) and \( O(\sqrt{\kappa/\epsilon}) \).

Remark 1 (i) From the complexity statements in Propositions 1 and 2, the dependency on the condition number is improved from \( \kappa \) (in Algorithm 1) to \( \sqrt{\kappa} \) by the accelerated Algorithm 2. (ii) The rate and oracle complexity for the heavy ball method (Algorithm 3) is similar to that of Algorithm 2 but Algorithm 3 requires stronger assumptions than Algorithms 1–2 in that \( f(\cdot) \) is twice continuously differentiable. (iii) Propositions 1-3 imply that a smaller \( \kappa \) leads to a smaller constant in the oracle complexity.

Remark 2 (Convergence in Probability) Because the mean-squared convergence implies convergence in probability, on the basis of Propositions 1-3, we may conclude that the sequences \( \{x_k\} \) and \( \{y_k\} \) generated by Algorithms 1-3 converge in probability to the optimal solution \( x^* \).

3.2 Preliminary Lemmas.

Before establishing CLTs for Algorithms 1-3, we first introduce a preliminary CLT on doubly-indexed random variables [12, Lemma 3.3.1]. We state it as Lemma 5, whose proof is found in [23, Chapter 12].

Lemma 5 Let \( \xi_{kt}, t = 1, \cdots, k \) be \( m \)-dimensional random vectors. Define

\[
S_{kt} \triangleq \mathbb{E}[\xi_{kt}^T \xi_{kt}], \quad R_{kt} \triangleq \mathbb{E}[\xi_{kt}^T \xi_{kt} | \xi_{k1}, \cdots, \xi_{k,t-1}], \quad \text{and} \quad S_k \triangleq \sum_{t=1}^{k} S_{kt}
\]

(23)

Assume that \( \mathbb{E}[\xi_{kt} | \xi_{k1}, \cdots, \xi_{k,t-1}] = 0 \), \( \sup_{k \geq 1} \sum_{t=1}^{k} \mathbb{E}[\|\xi_{kt}\|^2] < \infty \),

(24)

12
\[
\lim_{k \to \infty} S_k = S, \quad \lim_{k \to \infty} \sum_{t=1}^{k} E[\|S_{kt} - R_{kt}\|] = 0, \quad (25)
\]

and \[
\lim_{k \to \infty} \sum_{t=1}^{k} E[\|\xi_{kt}\|^2 I[\|\xi_{kt}\| \geq \epsilon]] = 0, \quad \forall \epsilon > 0. \quad (26)
\]

Then \[
\sum_{t=1}^{k} \xi_{kt} \xrightarrow{d_{k \to \infty}} N(0, S).
\]

To establish the CLTs for Algorithms 1–3, we further require the following conditions.

**Assumption 2** (i) \(\nabla f : \mathbb{R}^m \to \mathbb{R}^m\) is differentiable at \(x^*\) with Hessian matrix \(H\),

\[
\nabla f(x) = H(x - x^* + \delta(x)), \quad \text{where} \quad \delta(x^*) = 0 \quad \text{and} \quad \delta(x) = o(\|x - x^*\|).
\]

(ii) The noise sequence \(\{w_{k,N_k}\}\) further satisfies:

\[
\lim_{k \to \infty} \left( N_k E w_{k,N_k}^T F_k \right) = \lim_{k \to \infty} \left( N_k E w_{k,N_k} w_{k,N_k}^T \right) = S_0, \quad \text{a.s.,} \quad (28)
\]

and \[
\lim_{r \to \infty} \sup_k E \left[ \left\| \sqrt{N_k} w_{k,N_k} \right\|^2 I[\left\| \sqrt{N_k} w_{k,N_k} \right\| > r] \right] = 0, \quad (29)
\]

where \(I_{[a,b]} = 1\) if \(a > b\), and \(I_{[a,b]} = 0\), otherwise.

**Remark 3** Since the Lindeberg condition condition (29) is less easily verified, we provide a sufficient condition for the ease of understanding and verification in practice. Suppose there exists a constant \(\delta > 0\) and a finite value \(b > 0\) such that \(E[\|\sqrt{N_k} w_{k,N_k}\|^{2+\delta}] \leq b\) for any \(k \geq 0\). Therefore,

\[
E \left[ \left\| \sqrt{N_k} w_{k,N_k} \right\|^2 I[\|\sqrt{N_k} w_{k,N_k}\| > r] \right] \leq \frac{1}{r^\delta} E \left[ \left\| \sqrt{N_k} w_{k,N_k} \right\|^{2+\delta} I[\|\sqrt{N_k} w_{k,N_k}\| > r] \right] \leq \frac{1}{r^\delta} E \left[ \left\| \sqrt{N_k} w_{k,N_k} \right\|^{2+\delta} \right] \leq \frac{b}{r^\delta}.
\]

This implies that \sup_k E \left[ \left\| \sqrt{N_k} w_{k,N_k} \right\|^2 I[\|\sqrt{N_k} w_{k,N_k}\| > r] \right] \leq \frac{b}{r^\delta}. Hence (29) holds.

In the following, we establish the central limit theorem of a linear recursion, for which the proof is provided in Appendix B. This result will be applied in the proof of Theorems 1-3.

**Lemma 6** Suppose that \(P\) is a stable matrix (i.e., \(\|P\| < \rho\) for some \(\rho < 1\), \(N_k = \lfloor P^{-(k+1)} \rfloor\) with \(\rho \in (0, 1)\) for any \(k \geq 0\), and that \(\{w_{k,N_k}\}\) satisfies Assumption 2(ii). Let \(\{e_k\}\) be generated by

\[
e_{k+1} = P e_k - \alpha \rho^{-(k+1)}/2 G w_{k,N_k} + \zeta_{k+1}, \quad (30)
\]

where \(E[\|e_0\|^2] < \infty\) and \(\zeta_{k+1} \xrightarrow{p_{k \to \infty}} 0\). Then

\[
\alpha^{-1} e_k \xrightarrow{d_{k \to \infty}} N(0, \Sigma) \quad \text{with} \quad \Sigma \equiv \lim_{k \to \infty} \sum_{t=0}^{k} P^t G S_0 G^T (P^t)^T.
\]
3.3 Central Limit Theorems.

Based on Proposition 1 and Lemma 6, using Assumption 2, we present the central limit theorem for Algorithm 1 with geometrically increasing batch-sizes.

Theorem 1 (CLT of Algorithm 1 with Geometrically increasing $N_k$) Let Assumptions 1 and 2 hold. Consider Algorithm 1, where $\alpha \in (0, \frac{2}{q+L}]$. Set $q \triangleq 1 - \frac{2\alpha L}{q+L}$, $N_k \triangleq \lceil \rho^{-1}(k+1) \rceil$ with $\rho_1 \in (q, 1)$, and $P_1 \triangleq \rho_1^{-1/2}(I_m - \alpha H)$. Then

$$
\alpha^{-1} \rho_1^{-k/2}(x_k - x^*) \xrightarrow{d}{k\to\infty} N(0, \Sigma_1), \text{ where } \Sigma_1 \triangleq \lim_{k \to \infty} k \sum_{t=0}^{k} P_1^t S_0 P_1^t. \quad (31)
$$

Proof. By definition of $\delta(x)$ in (27), (4) can be rewritten as:

$$
x_{k+1} - x^* = x_k - x^* - \alpha H(x_k - x^*) - \alpha (\nabla f(x_k) - H(x_k - x^*) + w_k, N_k)
= (I_m - \alpha H)(x_k - x^*) - \alpha (\delta(x_k) + w_k, N_k). \quad (32)
$$

Define $D_k \triangleq 0_m$ if $x_k = x^*$, and $D_k \triangleq \frac{\delta(x_k)(x_k - x^*)^T}{\|x_k - x^*\|^2}$ when $x_k \neq x^*$. Since $\delta(x^*) = 0$, we obtain that

$$
\delta(x_k) = D_k(x_k - x^*). \quad (33)
$$

Note by Proposition 1 that $x_k \xrightarrow{P}{k\to\infty} x^*$. Hence $x_k \xrightarrow{P}{k\to\infty} x^*$ by Markov’s inequality. Then by using [60, Lemma 2.12(i)], $\delta(x^*) = 0$, $\delta(x) = o(\|x - x^*\|)$, and $x_k \xrightarrow{P}{k\to\infty} x^*$, we obtain $\delta(x_k) = o_P(\|x_k - x^*\|)$. This together with (33) implies $o_P(\|x_k - x^*\|) = D_k(x_k - x^*)$. Thus, by recalling the definition “$X_k = o_P(R_k) \iff X_k = Y_k R_k$ and $Y_k \xrightarrow{P}{k\to\infty} 0$” from [60, Section 2.2 in p.12], we may conclude that $D_k \xrightarrow{P}{k\to\infty} 0$.

Define $e_k \triangleq \rho_1^{-k/2}(x_k - x^*)$. By multiplying both sides of (32) by $\rho_1^{-k/2}$, using (33) and the definition $P_1 \triangleq \rho_1^{-1/2}(I_m - \alpha H)$,

$$
e_{k+1} = \rho_1^{-1/2}(I_m - \alpha H)\rho_1^{-k/2}(x_k - x^*) - \alpha \rho_1^{-k/2}(D_k(x_k - x^*) + w_k, N_k)
= P_1 e_k - \alpha \rho_1^{-k/2} w_k, N_k + \zeta_{k+1}, \quad (34)
$$

where $\zeta_{k+1} \triangleq -\alpha \rho_1^{-k/2} D_k(x_k - x^*)$. From (16), it follows that for any $k \geq 0$,

$$
\text{var}(e_k) \leq E[\|e_k\|^2] = E[\|x_k - x^*\|^2/\rho_1^{k/2}] \leq E[\|x_0 - x^*\|^2] + \frac{\alpha^2 \nu^2}{1 - q/\rho_1} \triangleq v_e^2 \text{ with } v_e > 0. \quad (35)
$$

Chebyshev’s inequality asserts that if $X$ is a random variable with mean $\mu$ and variance $\sigma^2$, then for any real number $h > 0$,

$$
P(\|X - \mu\| \leq h\sigma) \geq 1 - h^{-2}. \quad (36)
$$
By setting $h = \chi^{-1/2}$ for any $\chi \in (0, 1)$ and applying (36) to $e_k$, we have

$$
\mathbb{P}(\|e_k - \mathbb{E}[e_k]\| \leq \chi^{-1/2}\text{var}(e_k)) \geq 1 - \chi. 
$$

(37)

Define the events $A_1 \triangleq \{\|e_k - \mathbb{E}[e_k]\| \leq \chi^{-1/2}\text{var}(e_k)\}, A_2 \triangleq \{\|e_k - \mathbb{E}[e_k]\| \leq \chi^{-1/2}v_e^2\}, A_3 \triangleq \{\|e_k|| \leq \|\mathbb{E}[e_k]\| + \chi^{-1/2}v_e^2\}, A_4 \triangleq \{\|e_k|| \leq \|\mathbb{E}[e_k]\| + \chi^{-1/2}v_e^2\},$ and $A_5 \triangleq \{\|e_k\| \leq v_e + \chi^{-1/2}v_e^2\}.$ Note by (35), we have that $A_1 \subseteq A_2.$ We observe that $A_2 \subseteq A_3$ by the inequality $\|x_1 - x_2\| \geq \|x_1\| - \|x_2\|.$ Since $\|\cdot\|$ is convex in $x,$ by the Jensen’s inequality we have $\|\mathbb{E}[X]\| \leq \|\mathbb{E}[X]\|,$ and hence $A_3 \subseteq A_4.$ Since $\|\cdot\|$ is a convex function, by using (35) and Jensen’s inequality, we have $\mathbb{E}[\|e_k\|] \leq \sqrt{\mathbb{E}[\|e_k\|^2]} \leq v_e$ and hence $A_4 \subseteq A_5.$ This together with (37) implies that for any $k \geq 0,$

$$
\mathbb{P}(\|e_k\| \leq v_e + \chi^{-1/2}v_e^2) = \mathbb{P}(A_5) \geq \mathbb{P}(A_1) = \mathbb{P}(\|e_k - \mathbb{E}[e_k]\| \leq \chi^{-1/2}\text{var}(e_k)) \geq 1 - \chi. 
$$

Therefore, $e_k$ is bounded in probability (i.e., $e_k = O_P(1)$). Recall from [60, p.12] that symbols $o_P(\cdot)$ and $O_P(\cdot)$ satisfy $O_P(1) o_P(1) = o_P(1).$ Then by using $e_k = O_P(1)$ and $D_k = o_P(1)$, we obtain

$$
\zeta_{k+1} = -\alpha\rho_1^{-1/2}D_ke_k \xrightarrow{P_{k \to \infty}} 0. 
$$

(38)

Since $H$ is the Hessian matrix of $f(x)$ at $x = x^*$, by Assumptions 1(i) and 1(ii), we conclude that $H$ has eigenvalues $\lambda_1, \cdots, \lambda_m$ that satisfy $0 < \eta \leq \lambda_m \leq \lambda_{m-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 \leq L.$ Then $\|I_m - \alpha H\|_2 \leq \max\{|1 - \alpha\eta|, |1 - \alpha L|\}.$ We first show that $|1 - \alpha\eta| \geq |1 - \alpha L|.$ It is easily seen that $|1 - \alpha\eta| \leq |1 - \alpha L|$ when $\alpha \in (0, 1/L].$ While for any $\alpha \in [1/L, \frac{2}{\eta + L}],$ $|1 - \alpha\eta| = 1 - \alpha\eta, |1 - \alpha L| = \alpha L - 1,$ and hence $|1 - \alpha\eta| - |1 - \alpha L| = 2 - \alpha(\eta + L) \geq 0.$ Then for any $\alpha \in (0, \frac{2}{\eta + L}],$

$$
\|I_m - \alpha H\| \leq |1 - \alpha\eta| = 1 - \alpha\eta \leq \sqrt{1 - \frac{2\alpha\eta}{\eta + L}}. 
$$

This is because $\alpha \leq \frac{2}{\eta + L} \Rightarrow \frac{2L}{\eta + L} + \alpha\eta - 2 \leq 0 \Rightarrow \frac{2\alpha\eta}{\eta + L} + (\alpha\eta)^2 - 2\alpha\eta \leq 0 \Rightarrow (1 - \alpha\eta)^2 \leq 1 - \frac{2\alpha\eta}{\eta + L}.$ Since $q = 1 - \frac{2\alpha\eta}{\eta + L}$ and $\rho_1 \in (q, 1),$ we have that

$$
\|P_1\| = \rho_1^{-1/2}\|I_m - \alpha H\| \leq (q/\rho_1)^{1/2} < 1. 
$$

(39)

Then by invoking Lemma 6, setting $G = I_m$ and $P = P_1$ (symmetric), we conclude that the sequence $\alpha^{-1}e_k$ generated by (34) converges in distribution to a normally distributed random variable with zero mean and covariance $\Sigma_1$ defined as in (31). The result follows by $e_k = \rho_1^{-k/2}(x_k - x^*).$

Based on Lemma 4, by employing similar proof arguments as in Theorem 1, we conclude that the central limit result established in Theorem 1 also holds under Assumptions 1(i)-(ii), Assumption 2, and the state-dependent noise condition (18).

**Remark 4** Suppose we set $\alpha = \frac{2}{\eta + L}$ and $\rho_1 \triangleq \left(\frac{\alpha}{\alpha + 1}\right)^2$ in Algorithm 1. If $V \sim N(0, I_m),$ it follows from (31) that $\frac{\eta + L}{2} \left(\frac{\alpha + 1}{\alpha}\right)\kappa (x_k - x^*) \xrightarrow{d_{k \to \infty}} N(0, \Sigma_1).$ This implies that

$$
\frac{\eta + L}{2} \left(\frac{\alpha + 1}{\alpha}\right)^k \left(1 - \frac{1}{\kappa + 1}\right)^k \Sigma_1^{1/2} V \text{ for large } k. 
$$

(40)
This result has several implications.
(i) The sequence \( \{x_k\} \) converges in distribution to the optimal solution \( x^* \) at rate \((\frac{\kappa}{k+1})^k\), and \( \{x_k\} \) is asymptotically normally distributed for large \( k \). This provides the possibility of assessing confidence regions of the estimate from the normal distribution.
(ii) The estimation error for large \( k \) depends on the structure of the studied problem (including \( \eta, L \), and the Hessian matrix \( H \)) and probability distribution of the gradient noise, measured through the coupling matrix \( \Sigma_1 \). Thus, the problem’s difficulty is largely characterized by the covariance matrix \( \Sigma_1 \).

Based on the CLT established in Theorem 1, we proceed to use the Delta method [54] to derive the asymptotic distribution of the gradient map \( \nabla f(\cdot) \) and the objective function value \( f(\cdot) \) based on the estimation sequence \( \{x(k)\} \).

**Corollary 1** Consider Algorithm 1 and suppose all conditions of Theorem 1 hold. Then

(i) \( \alpha^{-1} \rho_{1}^{-k/2}(\nabla f(x_k) - \nabla f(x^*)) \xrightarrow{d_{k\to\infty}} N(0, H\Sigma_1H) \).

(ii) \( \alpha^{-2} \rho_{1}^{-k}(f(x_k) - f(x^*)) \xrightarrow{d_{k\to\infty}} \frac{1}{2}N(0, \Sigma_1)^THN(0, \Sigma_1) \).

**Proof.** (i) Note by Assumption 2(i) that the gradient mapping \( \nabla f(x) : \mathbb{R}^m \to \mathbb{R}^m \) is differentiable at \( x^* \) with Hessian matrix \( H \). By using (31) and the Delta theorem [54, Eqn. (7.182)], we have that

\[
\alpha^{-1} \rho_{1}^{-k/2}(\nabla f(x_k) - \nabla f(x^*)) \xrightarrow{d_{k\to\infty}} HN(0, \Sigma_1).
\]

Then by the fact that \( \nabla f(x^*) = 0 \) and \( H \) is symmetric, the assertion (i) holds.

(ii) By using (31), \( \nabla f(x^*) = 0 \), and the second-order Delta theorem [54, Theorem 7.70], one obtains

\[
\alpha^{-2} \rho_{1}^{-k}(f(x_k) - f(x^*)) \xrightarrow{d_{k\to\infty}} \frac{1}{2}f''(x^*) \text{ with } u = N(0, \Sigma_1),
\]

where \( f''(x^*) \) denotes the second order directional derivative at \( x^* \) along the direction \( u \). Since \( f : \mathbb{R}^m \to \mathbb{R} \) is twice continuously differentiable, \( f''(x^*) = u^THu \). Then result (ii) follows from \( u = N(0, \Sigma_1) \). \(\square\)

Akin to Theorem 1, based on Proposition 2 and Lemma 6, we can also establish the CLT for Algorithm 2 with geometrically increasing batch-sizes.

**Theorem 2 (CLT for Algorithm 2 with Geometrically Increasing \( N_k \))** Suppose Assumptions 1 and 2 hold. Consider Algorithm 2, where \( \alpha \in (0, \frac{1}{2}], \gamma \triangleq \sqrt{\alpha \eta}, \beta \triangleq \frac{1 - \gamma}{1 + \gamma}, \) and \( N_k \triangleq \lfloor \rho_2^{(k+1)} \rfloor \) with \( \rho_2 \in (1 - \gamma, 1) \). Then

\[
\alpha^{-1} \rho_{2}^{-k/2}\left(\frac{y_k - x^*}{y_{k-1} - x^*}\right) \xrightarrow{d_{k\to\infty}} N(0, \Sigma_2) \text{ with } \Sigma_2 \triangleq \lim_{k\to\infty} \sum_{t=0}^{k} P_2^t \begin{pmatrix} S_0 & 0_m \\ 0_m & 0_m \end{pmatrix} P_2^t^T,
\]

and \( P_2 \triangleq \rho_{2}^{-1/2}\begin{pmatrix} (1 + \beta)(I_m - \alpha H) & -\beta(I_m - \alpha H) \\ I_m & 0_m \end{pmatrix} \).

16
\textbf{Proof.} Similarly to (32), by using (27), we can rewrite (10a) as
\begin{equation}
y_{k+1} - x^* = (I_m - \alpha H)(x_k - x^*) - \alpha(\delta(x_k) + w_{k,N_k}). \tag{41}\end{equation}
Note by (10b) that \(x_k - x^* = (1 + \beta)(y_k - x^*) - \beta(y_{k-1} - x^*).\) This together with (41) produces
\begin{equation*}y_{k+1} - x^* = (I_m - \alpha H)((1 + \beta)(y_k - x^*) - \beta(y_{k-1} - x^*)) - \alpha(\delta(x_k) + w_{k,N_k})\end{equation*}
\begin{equation*}= (1 + \beta)(I_m - \alpha H)(y_k - x^*) - \beta(I_m - \alpha H)(y_{k-1} - x^*) - \alpha(\delta(x_k) + w_{k,N_k}).\end{equation*}
Define \(z_{k+1} \triangleq \begin{pmatrix} y_{k+1} - x^* \\ y_k - x^* \end{pmatrix} \) and \(H_2 \triangleq \begin{pmatrix} (1 + \beta)(I_m - \alpha H) & -\beta(I_m - \alpha H) \\ I_m & 0_m \end{pmatrix}.\) Then based on the above equation, we obtain the following recursion.
\begin{equation}
z_{k+1} = H_2z_k - \alpha \begin{pmatrix} \delta(x_k) + w_{k,N_k} \\ 0 \end{pmatrix}. \tag{42}\end{equation}

The eigenvalue decomposition of \(H\) is given by \(H = U\Lambda U^T,\) where \(\Lambda \triangleq \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\}\) and \(U\) is orthogonal. This allows us to rewrite \(H_2\) as
\begin{equation*}H_2 = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} (1 + \beta)(I_m - \alpha \Lambda) & -\beta(I_m - \alpha \Lambda) \\ I_m & 0 \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix}^T.\end{equation*}
As a result, the eigenvalues of the matrix \(H_2\) are the roots of the following characteristic equations for \(i = 1, \ldots, m.\)
\begin{equation}
v - (1 + \beta)(1 - \alpha \lambda_i) & \beta(1 - \alpha \lambda_i) \\ -1 & v \end{equation}
\begin{equation*}= v^2 - (1 + \beta)(1 - \alpha \lambda_i)v + \beta(1 - \alpha \lambda_i) = 0. \tag{43}\end{equation*}
By \(\gamma^2 = \alpha \eta\) and \(\beta = \frac{1 - \gamma}{1 + \gamma},\) we show that the discriminant of the above quadratic equation is nonpositive for each \(i = 1, \ldots, m.\)
\begin{equation*}\Delta_i = (1 + \beta)^2(1 - \alpha \lambda_i)^2 - 4\beta(1 - \alpha \lambda_i) = 4(1 - \alpha \lambda_i) \left(\frac{(1 - \alpha \lambda_i)}{(1 + \gamma)^2} - \frac{1 - \gamma}{1 + \gamma}\right)\end{equation*}
\begin{equation*}= \frac{4(1 - \alpha \lambda_i)}{(1 + \gamma)^2} (\gamma^2 - \alpha \lambda_i) = \frac{4\alpha(1 - \alpha \lambda_i)}{(1 + \gamma)^2} (\eta - \lambda_i) \leq 0,\end{equation*}
where the last inequality holds by the fact that for any \(i = 1, \ldots, m, 1 - \alpha \lambda_i \geq 1 - \lambda_i/L \geq 0\) and \(\eta - \lambda_i \leq 0.\) Consequently, (43) has two complex roots \(\frac{1 + \beta(1 - \alpha \lambda_i)}{2} \pm \sqrt{\frac{\Delta_i}{4}},\) where \(i\) denotes the imaginary part. Thus, the magnitude of the roots is \(\sqrt{\left(\frac{1 + \beta(1 - \alpha \lambda_i)}{2}\right)^2 - \frac{\Delta_i}{4}} = \sqrt{\beta(1 - \alpha \lambda_i)}.\) Since \(\gamma^2 = \alpha \eta,\)
\(\beta = \frac{1 - \gamma}{1 + \gamma}, \rho_2 \in (1 - \gamma, 1),\) and \(\lambda_i \in [\eta, L],\) we obtain that
\begin{equation}
\|H_2\| = \max_{i} \sqrt{\beta(1 - \alpha \lambda_i)} \leq \sqrt{\beta(1 - \alpha \eta)} = \sqrt{\frac{(1 - \gamma)(1 - \gamma^2)}{1 + \gamma}} = 1 - \gamma < \rho_2 < \rho_2^{1/2}. \tag{44}\end{equation}
Define $\varepsilon_k \triangleq \rho_2^{-k/2} z_k$. By multiplying both sides of (42) by $\rho_2^{-(k+1)/2}$, we obtain the following recursion with $s_{k+1} \triangleq -\alpha \rho_2^{-(k+1)/2} \delta(x_k)$:

$$
\varepsilon_{k+1} = \rho_2^{-1/2} H_2 \varepsilon_k - \alpha \rho_2^{-(k+1)/2} \left( \delta(x_k) + w_{k,N_k} \right) = P_2 \varepsilon_k - \alpha \rho_2^{-(k+1)/2} \left( w_{k,N_k} \right) + \left( s_{k+1} \right).
$$

By Proposition 2, we see that $\mathbb{E}[\rho_k^{-k} \|x_k - x^*\|^2] \leq c$ for some constant $c > 0$. Recall from Remark 2 that $x_k - x^* = O_P(1)$. Similarly to the procedures for proving (38), we can show that $\rho_2^{-k/2}(x_k - x^*) = O_P(1)$, and hence

$$
\delta(x_k) = O_P(\|x_k - x^*\|) = D_k(x_k - x^*) \text{ with } D_k = O_P(1),
$$

and hence

$$
\varepsilon_{k+1} = -\alpha \rho_2^{-(k+1)/2} \delta(x_k) = -\alpha \rho_2^{-(k+1)/2} D_k(x_k - x^*) \xrightarrow{P_{k \to \infty}} 0.
$$

Because $\|P_2\| = \rho_2^{-1/2} \|H_2\| < 1$ by (44), by setting $G = \left( \begin{array}{c} I_m \\ 0_m \end{array} \right)$ and using Lemma 6, we obtain that

$$
\alpha^{-1} \varepsilon_k \xrightarrow{d} N(0, \Sigma_2) \text{ with } \Sigma_2 \triangleq \lim_{k \to \infty} \sum_{t=0}^{k} P_2^t G S_0 G^T (P_2^t)^T. \text{ Then by } G S_0 G^T = \left( \begin{array}{cc} S_0 & 0_m \\ 0_m & 0_m \end{array} \right)
$$

and $\varepsilon_k = \rho_2^{-k/2} z_k = \rho_2^{-k/2} \left( \frac{y_k - x^*}{y_{k-1} - x^*} \right)$, we prove the result. \quad \Box

**Remark 5** By setting $\alpha = \frac{1}{L}$ and $\rho_2 = 1 - \frac{1}{a \sqrt{\kappa}}$ for some $a > 1$ in Algorithm 2, we have that $1 - \gamma = 1 - \frac{1}{\sqrt{\kappa}} < \rho_2$, and Theorem 2 holds. Thus, we obtain $L \left( 1 - \frac{1}{a \sqrt{\kappa}} \right)^{-k/2} \left( \frac{y_k - x^*}{y_{k-1} - x^*} \right) \xrightarrow{d} N(0, \Sigma_2)$. Hence the following holds with $V \sim N(0, I_{2m})$.

$$
\left( \begin{array}{c} y_k \\ y_{k-1} \end{array} \right) \xrightarrow{D} \left( \begin{array}{c} x^* \\ x^* \end{array} \right) + \frac{1}{L} \left( 1 - \frac{1}{a \sqrt{\kappa}} \right)^{k/2} \Sigma_2^{1/2} V \text{ for large } k.
$$

In a similar fashion, we establish the central limit theorem for the stochastic heavy ball method (Algorithm 3) with geometrically increasing batch-sizes based on Proposition 3 and Lemma 6.

**Theorem 3 (CLT for Algorithm 3 with Geometrically Increasing $N_k$)** Suppose that Assumptions 1 and 2 hold, and $f(\cdot)$ is twice continuously differentiable. Consider Algorithm 3, where $\alpha \triangleq \frac{1}{(\sqrt{\kappa} + L)^2}$, $\beta \triangleq \left( 1 - \frac{2}{\sqrt{\kappa} + 1} \right)^2$, and $N_k \triangleq \rho_3^{-(k+1)}$ with $\rho_3 \in (\beta, 1)$ for any $k \geq 0$. Then

$$
\alpha^{-1} \rho_3^{-k/2} \left( \frac{x_k - x^*}{x_{k-1} - x^*} \right) \xrightarrow{d} N(0, \Sigma_3) \text{ with } \Sigma_3 \triangleq \lim_{k \to \infty} \sum_{t=0}^{k} P_3^t \left( \begin{array}{cc} S_0 & 0_m \\ 0_m & 0_m \end{array} \right) (P_3^t)^T,
$$

where $P_3 \triangleq \rho_3^{-1/2} \left( \begin{array}{cc} (1 + \beta) I_m - \alpha H & -\beta I_m \\ I_m & 0_m \end{array} \right)$.
Proof. From (12) and Assumption 2(a) it follows that

\[ x_{k+1} - x^* = x_k - x^* + \beta(x_k - x_{k-1}) - \alpha H(x_k - x^*) - \alpha \delta(x_k) - \alpha w_{k,N_k}. \]

Then we may rewrite the recursion in a matrix form as follows:

\[
\begin{pmatrix}
  x_{k+1} - x^* \\
  x_k - x^*
\end{pmatrix} = \begin{pmatrix}
(1 + \beta)I_m - \alpha H & -\beta I_m \\
I_m & 0_m
\end{pmatrix} \begin{pmatrix}
x_k - x^* \\
x_{k-1} - x^*
\end{pmatrix} - \alpha \begin{pmatrix}
\delta(x_k) + w_{k,N_k} \\
0
\end{pmatrix}.
\]

Define \( \varepsilon_k \triangleq k^{-k/2} \begin{pmatrix} x_k - x^* \\ x_{k-1} - x^* \end{pmatrix} \). By multiplying the above equation with \( \rho_3^{-(k+1)/2} \), one obtains

\[
\varepsilon_{k+1} = \mathbf{P}_3 \varepsilon_k - \alpha \rho_3^{-(k+1)/2} \begin{pmatrix} I_m \\ 0_m \end{pmatrix} \mathbf{w}_{k,N_k} + \begin{pmatrix} \zeta_{k+1} \\ 0 \end{pmatrix} \text{ with } \zeta_{k+1} \triangleq -\alpha \rho_3^{-(k+1)/2} \delta(x_k).
\]

From Proposition 3 it follows that \( \mathbb{E}[\rho_3^{-k} \|x_k - x^*\|^2] \leq 2 \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\sigma^2}{1 - \beta/\rho_3} \) and \( x_k - x^* = \mathcal{O}(1) \). Similarly to the procedures for proving (38), we can show that \( \rho_3^{-k/2} (x_k - x^*) = \mathcal{O}(1), \delta(x_k) = \mathcal{O}(\|x_k - x^*\|) \) is \( \mathcal{D}_k(x_k - x^*) \) with \( \mathcal{D}_k = \mathcal{O}(1) \), and hence

\[
\zeta_{k+1} = -\alpha \rho_3^{-(k+1)/2} \delta(x_k) = -\alpha \rho_3^{-(k+1)/2} \mathcal{D}_k(x_k - x^*) \xrightarrow{P} 0.
\]

Since \( \mathbf{P}_3 = \rho_3^{-1/2} \sqrt{\beta} < 1 \) by (58), the result follows by invoking Lemma 6.

Remark 6 Suppose we set \( \rho_3 = \left(1 - \frac{1}{\sqrt{k+1}}\right)^2 \) in Algorithm 3, then \( \rho_3 > \beta \) and Theorem 3 holds; i.e.,

\[
\frac{(\sqrt{\pi} + \sqrt{L})^2}{4} \left(1 - \frac{1}{\sqrt{k+1}}\right)^{-k} \begin{pmatrix} x_k - x^* \\ x_{k-1} - x^* \end{pmatrix} \xrightarrow{d} N(0, \Sigma_3). \text{ If we denote } V \sim N(0, \mathbf{I}_{2m}), \text{ then}
\]

\[
\begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix} \mathcal{D} \begin{pmatrix} x^* \\ x^* \end{pmatrix} + \frac{4}{(\sqrt{\pi} + \sqrt{L})^2} \left(1 - \frac{1}{\sqrt{k+1}}\right)^k \Sigma_3^{1/2} V \text{ for large } k.
\]

4 Central Limit Theorems on Polynomial Batch-size.

There are many settings where a geometric increase in \( N_k \) is impractical. For instance, this is true when the generation of a sampled gradient is computationally expensive; an example of this commonly arises in simulation optimization problems in the context of large manufacturing or queueing simulations. To this end, we consider the use of polynomial increases in sample-size, an avenue that allows for more gentle growth, and proceed to investigate the rate of convergence and the associated central limit statements for Algorithms 1–3. Polynomial rates have been studied in [33] as well as [45] but remain unaware of rate and complexity statements as well as the ensuing CLTs in accelerated and heavy-ball regimes.
4.1 Rate of Convergence.

We first recall some preliminary results that find utility in the rate analysis of the proposed algorithms with the polynomially increasing batch-sizes.

Lemma 7  (i) [34, Eqn. (17)] For any \( q \in (0, 1) \) and \( v > 0 \), there holds

\[
\sum_{t=1}^{k} q^{k-t} t^{-v} \leq q^k \frac{2vq^{-1} - 1}{1 - q} + \frac{2k^{-v}}{q \ln(1/q)}.
\]

(ii) [34, Lemma 4] For any \( q \in (0, 1) \) and \( v > 0 \), \( q^v \leq c_{q,v} x^{-v} \) for all \( x > 0 \) where \( c_{q,v} \triangleq e^{-v \left( \frac{v}{\ln(1/q)} \right)^v} \).

Based on Lemmas 1-3 in Section 2 and Lemma 7, we can establish polynomial rates of convergence of the iterates generated by the three proposed methods. Omitted proofs are included in the supplementary material for purposes of completeness.

Proposition 4 (Rate statement for Algorithm 1 under polynomially increasing \( N_k \)) Suppose Assumption 1 holds and that \( N_k \triangleq \lceil (k + 1)^v \rceil \) for some \( v > 0 \). Consider Algorithm 1 with \( \alpha \in (0, \frac{2}{\eta + L}] \). Define \( q \triangleq 1 - \frac{2\alpha \eta L}{\eta + L} \) and \( c_{q,v} \triangleq e^{-v \left( \frac{v}{\ln(1/q)} \right)^v} \). Then

\[
\mathbb{E}[\|x_k - x^*\|^2] \leq \left( c_{q,v} \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\alpha^2 v^2 c_{q,v} (e^{2vq^{-1} - 1})}{1 - q} + \frac{2\alpha^2 v^2}{q \ln(1/q)} \right) k^{-v}, \quad \forall k \geq 1.
\]

Then the number of iterations and sampled gradients required to obtain an \( \epsilon \)-optimal solution in the mean-squared sense (i.e. \( \mathbb{E}[\|x - x^*\|^2] \leq \epsilon \)) are \( \mathcal{O}(v(1/\epsilon)^{1/v}) \) and \( \mathcal{O}(e^{v/v} (1/\epsilon)^{1+1/v}) \), respectively.

Proposition 5 (Rate statement for Algorithm 2 under polynomially increasing \( N_k \)) Let Algorithm 2 be applied to (1), where \( N_k \triangleq \lceil (k + 1)^v \rceil \) with some \( v > 0 \). Suppose Assumption 1 holds and \( \alpha \in (0, \frac{1}{L}] \). Define \( \gamma \triangleq \sqrt{\alpha \eta} \) and \( \beta \triangleq \frac{1 - \gamma}{1 + \gamma} \). Then there exists a constant \( C(v) > 0 \) such that

\[
\mathbb{E}[f(y_k)] - f^* \leq C(v) k^{-v}, \quad \forall k \geq 1.
\]

Remark 7 Since \( f(x) - f(x^*) \geq \frac{1}{2} \|x - x^*\|^2 \), we have \( \mathbb{E}[\|y_k - x^*\|^2] \leq \frac{2C(v)}{\eta} k^{-v} \). Then from (10b) it follows that \( \mathbb{E}[\|x_k - x^*\|^2] \leq 2(1 + \beta)^2 \mathbb{E}[\|y_k - x^*\|^2] + 2\beta^2 \mathbb{E}[\|y_{k-1} - x^*\|^2] \leq \frac{c}{k^{-v}} \) for some \( c > 0 \). Thus, the mean-squared error of Algorithm 2 also displays the polynomial rate of convergence similar to that shown in Proposition 4 for Algorithm 1. Because the mean-squared convergence implies converges in probability, the sequences \( \{x_k\} \) and \( \{y_k\} \) generated by Algorithm 2 satisfy \( x_k \xrightarrow{P} x^* \) and \( y_k \xrightarrow{P} x^* \) when the conditions of Proposition 5 hold.

Proposition 6 (Rate Statement for Algorithm 3 under polynomially increasing \( N_k \)) Suppose that Assumption 1 holds and that \( f(\cdot) \) is twice continuously differentiable. Consider Algorithm 3, where \( \alpha \triangleq \frac{4}{(\sqrt{\eta} + \sqrt{L})^2}, \) \( \beta \triangleq (\frac{\sqrt{\eta} - 1}{\sqrt{\eta} + 1})^2 \), and \( N_k \triangleq \lceil (k + 1)^v \rceil \) with \( v > 0 \). Then there exists a constant \( C(v) > 0 \) such that

\[
\mathbb{E} \left[ \left\| \frac{x_{k+1} - x^*}{x_k - x^*} \right\|^2 \right] \leq C(v) (k + 1)^{-v}, \quad \forall k \geq 0.
\]
4.2 Central Limit Theorems under Polynomially Increasing $N_k$.

We now prove a preliminary lemma that will be used in establishing the CLT.

**Lemma 8** For any $k \geq 1$, define $A_0 = A$ and $A_k \triangleq \left(\frac{k+1}{k}\right)^{v/2} A$. Suppose there exists $\varrho \in (0, 1)$ such that $\|A\|_2 \leq \varrho$. Define $\Phi_{k,j} \triangleq A_k \cdots A_j$ with $\Phi_{j+1} \triangleq I_m$. Then for any $a > 0$, there holds

$$\sup_{k \geq 1} \sum_{t=0}^{k} \|\Phi_{k,t} \|^a \leq c_{a,v} a^{v/2} \frac{e^{a\varrho} - 1}{\varrho^a(1 - \varrho^a)} + \frac{2}{a\varrho^{2a} \ln(1/\varrho)}. \quad (46)$$

**Proof.** By the definition of $\Phi_{k,j}$ and $A_k$, we have $\Phi_{k,t+1} = \left(\frac{k+1}{t+1}\right)^{v/2} A^{k-t}$. Hence $\|\Phi_{k,t+1}\| \leq \left(\frac{k+1}{t+1}\right)^{v/2} \varrho^{k-t}$.

By defining $\tilde{q} \triangleq q^a$, $\tilde{\alpha} \triangleq av/2$, and using Lemma 7(i), we obtain that

$$\sum_{t=0}^{k} \|\Phi_{k,t} \|^a \leq \sum_{t=0}^{k} \left(\frac{k+1}{t+1}\right)^{av/2} q^{a(k+1)-(t+1)} \leq \tilde{q}^{-1} \sum_{t=0}^{k+1} t^{-\tilde{\alpha}} \tilde{\varrho}^{k+1-t} \leq \tilde{q}^{-1} (k+1)^{\tilde{\alpha}} \tilde{\varrho}^{k+1} \sum_{t=1}^{k+1} t^{-\tilde{\alpha}} \tilde{\varrho}^{k+1-t} \leq \tilde{q}^{-1} (k+1)^{\tilde{\alpha}} \tilde{\varrho}^{k+1} \sum_{t=1}^{k+1} t^{-1} \frac{2}{\tilde{\varrho} \ln(1/\tilde{\varrho})} \leq c_{\tilde{q},\tilde{\alpha}} \frac{e^{2\tilde{\varrho}-1}}{\tilde{\varrho} \ln(1/\tilde{\varrho})} \varrho^{k+1} \sum_{t=1}^{k+1} t^{-1} \frac{2}{\tilde{\varrho} \ln(1/\tilde{\varrho})}, \quad \forall k \geq 1.

Then by substituting $\tilde{q} \triangleq q^a$ and $\tilde{\alpha} \triangleq av/2$ into the above inequality, we prove the result. \[\square\]

Next, we establish the asymptotic normality of a time-varying linear recursion and provide the proof in Appendix C. This result will be applied in proving Theorems 4-6.

**Lemma 9** Suppose that $\|A\| \leq \varrho$ for some $\varrho \in (0, 1)$, $N_k = \lceil (k+1)^v \rceil$ with $v > 0$, and $\{w_{k,N_k}\}$ satisfies Assumption 2(ii). Let the sequence $\{e_k\}$ be generated by

$$e_{k+1} = A_k e_k - \alpha(k+1)^{v/2} G w_{k,N_k} + \zeta_{k+1}, \quad E[\|e_0\|^2] < \infty, \quad (47)$$

where $A_0 = A$, $A_k \triangleq \left(\frac{k+1}{k}\right)^{v/2} A$ for any $k \geq 1$, and $\zeta_{k+1} \xrightarrow{\mathbb{P}} 0$. Then

$$\alpha^{-1} e_k \xrightarrow{\mathbb{P}, k \to \infty} N(0, \Sigma) \text{ with } \Sigma \triangleq \lim_{k \to \infty} \sum_{t=1}^{k} \left(\frac{k}{t}\right)^{v} A^{-t} G S_0 G^T \left(A^T\right)^{k-t}.$$

Based on Proposition 4 and Lemma 9, by using Assumption 2, we are now ready to derive the associated central limit theorem for Algorithm 1 with polynomially increasing batch-sizes.

**Theorem 4** (CLT of Algorithm 1 with Polynomially increasing Batch-sizes) Suppose Assumptions 1 and 2 hold. Consider Algorithm 1, where $\alpha \in (0, \frac{2}{v+1}]$ and $N_k \triangleq \lceil (k+1)^v \rceil$ with some $v > 0$. Define $A \triangleq I_m - \alpha H$. Then

$$\alpha^{-1} k^{v/2} (x_k - x^*) \xrightarrow{d, k \to \infty} N(0, \Sigma) \text{ with } \Sigma \triangleq \lim_{k \to \infty} \sum_{t=1}^{k} \left(\frac{k}{t}\right)^{v} A^{-t} S_0 A^{k-t}.$$
Proof. We begin by noting that \( x_k \xrightarrow{k \to \infty} x^* \) and (32) holds. Define \( e_0 \triangleq x_0 - x^* \), \( e_k \triangleq k^{v/2}(x_k - x^*) \) for any \( k \geq 1 \), and \( \zeta_{k+1} \triangleq -\alpha(k+1)^{v/2}\delta(x_k) \) for any \( k \geq 0 \). By multiplying both sides of (32) with \((k+1)^{v/2}\), and using \( A_k = (k+1)^{v/2}(I_m - \alpha H) \), we achieve

\[
\begin{align*}
e_{k+1} &= \left(\frac{k+1}{k}\right)^{v/2} (I_m - \alpha H)k^{v/2}(x_k - x^*) - \alpha(k+1)^{v/2}\delta(x_k) + w_{k,N_k} \\
&= A_k e_k - \alpha(k+1)^{v/2} w_{k,N_k} + \zeta_{k+1}, \quad \forall k \geq 1.
\end{align*}
\]

By setting \( k = 0 \) in (32), and using \( A_0 = A = I_m - \alpha H \), we see that

\[
e_1 = x_1 - x^* = A_0(x_0 - x^*) - \alpha w_{1,N_1} - \alpha \delta(x_0).
\]

Then by \( \zeta_1 = -\alpha \delta(x_0) \), we see that the (48) also holds for \( k = 0 \). Hence the recursion (48) holds for any \( k \geq 0 \). From (39) it follows that the symmetric matrix \( A \) satisfies

\[
\|A\|_2 = \|I_m - \alpha H\|_2 = \rho^{1/2}\|P_1\|_2 \leq q^{1/2} = \left(1 - \frac{2\alpha\eta L}{\eta + L}\right)^{1/2} \triangleq q.
\]

We conclude from Proposition 4 that \( \mathbb{E}[k^{v}\|x_k - x^*\|^2] \leq c \) for some constant \( c > 0 \), and \( x_k - x^* = o_P(1) \). Similarly to the procedures for proving (38), we can show that \( e_k = k^{v/2}(x_k - x^*) = O_P(1) \), \( \delta(x_k) = o_P(\|x_k - x^*\|) = D_k(x_k - x^*) \) with \( D_k = o_P(1) \), and hence

\[
\zeta_{k+1} = -\alpha(k+1)^{v/2}\delta(x_k) = -\alpha \left(1 + \frac{1}{k}\right)^{v/2} D_k e_k \xrightarrow{k \to \infty} 0.
\]

Therefore, by using Lemma 9 with \( G = I_m \), we conclude that \( \alpha^{-1}e_k \xrightarrow{k \to \infty} N(0, \Sigma_4) \). By the definition \( e_k = k^{v/2}(x_k - x^*) \), we obtain the result.

Similarly to Theorem 4, based on Proposition 5 and Lemma 9, we now establish the central limit theorem for Algorithm 2 under the assumption of polynomially increasing batch-sizes.

**Theorem 5 (CLT of Algorithm 2 under Polynomially increasing Batch-sizes)** Suppose Assumptions 1 and 2 hold. Consider Algorithm 2, where \( \alpha \in (0, \frac{1}{\gamma}] \) and \( N_k = \lceil(k+1)^v\rceil \) with some \( v > 0 \). Define \( \gamma \triangleq \sqrt{\alpha \eta} \), \( \beta \triangleq \frac{1 - \gamma}{1 + \gamma} \), and \( H_2 \triangleq \frac{(1 + \beta)(I_m - \alpha H) - \beta(I_m - \alpha H)}{I_m} \).

Then

\[
\alpha^{-1}k^{v/2} \begin{pmatrix} y_k - x^* \\ y_{k+1} - x^* \end{pmatrix} \xrightarrow{k \to \infty} N(0, \Sigma_5) \text{ with } \Sigma_5 \triangleq \lim_{t \to \infty} \sum_{t=0}^k \left(\frac{k}{t}\right)^v H_2^{k-t} \left(\begin{array}{cc} S_0 & 0_m \\ 0_m & 0_m \end{array}\right) (H_2^T)^{k-t}.
\]

Proof. Define \( z_{k+1} \triangleq \begin{pmatrix} y_{k+1} - x^* \\ y_k - x^* \end{pmatrix} \), \( e_0 \triangleq z_0 \), and \( e_k \triangleq k^{v/2}z_k \) for any \( k \geq 1 \). Therefore, by defining \( \varsigma_{k+1} \triangleq -\alpha(k+1)^{v/2}\delta(x_k) \), and multiplying both sides of (42) by \((k+1)^{v/2}\), we obtain that

\[
\epsilon_{k+1} = \left(\frac{k+1}{k}\right)^{v/2} H_2 \epsilon_k - \alpha(k+1)^{v/2} \begin{pmatrix} I_m \\ 0_m \end{pmatrix} w_{k,N_k} + \begin{pmatrix} \varsigma_{k+1} \\ 0 \end{pmatrix}.
\]
By defining $G \triangleq \left( \begin{array}{c} I_m \\ 0_m \end{array} \right)$, $A_0 \triangleq H_2$, and $A_k \triangleq (\frac{k+1}{k})^{v/2}H_2$, there holds

$$\varepsilon_{k+1} = A_k \varepsilon_k - \alpha (k+1)^{v/2}Gw_{k,N_k} + \left( \frac{\xi_{k+1}}{0} \right), \quad \forall k \geq 1. \quad (50)$$

From (42) it is seen that (50) holds for $k = 0$ as well. Thus, the recursion (50) holds for any $k \geq 0$.

Note from Remark 7 that $\mathbb{E}[k^v \|x_k - x^*\|^2] \leq c$ for some constant $c > 0$, and $x_k - x^* = o_P(1)$. Similarly to the procedures for proving (38), we can show that $e_k = k^{v/2}(x_k - x^*) = O_P(1)$, $\delta(x_k) = o_P(\|x_k - x^*\|)$, and hence $\zeta_{k+1} = -\alpha (k+1)^{v/2}\delta(x_k) = -\alpha (1 + \frac{1}{k})^{v/2}D_k e_k \frac{P}{k \rightarrow \infty} 0$. Then by using Lemma 9 and (44), we obtain that

$$\alpha^{-1} \varepsilon_k \xrightarrow{d} N(0, \Sigma_6), \quad \text{where} \quad \Sigma_6 \triangleq \lim_{k \rightarrow \infty} \sum_{t=0}^{k} \left( \frac{k}{t} \right)^v H_2^{k-t}G_0 G^T (H_2^T)^{t-k}.$$

Then by the fact that $G_0 G^T = \left( \begin{array}{cc} S_0 & 0_m \\ 0_m & 0_m \end{array} \right)$ and $\varepsilon_k = k^{v/2} \left( \begin{array}{c} y_k - x^* \\ y_{k-1} - x^* \end{array} \right)$, we prove the result. \square

Remark 8 (i) Suppose we set $\alpha = \frac{2}{\sqrt{\kappa + \eta}}$ in Algorithm 1. Then $q = 1 - \frac{2\eta L}{\eta + L} = \left( \frac{\kappa - 1}{\kappa + 1} \right)^2$. By (39) it is seen that the matrix $A$, defined in Theorem 4, satisfies $\|A\|_2 \leq q^{1/2} = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\sqrt{\kappa + 1}}$.

(ii) Suppose $\alpha = \frac{1}{L}$ in Algorithm 2. Then $\gamma = 1/\sqrt{\kappa}$ and by (44) we know that $H_2$, defined in Theorem 2, satisfies $\|H_2\|_2 = 1 - \gamma = 1 - \frac{1}{\sqrt{\kappa}}$.

(iii) From Theorems 4 and 5, we note that both the unaccelerated gradient method (Algorithm 1) and its accelerated counterpart (Algorithm 2) have convergence rates with the same order given by $\|x_k - x^*\| = \mathcal{O}(k^{-v/2})$; however the accelerated scheme has a smaller constant than its unaccelerated counterpart since $\|H_2\|_2 < \|A\|_2$ due to the fact that $\frac{2}{\kappa + 1} \leq \frac{1}{\sqrt{\kappa}}$.

Akin to the proof of Theorem 2, based on Proposition 3 and Lemma 6, we now state the central limit theorem for Algorithm 3 with polynomially increasing batch-size (proof omitted).

Theorem 6 (CLT of Algorithm 3 with polynomially increasing $N_k$) Let Assumptions 1 and 2 hold, and $f(\cdot)$ be twice continuously differentiable. Consider Algorithm 3, where $\alpha \triangleq \frac{4}{\left(\sqrt{\kappa + \sqrt{\zeta}}\right)^2}$, $\beta \triangleq \left( \frac{\sqrt{\kappa - 1}}{\sqrt{\kappa + 1}} \right)^2$,

and $N_k = \left\lceil (k+1)^v \right\rceil$, $v > 0$. Set $H_3 \triangleq \left( \begin{array}{cc} (1+\beta)I_m - \alpha H_m \quad -\beta I_m \\ I_m \quad 0_m \end{array} \right)$. Then

$$\alpha^{-1} k^{v/2} \left( \frac{x_k - x^*}{x_{k-1} - x^*} \right) \xrightarrow{d} N(0, \Sigma_6) \quad \text{with} \quad \Sigma_6 \triangleq \lim_{k \rightarrow \infty} \sum_{t=0}^{k} \left( \frac{k}{t} \right)^v H_3^{k-t} \left( \begin{array}{cc} S_0 & 0_m \\ 0_m & 0_m \end{array} \right) (H_3^T)^{t-k}.$$

5 Confidence Regions of the Optimal Solution.

A crucial motivation for developing CLTs lies in developing confidence statements. In this section, we proceed to construct the confidence regions for the optimal solution $x^*$. Note that the limiting covariance
matrix is dependent on the Hessian at the solution, which is unavailable. Furthermore, we do not have a consistent estimate of the covariance matrix. Yet, in the absence of such an estimate, we proceed to develop rigorous confidence statements, adopting an approach developed in [26].

Since Algorithms 1, 2, and 3 lead to similar central limit results, we show how to construct confidence regions merely for the sequence \( \{x_k\} \) generated by Algorithm 1. The simulation framework in [26] lies in generating \( n \) independent replications of Algorithm 1, leading to \( n \) copies of the random iterate \( x_k \), denoted by \( x_{1k}, \ldots, x_{nk} \). Then the sample mean and the covariance estimator are respectively defined as

\[
\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \quad \text{and} \quad S_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)^T.
\] (51)

Based on [6, Theorem 4 and Corollary 1] and [26, Theorem 2], we may achieve the following result. The proof is given in Appendix D for completeness.

**Proposition 7** Consider Algorithm 1. Suppose that the conditions of Theorem 1 hold and \( n \geq m + 1 \). Then

(i) \( \sqrt{n} \alpha^{-1} \beta_k^{k/2} (\bar{x}_k - x^*) \xrightarrow{d} N(0, \Sigma_1) \).

(ii) \( n(\bar{x}_k - x^*)^T S_k^{-1} (\bar{x}_k - x^*) \xrightarrow{d} \frac{m(n-1)}{n-m} F(m, n-m) \), where \( F(m, n-m) \) denotes the \( F \)-distribution with \( (m, n-m) \) degrees of freedom.

Proposition 7 can be used to construct the confidence region of the optimal solution. Define

\[
X_{mk}(z) \triangleq \left\{ x : n(\bar{x}_k - x)^T S_k^{-1} (\bar{x}_k - x) \leq \frac{m(n-1)}{n-m} z \right\},
\] (52)

where \( z \) is selected such that \( \mathbb{P}(F(m, n-m) \leq z) \geq 1 - \delta \) with some \( \delta \in (0, 1) \). Then we have the following corollary.

**Corollary 2** (26, Proposition 3) Consider Algorithm 1. Suppose that \( n \geq m + 1 \) and the conditions of Theorem 1 hold. Then the confidence region \( X_{mk}(z) \) defined in (52) is asymptotically correct, i.e.,

\[
\lim_{k \to \infty} \mathbb{P}(x^* \in X_{mk}(z)) = 1 - \delta.
\]

The above result asserts that the estimated confidence region \( X_{mk}(z) \) asymptotically covers the optimal solution \( x^* \) with probability \( 100(1 - \delta)\% \). The approach is easily implementable because it merely requires \( n \) independent replications of Algorithm 1, while without requiring a consistent estimator of the covariance matrix of the stationary normal distribution. The confidence regions of Algorithm 2 and Algorithm 3 can be constructed in a similar way.

In the following, we show that the sequence \( \sqrt{\frac{n(n-m)}{n-1}} S_k^{-1/2} (\bar{x}_k - x^*) \) converges to a suitably defined multivariate \( t \) distribution, the proof of which is given in Appendix E.

**Proposition 8** Consider Algorithm 1. Suppose that the conditions of Theorem 1 hold and \( n > m + 2 \). Then

\[
\sqrt{\frac{n(n-m)}{n-1}} S_k^{-1/2} (\bar{x}_k - x^*) \xrightarrow{d} T_{n-m}(0, I_m, m),
\] (53)
where a $m$-variate random vector $X \sim T_n(\mu, \Lambda, m)$ (i.e., $X$ is a $t$ distribution with mean $\mu$, covariance matrix $v(v-2)^{-1}\Lambda$, $v > 2$) if it has a probability density function $f$ given by

$$
f(x) = \frac{\Gamma((v + m)/2)}{(\pi v)^{v/2}\Gamma(v/2)|\Lambda|^{1/2}} \left\{ 1 + \frac{(x-\mu)^T \Lambda^{-1} (x-\mu)}{v} \right\}^{-\frac{v+m}{2}}, v > 2. \tag{54}
$$

The above result can also be adopted to construct the confidence regions. Define

$$
\tilde{X}_{mk}(U) \triangleq \left\{ x : \sqrt{\frac{n(n-m)}{n-1}} S_k^{-1/2}(\bar{x}_k-x) \in U \right\},
$$

where the region $U$ is selected such that $\mathbb{P}(T_{n-m}(0, I_m, m) \in U) \geq 1 - \delta$ with some $\delta \in (0, 1)$. Similarly to Corollary 2, the confidence region $\tilde{X}_{mk}(U)$ defined above is asymptotically correct, i.e., $\lim_{k \to \infty} \mathbb{P}(x^* \in \tilde{X}_{mk}(U)) = 1 - \delta$. However, it might be harder to construct the confidence regions from the multivariate $t$ distribution than that from the $F$-distribution since the region $U$ might not be easily obtained.

### 6 Numerical Simulations.

In this section, we carry out simulations for the parameter estimation problem. We aim to estimate the unknown $m$-dimensional parameter $x^*$ based on the gathered scalar measurements $\{d_k\}_{k \geq 1}$ given by $d_k = u_k^T x^* + \nu_k$, where $u_k \in \mathbb{R}^m$ denotes the regression vector and $\nu_k \in \mathbb{R}$ denotes the local observation noise. Assume that $\{u_k\}$ and $\{\nu_k\}$ are mutually independent i.i.d. Gaussian sequences with distributions $N(0, R_u)$ and $N(0, \sigma^2_u)$, respectively. Suppose the covariance matrix $R_u$ is positive definite. Then we might model the parameter estimation problem as the following stochastic optimization problem:

$$
\min_{x \in \mathbb{R}^m} f(x) \triangleq \mathbb{E}\left[ \|d_k - u_k^T x\|^2 \right]. \tag{55}
$$

Thus, $f(x) = (x-x^*)^T R_u (x-x^*) + \sigma_u^2$ and $\nabla f(x) = R_u (x-x^*)$. Because the Hessian matrix $R_u$ of the objective function $f(\cdot)$ is positive definite, $x^*$ is the unique optimal solution to (55). Suppose we are able to observe the regressor $u_k$ and the measurement $d_k$, then the noisy observation of the gradient $\nabla f(x)$ might be constructed as $u_k^T x - d_k u_k$. Set the dimension of $x^*$ as $m = 5$. We run Algorithm 1 with $\alpha = \frac{2}{L + \eta}$, Algorithm 2 with $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{\eta - 1}}{\sqrt{\eta + 1}}$, and Algorithm 3 with $\alpha = \frac{4}{(\sqrt{\eta + L})^2}$ and $\beta = \left(\frac{\sqrt{\eta - 1}}{\sqrt{\eta + 1}}\right)^2$, where the initial values $x_0 = y_0 = 0$ and the batch-size $N_k = \lfloor \rho^{-k} \rfloor$ with $\rho = \frac{k^2}{(k+1)^2}$.

**Convergence rate, iteration and oracle complexity.** We run Algorithms 1, 2, 3, and the standard SGD algorithm $x_{k+1} = x_k - \alpha_k \nabla f(x_k, \xi_k)$ with $\alpha_k = R_u^{-1}/k$ setting to be the optimal tuning steplength, and terminate the schemes when $\frac{\mathbb{E}\|x_k - x^*\|^2}{\|x^*\|^2} \leq 10^{-3}$. We then examine their empirical rate of convergence, iteration and oracle complexity. Here the empirical mean is calculated by averaging across 100 trajectories. The convergence rate of the relative error $\frac{\mathbb{E}\|x_k - x^*\|^2}{\|x^*\|^2}$ is shown in Figure 1, which demonstrates that the iterates generated by Algorithms 1-3 converge in mean to the optimal solution at a linear rate. We see that the accelerated scheme (Algorithm 2) has the fastest empirical rate, while the heavy ball method (Alg. 3)
tends to stabilize later in the process. The empirical relationship between the accuracy $\epsilon$ and $K(\epsilon)$ is shown in Figure 2, where $K(\epsilon)$ denotes the number of iterations required to make $\frac{\mathbb{E}[\|x_k - x^*\|_2]}{\|x^*\|_2} < \epsilon$. It is seen that the standard SGD algorithm requires far more iterations than the proposed variance reduced schemes for obtaining an approximate solution with the same accuracy. The empirical relationship between $\epsilon$ and $O(\epsilon)$ is shown in Figure 3, where $O(\epsilon)$ denotes the number of sampled gradients required to make $\frac{\mathbb{E}[\|x_k - x^*\|_2]}{\|x^*\|_2} < \epsilon$. We observe that for obtaining an estimate with the same accuracy, the accelerated scheme (Alg. 2) requires the smallest number of sampled gradients, while the variable sample-size SGD method (Alg. 1) requires more sampled gradients than standard SGD.

**Limiting distributions.** We run Algorithms 1, 2, and 3 by 1000 independent sample paths and terminate the schemes at $k = 50$. The empirical largest eigenvalue of the covariance matrix $S_k$ (estimated by (51)) at $k = 50$ are $1.648 \times 10^{-5}$, $1.649 \times 10^{-6}$, $1.825 \times 10^{-6}$, respectively. This might imply that the accelerated scheme (Algorithm 2) has the best performance measured by the covariance of the stationary distribution.
Since the unknown parameter $x^*$ is multi-dimensional, we merely show the limiting distribution of one component of the rescaled error $\rho^{-k/2}(x_k - x^*)$. The histograms of $\rho^{-k/2}(x_k^5 - x^*)$ at $k = 50$ are shown in Figure 4 along with the fitted normal distribution (the red curve), where $x_k^5$ denotes the fifth component of $x_k$. It is also seen that the rescaled error is (approximately) normally distributed and among them, the accelerated scheme has the smallest variance. In addition, the histograms of the rescaled suboptimality gap $\rho^{-k}(f(x_k) - f(x^*))$ at $k = 50$ are shown in Figure 5. We can further conclude that the empirical sub-optimality gap of the accelerated scheme is the best among the proposed variable sample-size schemes.

We further run the SGD method $x_{k+1} = x_k - R_u^{-1}\nabla f(x_k, \xi_{j,k})/k$. The histogram along with the fitted normal distribution of $\sqrt{N}(x_N^5 - x^*)$ is displayed in Figure 6, while the empirical sub-optimality gap $N(f(x_N) - f(x^*))$ is shown in Figure 7, where $N \triangleq \sum_{k=1}^{50} N_k$ denotes the number of sampled gradients utilized by Algorithms 1-3.

![Figure 6: Hist. along fitted normal for $\sqrt{N}(x_N^5 - x^*)$](image)

![Figure 7: Hist. of $N(f(x_N) - f(x^*))$](image)

**Coverage probability of the constructed confidence region.** In a single replication, we generate $n$ independent sample paths for Algorithms 1, 2 and 3 and terminate each sample path when the total number of sampled gradients used reaches $N_{\text{max}}$. Then we can construct the 95% confidence region by (52) and check whether the true parameter lies in the constructed confidence region. We estimate the coverage probability (i.e., the proportion of replications that the confidence region contains the true value) by conducting 1000 replications. The estimated coverage probability for Algorithms 1-3 with different selection of sample paths $n$ and simulation budget $N_{\text{max}}$ is shown in Table 1. It can be seen that the coverage probabilities are getting closer to the nominal level of 95% when the number of sample paths $n$ grows larger. It seems that the simulation budget $N_{\text{max}}$ does not significantly impact the coverage probability; however, a larger $N_{\text{max}}$ does lead to narrowing of the confidence region.

**Polynomially increasing batch-sizes.** We run Algorithms 1-3 with the batch-size increasing at a polynomial rate $N_k = \lceil k^v \rceil$, $v = 2$. The convergence rate of the relative error $\frac{E[\|x_k - x^*\|_2]}{\|x^*\|_2}$ is shown in Figure 8, while the histograms of $k^{v/2}(x_k^5 - x^*)$ at $k = 100$ along with the fitted normal distribution are shown in Figure 10. It can be seen that the accelerated scheme has the best performance because it displays
| \( N_{\text{max}} \) | VSS-SGD \( (n=6) \) | VSS-ACC \( (n=6) \) | VSS-HB \( (n=6) \) |
|---|---|---|---|
| \( 10^3 \) | 0.671 ± 0.221 | 0.652 ± 0.2259 | 0.665 ± 0.223 |
| \( 10^4 \) | 0.904 ± 0.0869 | 0.872 ± 0.1117 | 0.893 ± 0.0995 |
| \( 10^5 \) | 0.946 ± 0.0511 | 0.923 ± 0.0591 | 0.951 ± 0.0466 |

Table 1: The estimated coverage probability of Algorithms 1-3. The ideal coverage probability is 0.95.

Figure 8: Poly. Batch-size

Figure 9: Convergence rate of VSS-ACC (Alg. 2) for different problem dimension \( m \)

Figure 10: Histograms of \( k^{\|/2} (x_k^5 - x^*) \) at \( k = 100 \)
The fastest convergence rate and the rescaled error has smallest variance.

**The effect of the problem dimension.** We take Algorithm 2 as an example. By adding further numerical simulations, we examine the impact of $m$ on algorithm performance. In particular, we implement Algorithm 2 for $m = 5, 10, 20$, where $\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa-1}}{\sqrt{\kappa+1}}$, $N_k = \max \{ \lceil m^{3/4} \rceil, \lceil \rho^{-k} \rceil \}$ with $\rho = \frac{\kappa^2}{(\kappa+1)^2}$, and the initial values $x_0 = y_0 = 0$. The empirical counterpart of the convergence rate of the relative error $E[\|x_k - x^*\|_2] \|x^*\|_2$ is shown in Figure 9, demonstrating that the larger $m$ leads to a worse rate of convergence. We additionally test the coverage probability of the constructed confidence region for Algorithm 2 for different choices of $m$. In a single replication, we run Algorithm 2 for $n = 30$ independent sample paths and terminate them when the total number of sampled gradients used reaches 5000. Then we can construct the 95% confidence region by (52) and check whether the true parameter lies in the constructed confidence region. Finally, we estimate the coverage probability by 1000 replications. The estimated coverage probabilities for $m = 5, 10, 20$ are respectively $0.956 \pm 0.0421$, $0.953 \pm 0.0448$, and $0.945 \pm 0.052$.

7 Conclusions

In this work, we considered the strongly convex stochastic optimization problem and proposed three classes of variance reduced stochastic gradient descent algorithms (unaccelerated, accelerated, and heavy ball methods), where the unavailable exact gradient is approximated by an increasing batch of sampled gradients. We then establish rate and complexity guarantees. Further, we establish amongst the first formal central limit theorems for all the three schemes when the batch-size increased at either a geometric or a polynomial rate. The covariance matrix specifies how problem structure (including the strong convexity parameter, the Lipschitz constant, and the Hessian matrix) and distribution of gradient noise influences the algorithm performance. In addition, we provide an avenue to construct the confidence region of the optimal solution based on the central limit theorems. The paper concludes with an application of the proposed schemes to the stochastic parameter estimation problem to validate our theoretical findings. Yet much remains to be understood about the how one may develop analogs of such statements in constrained and nonsmooth regimes. One avenue for addressing such challenges is via a smoothing framework [1] whereby a stochastic gradient update is taken with respect to a smoothed objective and the smoothing parameter is progressively reduced (cf. [27]). It is also of interest to explore whether the central limit result on the function value can be obtained under weaker condition. Finally, we intend to examine if confidence regions can be designed based on the batch means method [65] when generating independent trials is expensive.

References

[1] Beck, A.: First-Order Methods in Optimization. SIAM, Philadelphia, PA (2017)

[2] Bertsekas, D., Tsitsiklis, J.: Parallel and Distributed Computation: Numerical Methods. Prentice-Hall International editions. Prentice-Hall International (1989)
[3] Bickel, P., Choi, D., Chang, X., Zhang, H., et al.: Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. The Annals of Statistics 41(4), 1922–1943 (2013)

[4] Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming: Springer Series in Operations Research. Springer (1997)

[5] Blum, J.R.: Multidimensional stochastic approximation methods. The Annals of Mathematical Statistics pp. 737–744 (1954)

[6] Bodnar, T., Mazur, S., Podgórski, K.: Singular inverse wishart distribution and its application to portfolio theory. Journal of Multivariate Analysis 143, 314–326 (2016)

[7] Borkar, V.S.: Stochastic approximation: a dynamical systems viewpoint, vol. 48. Springer (2009)

[8] Bottou, L.: Online learning and stochastic approximations. On-line learning in neural networks 17(9), 142 (1998)

[9] Bubeck, S.: Theory of convex optimization for machine learning. arXiv preprint arXiv:1405.4980 15 (2014)

[10] Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. 134(1), 127–155 (2012)

[11] Chee, J., Toulis, P.: Convergence diagnostics for stochastic gradient descent with constant learning rate. In: International Conference on Artificial Intelligence and Statistics, pp. 1476–1485. PMLR (2018)

[12] Chen, H.F.: Stochastic approximation and its applications, vol. 64. Springer Science & Business Media (2006)

[13] Chen, X., Lee, J.D., Tong, X.T., Zhang, Y.: Statistical inference for model parameters in stochastic gradient descent. The Annals of Statistics 48(1), 251–273 (2020)

[14] DasGupta, A.: Asymptotic theory of statistics and probability. Springer Science & Business Media (2008)

[15] Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in neural information processing systems, pp. 1646–1654 (2014)

[16] DeGroot, M.H., Schervish, M.J.: Probability and statistics. Pearson Education (2012)

[17] Deng, G., Ferris, M.C.: Variable-number sample-path optimization. Math. Program. 117(1-2), 81–109 (2009)
[18] Dieuleveut, A., Durmus, A., Bach, F.: Bridging the gap between constant step size stochastic gradient descent and markov chains. The Annals of Statistics 48(3), 1348–1382 (2020)

[19] Fabian, V., et al.: On asymptotic normality in stochastic approximation. The Annals of Mathematical Statistics 39(4), 1327–1332 (1968)

[20] Friedlander, M.P., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM J. Scientific Computing 34(3), A1380–A1405 (2012)

[21] Gadat, S., Panloup, F., Saadane, S., et al.: Stochastic heavy ball. Electronic Journal of Statistics 12(1), 461–529 (2018)

[22] Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming 156(1-2), 59–99 (2016)

[23] Hall, P., Heyde, C.C.: Martingale limit theory and its application. Academic press (2014)

[24] Hasminskii, R.Z., Silver, B.: Stochastic approximation and recursive estimation, vol. 47. American Mathematical Soc. (1973)

[25] Homem-de-Mello, T.: Variable-sample methods for stochastic optimization. ACM Trans. Model. Comput. Simul. 13(2), 108–133 (2003)

[26] Hsieh, M.H., Glynn, P.W.: Recent advances in simulation optimization: confidence regions for stochastic approximation algorithms. In: Proceedings of the 34th conference on Winter simulation: exploring new frontiers, pp. 370–376. Winter Simulation Conference (2002)

[27] Jalilzadeh, A., Shanbhag, U.V., Blanchet, J.H., Glynn, P.W.: Optimal smoothed variable sample-size accelerated proximal methods for structured nonsmooth stochastic convex programs. arXiv:1803.00718v2 (2022 (To appear in Stochastic Systems))

[28] Jofré, A., Thompson, P.: On variance reduction for stochastic smooth convex optimization with multiplicative noise. Math. Program. 174(1-2), 253–292 (2019)

[29] Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in neural information processing systems, pp. 315–323 (2013)

[30] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)

[31] Kushner, H., Yin, G.G.: Stochastic approximation and recursive algorithms and applications, vol. 35. Springer Science & Business Media (2003)

[32] Lamm, M., Lu, S., Budhiraja, A.: Individual confidence intervals for solutions to expected value formulations of stochastic variational inequalities. Math. Program. 165(1), 151–196 (2017)
[33] Lei, J., Shanbhag, U.V.: Asynchronous variance-reduced block schemes for composite non-convex stochastic optimization: block-specific steplengths and adapted batch-sizes. Optimization Methods and Software pp. 1–31 (2020)

[34] Lei, J., V. Shanbhag, U.: Distributed variable sample-size gradient-response and best-response schemes for stochastic Nash equilibrium problems over graphs. arXiv:1811.11246v2 (2020)

[35] Lin, P.E.: Some characterizations of the multivariate t distribution. Journal of Multivariate Analysis 2(3), 339–344 (1972)

[36] Lu, S.: A new method to build confidence regions for solutions of stochastic variational inequalities. Optimization 63(9), 1431–1443 (2014)

[37] Mandt, S., Hoffman, M.D., Blei, D.M.: Stochastic gradient descent as approximate Bayesian inference. The Journal of Machine Learning Research 18(1), 4873–4907 (2017)

[38] Meyn, S.P., Tweedie, R.L.: Markov chains and stochastic stability. Springer Science & Business Media (2012)

[39] Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization 19(4), 1574–1609 (2009)

[40] Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience publication. Wiley (1983)

[41] Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2013)

[42] Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In: Dokl. akad. nauk Sssr, vol. 269, pp. 543–547 (1983)

[43] Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research. Springer-Verlag, New York (1999)

[44] Pasupathy, R.: On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. Oper. Res. 58(4-Part-1), 889–901 (2010)

[45] Pasupathy, R., Glynn, P.W., Ghosh, S., Hashemi, F.S.: On sampling rates in simulation-based recursions. SIAM J. Optimization 28(1), 45–73 (2018)

[46] Pasupathy, R., Kim, S.: The stochastic root-finding problem: Overview, solutions, and open questions. ACM Transactions on Modeling and Computer Simulation (TOMACS) 21(3), 19 (2011)

[47] Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics 4(5), 1–17 (1964)
[48] Polyak, B.T.: Introduction to optimization. Optimization Software, Inc., Publications Division, New York (1987)

[49] Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30(4), 838–855 (1992)

[50] Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)

[51] Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Optimizing methods in statistics, pp. 233–257. Elsevier (1971)

[52] Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in neural information processing systems, pp. 1458–1466 (2011)

[53] Shanbhag, U.V., Blanchet, J.H.: Budget-constrained stochastic approximation. In: Proceedings of the 2015 Winter Simulation Conference, pp. 368–379. IEEE Press (2015)

[54] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming: modeling and theory. SIAM (2009)

[55] Spall, J.C.: Adaptive stochastic approximation by the simultaneous perturbation method. IEEE transactions on automatic control 45(10), 1839–1853 (2000)

[56] Su, W.J., Zhu, Y.: Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. arXiv preprint arXiv:1802.04876 (2018)

[57] Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International conference on machine learning, pp. 1139–1147 (2013)

[58] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)

[59] Toulis, P., Airoldi, E., Rennie, J.: Statistical analysis of stochastic gradient methods for generalized linear models. In: International Conference on Machine Learning, pp. 667–675 (2014)

[60] Van der Vaart, A.W.: Asymptotic statistics, vol. 3. Cambridge university press (2000)

[61] Venter, J., et al.: An extension of the robbins-monro procedure. The Annals of Mathematical Statistics 38(1), 181–190 (1967)

[62] Wei, C., et al.: Multivariate adaptive stochastic approximation. The Annals of Statistics 15(3), 1115–1130 (1987)

[63] Yang, T., Lin, Q., Li, Z.: Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. arXiv preprint arXiv:1604.03257 (2016)
It is noticed from (12), \( \nabla f(x^*) = 0 \), and the mean-value theorem that

\[
x_{k+1} - x^* = x_k - x^* + \beta(x_k - x_{k-1}) - \alpha(\nabla f(x_k) - \nabla f(x^*)) - \alpha w_{k,N_k}
\]

\[
x_k - x^* + \beta(x_k - x_{k-1}) - \alpha \nabla^2 f(z_k)(x_k - x^*) - \alpha w_{k,N_k}
\]

\[
= ((1 + \beta)I_m - \alpha \nabla^2 f(z_k))(x_k - x^*) - \beta(x_{k-1} - x^*) - \alpha w_{k,N_k}
\]

for some \( z_k \) on the line segment between \( x_k \) and \( x^* \). We then write this recursion in matrix form.

\[
\begin{pmatrix}
x_{k+1} - x^*
\end{pmatrix} =
\begin{pmatrix}
(1 + \beta)I_m - \alpha \nabla^2 f(z_k) & -\beta I_m \\
I_m & 0_m
\end{pmatrix}
\begin{pmatrix}
x_k - x^*
\end{pmatrix} + \begin{pmatrix}
-\alpha w_{k,N_k}
\end{pmatrix}.
\]

By the eigenvalue decomposition, \( \nabla^2 f(z_k) = U \Lambda U^T \), where \( U \) is orthogonal and \( \Lambda = \text{diag} \{ \lambda_1, \lambda_2, \cdots, \lambda_m \} \) with \( \lambda_i \in [\eta, L], i = 1, \cdots, n \) being the eigenvalues of \( \nabla^2 f(z_k) \). This allows us to get

\[
\|T_k\| = \left\| \begin{pmatrix}
(1 + \beta)I_m - \alpha U \Lambda U^T & -\beta I_m \\
I_m & 0_m
\end{pmatrix} \right\|
\]

\[
= \left\| \begin{pmatrix}
(1 + \beta)I_m - \alpha \Lambda & -\beta I_m \\
I_m & 0_m
\end{pmatrix} \right\| = \max_{i \in [1:m]} \left\| \begin{pmatrix}
1 + \beta - \alpha \lambda_i & -\beta \\
1 & 0
\end{pmatrix} \right\|,
\]

where \( \| \cdot \| \) denotes the Euclidean norm of a vector or a matrix, and the last equality holds because it is possible to permute the matrix to a block diagonal matrix with \( 2 \times 2 \) blocks. For each \( i = 1, \cdots, m \), the eigenvalues of the \( 2 \times 2 \) matrices are given by the roots of the characteristic equation \( p_i(v) = (v - (1 + \beta - \alpha \lambda_i))v + \beta = v^2 - (1 + \beta - \alpha \lambda_i)v + \beta = 0 \). From \( \eta \leq \lambda_i \leq L \) and \( \alpha < 4/L \) it follows that for any \( i = 1, \cdots, m : 1 - \sqrt{\alpha L} \leq 1 - \sqrt{\alpha \lambda_i} \leq 1 - \sqrt{\alpha \eta} \), hence \( |1 - \sqrt{\alpha \lambda_i}|^2 \leq \max \{|1 - \sqrt{\alpha \eta}|^2, |1 - \sqrt{\alpha L}|^2\} = \beta < 1 \). Thus, the discriminant of the equation \( p_i(v) \), denoted by \( \Delta_i \), is nonpositive.

\[
\Delta_i = (1 + \beta - \alpha \lambda_i)^2 - 4\beta = (1 - \beta)^2 - 2(1 + \beta)\alpha \lambda_i + (\alpha \lambda_i)^2
\]

\[
\leq (1 - |1 - \sqrt{\alpha \lambda_i}|^2)^2 - 2(1 + |1 - \sqrt{\alpha \lambda_i}|^2)\alpha \lambda_i + (\alpha \lambda_i)^2
\]

\[
= (1 + |1 - \sqrt{\alpha \lambda_i}|^2)^2 - 4|1 - \sqrt{\alpha \lambda_i}|^2 - 2(1 + |1 - \sqrt{\alpha \lambda_i}|^2)\alpha \lambda_i + (\alpha \lambda_i)^2
\]

\[
= (1 - \alpha \lambda_i + |1 - \sqrt{\alpha \lambda_i}|^2)^2 - 4|1 - \sqrt{\alpha \lambda_i}|^2 = 0.
\]
Hence \( p_i(v) = 0 \) has two complex roots \( \frac{1+\beta-\alpha\lambda_i}{2} \pm \frac{\sqrt{-\Delta}}{2i} \), where \( i = \sqrt{-1} \). Thus, the magnitude of the roots is \( \sqrt{\left(\frac{1+\beta-\alpha\lambda_i}{2}\right)^2 - \frac{\Delta}{4}} = \sqrt{\beta} \). Then by (57),

\[
\|T_k\| = \left\| \begin{pmatrix} (1 + \beta)I_m - \alpha A & -\beta I_m \\ I_m & 0 \end{pmatrix} \right\| \leq \sqrt{\beta}.
\]

(58)

By taking two-norm square with respect to both sides of (56), we obtain that

\[
\left\| \begin{pmatrix} x_{k+1} - x^* \\ x_k - x^* \end{pmatrix} \right\|^2 = \left( x_k - x^* \right)^T T_k^2 T_k \begin{pmatrix} x_k - x^* \\ x_{k-1} - x^* \end{pmatrix} + \left( -\alpha w_{k,N_k} \right)^T T_k \begin{pmatrix} x_k - x^* \\ x_{k-1} - x^* \end{pmatrix} + \alpha^2 \|w_{k,N_k}\|^2.
\]

Because \( x_k \) and \( x_{k-1} \) are adapted to \( F_k \), by taking expectations conditioned on \( F_k \) and using (5), (58), we obtain that

\[
E \left[ \left\| \begin{pmatrix} x_{k+1} - x^* \\ x_k - x^* \end{pmatrix} \right\|^2 \middle| F_k \right] \leq \beta \left\| \begin{pmatrix} x_k - x^* \\ x_{k-1} - x^* \end{pmatrix} \right\|^2 + \alpha^2 \frac{\nu^2}{N_k}, \text{ a.s.}
\]

(59)

Thus, by taking the unconditional expectations of (59), we obtain (13).

\[\square\]

### B Proof of Lemma 6

Define the sequence \( \{u_k\} \) as follows:

\[
u_{k+1} = Pu_k - \alpha \rho^{-(k+1)/2} G w_{k,N_k}, \quad u_0 = 0.
\]

(60)

By combining (60) with (30), we obtain the following recursion:

\[
u_{k+1} = P(u_k - \alpha \rho^{-(k+1)/2} G w_{k,N_k}) + \sum_{t=0}^{k} \rho^{k-t} \zeta_{t+1}.
\]

By using \( \|P\| \leq \rho \) and \( \zeta_k \xrightarrow[k \to \infty]{P} 0 \), we obtain the following bound:

\[
u_{k+1} - u_{k+1} \leq \|P\|^{k+1} \|e_0 - u_0\| + \sum_{t=0}^{k} \|P\|^{k-t} \|\zeta_{t+1}\| \leq \rho^{k+1} \|e_0 - u_0\| + \sum_{t=0}^{k} \rho^{k-t} o(1) \xrightarrow[k \to \infty]{P} 0.
\]

Thus, \( u_k \) defined by (30) and \( u_k \) produced by (60) have the same limit distribution if exists. This follows from the fact that \( X_k \xrightarrow{d} X \), \( \|X_k - Y_k\| \xrightarrow{P} 0 \Rightarrow Y_k \xrightarrow{d} X \) (see [60, Theorem 2.7(iv)]).

From (60) and \( u_0 = 0 \) it follows that

\[
u_{k+1} = P^{k+1} u_0 - \alpha \sum_{t=0}^{k} P^{k-t} G \rho^{-(t+1)/2} w_{t,N_t} = -\alpha \sum_{t=0}^{k} P^{k-t} G \rho^{-(t+1)/2} w_{t,N_t}.
\]

35
This implies that

$$\alpha^{-1}u_k = -\sum_{t=0}^{k-1} P^{k-1-t}G\rho^{-(t+1)/2}w_{t,N_t} = -\sum_{t=1}^{k} P^{k-t}G\rho^{-t/2}w_{t-1,N_{t-1}}. \quad (61)$$

Define $\xi_{kt} \triangleq -P^{k-t}G\rho^{-t/2}w_{t-1,N_{t-1}}$ for any $t : 1 \leq t \leq k$. We intend to apply Lemma 5. Therefore, we have to check conditions (24)-(26). By using Assumption 1(iii), $N_t \geq \rho^{-(t+1)}$ and $\|P\| < \rho$, we obtain that $\mathbb{E}[\xi_{kt}|\xi_{k1}, \ldots, \xi_{kt-1}] = 0$ and

$$\mathbb{E}[\|\xi_{kt}\|^2|\xi_{k1}, \ldots, \xi_{kt-1}] \leq \|P^{k-t}\|^2\|G\|^2\rho^{-t}\mathbb{E}[\|w_{t-1,N_{t-1}}\|^2|\mathcal{F}_{t-1}] \leq \|G\|^2\rho^{-t}N_{t-1} \leq \nu^2\|G\|^2\rho^{2(k-t)} \text{ a.s.}$$

$$\implies \sum_{t=1}^{k} \mathbb{E}[\|\xi_{kt}\|^2] \leq \nu^2\|G\|^2\sum_{t=1}^{k} \rho^{2(k-t)} = \frac{\nu^2\|G\|^2}{1 - \rho^2} < \infty.$$ 

Thus, (24) holds. By $S_{kt}$ and $R_{kt}$ defined in (23), (28), $\|P\| < \rho$, and $\rho^{-t} \leq N_{t-1}$, we have

$$\sum_{t=1}^{k} \mathbb{E}[\|R_{kt} - S_{kt}\|]$$

$$\leq \sum_{t=1}^{k} \|P^{k-t}G\|^2\mathbb{E}\left[\left\|N_{t-1}\mathbb{E}[w_{t-1,N_{t-1}}w_{t-1,N_{t-1}}^T|\mathcal{F}_{t-1}] - N_{t-1}\mathbb{E}[w_{t-1,N_{t-1}}w_{t-1,N_{t-1}}^T]\right\|\right]$$

$$\leq \|G\|^2\sum_{t=1}^{k} \rho^{2(k-t)} o(1) \xrightarrow{k \to \infty} 0.$$ 

This verifies the second equality in (25). We now verify the first equality in (25).

$$S_k = \sum_{t=1}^{k} S_{kt} = \sum_{t=1}^{k} P^{k-t}G\mathbb{E}\left[\rho^{-t}w_{t-1,N_{t-1}}w_{t-1,N_{t-1}}^T\right] (P^{k-t}G)^T$$

$$= \sum_{t=1}^{k} P^{k-t}GS_0(P^{k-t}G)^T + \sum_{t=1}^{k} P^{k-t}GS_0(P^{k-t}G)^T \left(\rho^{-t}/N_{t-1} - 1\right) \quad (62)$$

$$+ \sum_{t=1}^{k} P^{k-t}G \left(N_{t-1}\mathbb{E}[w_{t-1,N_{t-1}}w_{t-1,N_{t-1}}^T] - S_0\right) (P^{k-t}G)^T \rho^{-t}/N_{t-1},$$

where the second and last terms tend to zero by using $\sum_{t=1}^{k} \|P^{k-t}\|^2 \leq \sum_{t=1}^{k} \rho^{2t} = \frac{1}{1 - \rho^2}$, $\rho^{-t}/N_{t-1} \to 1$, and (28). The first term on the right-hand side of (62) can be written as

$$\sum_{t=0}^{k} P^{k-t}GS_0(P^{k-t}G)^T = \sum_{t=0}^{k-1} P^tGS_0G^T(P^t)^T. \quad (63)$$

By $\|P\| < \rho < 1$, we have $\sum_{t=0}^{k} \|P^tGS_0G^T(P^t)^T\| \leq \|GS_0G^T\| \sum_{t=0}^{k} \|P\|^{2t} \leq \|GS_0G^T\||(1 - \rho^2)$. Since $\sum_{t=0}^{k} \|P^tGS_0G^T(P^t)^T\|$ is a monotonically increasing and bounded series, its limit exists. As a
result, the limit of (63) exists and is denoted by $\Sigma$. Therefore,

$$
\lim_{k \to \infty} S_k \triangleq \Sigma = \lim_{k \to \infty} \sum_{t=0}^{k} P^t GS_0 G^T (P^t)^T.
$$

Finally, we have to verify the Lindeberg condition (26). By using $N_{t-1} = \lceil \rho^{-t} \rceil$, $\xi_{kt} = -P^{k-t} G \rho^{-t/2} w_{t-1,N_{t-1}}$, and $\|P\|_2 < \rho$, we obtain that

$$
\|\xi_{kt}\| \leq \|P\|^{k-t} \|G\| \rho^{-t/2} \|w_{t-1,N_{t-1}}\| \leq \|G\| \rho^{-t} \sqrt{N_{t-1}} \|w_{t-1,N_{t-1}}\|.
$$

(64)

Hence for any $\epsilon > 0$,

$$
\left\{ \|\xi_{kt}\| > \epsilon \right\} \subset \left\{ \sqrt{N_{t-1}} \|w_{t-1,N_{t-1}}\| > \epsilon \|G\|^{-1} \rho^{-(k-t)} \right\}.
$$

(65)

Because for any $t \geq 1$, $\rho^{-(k-t)} \xrightarrow{k \to \infty} \infty$. Then using (29), we obtain that

$$
\sup_{t \geq 1} \mathbb{E} \left[ N_{t-1} \|w_{t-1,N_{t-1}}\|^2 I_{\left\{ \sqrt{N_{t-1}} \|w_{t-1,N_{t-1}}\| > \epsilon \|G\|^{-1} \rho^{-(k-t)} \right\}} \right] \xrightarrow{k \to \infty} 0.
$$

Consequently, for any $\epsilon > 0$, by using (64) and (65), the following holds:

$$
\sum_{t=1}^{k} \mathbb{E} \left[ \|\xi_{kt}\|^2 I_{\|\xi_{kt}\| \geq \epsilon} \right] \leq \sum_{t=1}^{k} \|G\|^2 \rho^{2(k-t)} \mathbb{E} \left[ N_{t-1} \|w_{t-1,N_{t-1}}\|^2 I_{\left\{ \sqrt{N_{t-1}} \|w_{t-1,N_{t-1}}\| \geq \epsilon \|G\|^{-1} \rho^{-(k-t)} \right\}} \right]
$$

$$
= \sum_{t=1}^{k} \|G\|^2 \rho^{2(k-t)} \alpha(1) \xrightarrow{k \to \infty} 0.
$$

Thus, the conditions (24)-(26) hold. Then by using Lemma 5, the fact that $e_0$ and $u_0$ have the same limit distribution, and $\alpha^{-1} u_k = -\sum_{t=1}^{k} \xi_{kt}$, we proves Lemma 6. \hfill \Box

### C  Proof of Lemma 9.

Define an auxiliary sequence $\{u_k\}$ by

$$
u_{k+1} = A_k u_k - \alpha (k+1)^{t/2} G w_{k,N_k}, \quad u_0 = 0.
$$

(66)

This combined with (47) produces the following recursion:

$$
e_{k+1} - u_{k+1} = A_k (e_k - u_k) + \zeta_{k+1}
$$

$$
= \Phi_{k,0} (e_0 - u_0) + \sum_{t=0}^{k} \Phi_{k,t+1} G \zeta_{t+1} = \Phi_{k,0} e_0 + \sum_{t=0}^{k} \Phi_{k,t+1} \zeta_{t+1}.
$$

Then by using (46), $\mathbb{E} \left[ \|e_0\|^2 \right] < \infty$, and $\zeta_{k+1} \xrightarrow{P_{k \to \infty}} 0$, we conclude that

$$
\|e_{k+1} - u_{k+1}\| = \|\Phi_{k,0}\| \|e_0\| + \sum_{t=0}^{k} \|\Phi_{k,t+1}\| \|\zeta_{t+1}\| \xrightarrow{P_{k \to \infty}} 0.
$$

37
This implies that $e_k$ defined by (47) and $u_k$ defined as in (66) have the same limit distribution if exists. Thus, we aim to find the stationary distribution of $u_k$. From (66) it follows that

$$u_{k+1} = \Phi_{k,0} u_0 - \alpha \sum_{t=0}^{k} \Phi_{k,t+1} G(t + 1)^{\gamma/2} w_{t,N_t} = -\alpha \sum_{t=0}^{k} \Phi_{k,t+1} G(t + 1)^{\gamma/2} w_{t,N_t}.$$

This implies that

$$\alpha^{-1} u_k = -\alpha \sum_{t=0}^{k-1} \Phi_{k-1,t+1} G(t + 1)^{\gamma/2} w_{t,N_t} = -\sum_{t=1}^{k} \Phi_{k-1,t} G(t + 1)^{\gamma/2} w_{t-1,N_{t-1}}. \quad (67)$$

We intend to apply Lemma 5 by defining $\xi_{k,t} \triangleq -\Phi_{k-1,t} G(t^{\gamma/2} w_{t-1,N_{t-1}}$ and check conditions (24), (25), and (26). Using $N_{t-1} \triangleq \lceil t^\gamma \rceil \geq t^\gamma$, (46), and Assumption 1(iii), we can verify (24). Also, using (28) and (46), the definitions of $S_{k,t}$ and $R_{k,t}$ in (23), the second equality of (25) holds. We now verify the first equality in (25).

$$S_k = \sum_{t=1}^{k} S_{k,t} = \sum_{t=1}^{k} \Phi_{k-1,t} G S_0 G^T \Phi_{k-1,t} + \sum_{t=1}^{k} \Phi_{k-1,t} G S_0 G^T \Phi_{k-1,t} (t^{\gamma/N_t - 1}) \quad (68)$$

where the second and last terms tend to zero by (28), $t^{\gamma/N_t - 1} \xrightarrow{t \to \infty} 1$, and $\sum_{t=1}^{k} \| \Phi_{k-1,t} \|^2 < \infty$ from (46). While for first term on the right-hand side of (68), by using (46) one obtains

$$\sum_{t=1}^{k} \| \Phi_{k-1,t} G S_0 G^T \Phi_{k-1,t} \| \leq \| G S_0 G^T \| \sum_{t=1}^{k} \| \Phi_{k-1,t} \|^2 \leq \| G S_0 G^T \| \left( e^{2\gamma \ln(1/\rho)} - 1 + \frac{1}{\rho^2 \ln(1/\rho)} \right).$$

Because $\sum_{t=1}^{k} \| \Phi_{k-1,t} G S_0 G^T \Phi_{k-1,t} \|$ is monotonically increasing and bounded, its limit exists. Thus, the limit of $\sum_{t=1}^{k} \Phi_{k-1,t} G S_0 G^T \Phi_{k-1,t} = \sum_{t=1}^{k} \left( \frac{k}{t} \right)^\gamma A^{k-t} G S_0 G^T (A^{k-t})^T$ exists and is denoted by $\Sigma$. Then $S_k$ as defined as in (68) satisfies that

$$\lim_{k \to \infty} S_k = \lim_{k \to \infty} \sum_{t=1}^{k} \left( \frac{k}{t} \right)^\gamma A^{k-t} G S_0 G^T (A^{k-t})^T \triangleq \Sigma.$$

Thus, the first equality in (25) holds.

Finally, the Lindeberg condition (26) can similarly validated as that of Lemma 6. Then all conditions of Lemma 5 hold. Thus, by Lemma 5 and (67), we conclude that $\alpha^{-1} u_k \xrightarrow{d} N(0, \Sigma)$. Because $e_k$ defined by (47) and $u_k$ defined as in (66) have the same limit distribution, Lemma 9 is then proved. \(\square\)
D Proof of Proposition 7

Based on Theorem 1 it is seen that

\[ e_{ik} \triangleq \alpha^{-1} \rho_1^{-k/2}(x_{ik} - x^*) - \frac{d}{k \to \infty} Y_i \sim N(0, \Sigma_1), \quad i = 1, \ldots, n, \]  

(69)

where \( Y_1, \ldots, Y_n \) are \( n \) i.i.d. random vectors with distribution \( N(0, \Sigma_1) \). Define

\[ \bar{e} \triangleq \frac{1}{n} \sum_{i=1}^{n} Y_i \text{ and } S \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{e})(Y_i - \bar{e})^T. \]

(70)

Recall from [6, Theorem 4] that \( \bar{e} \) and \( S \) are independently distributed, and

\[ \bar{e} \sim N(0, \Sigma_1/n) \text{ and } (n-1)S \sim W_m(\Sigma_1, n-1), \]

(71)

where \( W_m(\Sigma_1, n-1) \) denotes the \( m \)-dimensional Wishart distribution with \( n-1 \) degrees of freedom and the matrix parameter \( \Sigma_1 \). Because \( \Sigma_1 \) is invertible and \( n \geq m + 1 \), the random matrix \( S \) is almost surely invertible.

(i) Denote by \( e = \text{col}\{e_1, \ldots, e_n\} \triangleq (e_1^T, \ldots, e_n^T)^T \in \mathbb{R}^{mn} \) with \( e_i \in \mathbb{R}^m, \ i = 1, \ldots, n \). Note that \( g_1(e) \triangleq \frac{1}{n} \sum_{i=1}^{n} e_i \) is a continuous function. Since \( e_{ik}, \ i = 1, \ldots, m \) are mutually independent, from (69) it follows that

\[ \text{col}\{e_{1k}, \ldots, e_{nk}\} \xrightarrow{k \to \infty} \text{col}\{Y_1, \ldots, Y_n\}. \]

(72)

Then by the continuous mapping theorem [14, Theorem 1.14] and (71),

\[ \sqrt{n} \alpha^{-1} \rho_1^{-k/2}(\bar{x}_k - x^*) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} e_{ik} \xrightarrow{k \to \infty} \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} Y_i = \sqrt{ne} \sim N(0, \Sigma_1). \]

(73)

(ii) Since \( \text{rank}(\Sigma_1) = m \) and \( \sqrt{ne} \sim N(0, \Sigma_1) \) by (73), we obtain

\[ U_1 \triangleq (\sqrt{ne})^T \Sigma_1^{-1}(\sqrt{ne}) \sim \chi^2(m), \]

(74)

where \( \chi^2(m) \) denotes the chi-squared distribution with \( m \) degrees of freedom. Because \( \text{rank}(\Sigma_1) = m, \bar{e} \in \mathbb{R}^m \) and \( S \) are independently distributed with \( \bar{e}^T \Sigma_1 \) being non-zero with probability one, from [6, Corollary 1] it follows that

\[ U_2 \triangleq \frac{\bar{e}^T \Sigma_1^{-1} \bar{e}}{\bar{e}^T ((n-1)S)^{-1} \bar{e}} \sim \chi^2(n-m) \]

(75)

is independent of \( e \). Hence \( U_1 \) and \( U_2 \) are independent.

From (69) and (51) it follows that

\[ \bar{x}_k - x^* = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - x^*) = \alpha \rho_1^{k/2} \sum_{i=1}^{n} e_{ik}, \]

\[ S_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)^T = \frac{\alpha^2 \rho_1^k}{n-1} \sum_{i=1}^{n} (e_{ik} - \frac{1}{n} \sum_{i=1}^{n} e_{ik}) (e_{ik} - \frac{1}{n} \sum_{i=1}^{n} e_{ik}). \]

(76)
Note that \( g_2(e) \triangleq \frac{1}{n-1} \sum_{i=1}^{n} \left( e_i - \frac{1}{n} \sum_{i=1}^{n} e_i \right) \left( e_i - \frac{1}{n} \sum_{i=1}^{n} e_i \right)^T \) is a continuous function. Because the matrix inverse functional is continuous in a neighborhood of any non-singular matrix, and \( \sum_{i=1}^{n} \left( Y_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \right) \left( Y_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \right)^T \) is almost surely invertible from (71), we conclude that \( (g_2(e))^{-1} \) is almost surely continuous in a neighborhood of \( \text{col}\{Y_1, \ldots, Y_n\} \). Hence, \( g(e) = g_1(e)^T (g_2(e))^{-1} g_1(e) \) is almost surely continuous in a neighborhood of \( \text{col}\{Y_1, \ldots, Y_n\} \). Therefore, by the continuous mapping theorem [14, Theorem 1.14], and (72), we have
\[
\begin{align*}
&n(\bar{x}_k - x^*)^T S_k^{-1}(\bar{x}_k - x^*) \\
&\quad \stackrel{(76)}{=} n \left( \frac{\sum_{i=1}^{n} e_{ik}}{n} \right)^T \left( \frac{1}{n-1} \sum_{i=1}^{n} \left( e_{ik} - \frac{1}{n} \sum_{i=1}^{n} e_{ik} \right) \left( e_{ik} - \frac{1}{n} \sum_{i=1}^{n} e_{ik} \right)^T \right)^{-1} \sum_{i=1}^{n} e_{ik} \\
&\quad \rightarrow \frac{d}{k \to \infty} n \sum_{i=1}^{n} Y_i \left( \frac{1}{n-1} \sum_{i=1}^{n} \left( Y_i - \frac{\sum_{i=1}^{n} Y_i}{n} \right) \left( Y_i - \frac{\sum_{i=1}^{n} Y_i}{n} \right)^T \right)^{-1} \sum_{i=1}^{n} Y_i \\
&\quad \stackrel{(70)}{=} n(n-1)e^T ((n-1)S)^{-1} e = (n-1) \frac{ne^T \Sigma_1^{-1} e}{e^T \Sigma_1^{-1} e} \\
&\quad = (n-1) \frac{U_1}{U_2} \sim \frac{m(n-1)}{n-m} F(m, n-m),
\end{align*}
\]
where the last one holds because \( U_1 = ne^T \Sigma_1^{-1} e \sim \chi^2(m) \) by (74), \( U_2 = \frac{e^T \Sigma_1^{-1} e}{e^T ((n-1)S)^{-1} e} \sim \chi^2(n-m) \) by (75), \( U_1 \) and \( U_2 \) are independent. Thus, \( F(d_1, d_2) \) arises as the ratio of two appropriately scaled chi-squared variates [16]. \( \square \)

### E Proof of Proposition 8

Since \( \Sigma_1 \) is symmetric and invertible, by the property of Wishart distribution, we have that
\[
\Sigma_1^{-1} (n-1)S \Sigma_1^{-1} \sim \mathcal{W}(n, m + (n-m) - 1).
\]
Note by (73) that \( \sqrt{n(n-m)} \Sigma_1^{-1/2} \bar{e} \sim \mathcal{N}(0, (n-m)I_m) \). Recall from [6, Theorem 4] that \( \bar{e} \) and \( S \) are independent. Then by [35, Representation B], we conclude that
\[
\left((n-1) \Sigma_1^{-1} S \Sigma_1^{-1}\right)^{-1/2} \sqrt{n(n-m)} \Sigma_1^{-1/2} \bar{e} \sim T_{n-m}(0, \Sigma_1, m).
\]
Thus, by the probability density function of \( T_{n}((\mu, \Lambda, m) \) defined in (54), we see that
\[
\Sigma_1^{-1/2} \left((n-1) \Sigma_1^{-1} S \Sigma_1^{-1}\right)^{-1/2} \sqrt{n(n-m)} \Sigma_1^{-1/2} \bar{e} \sim T_{n-m}(0, I_m, m).
\]
Hence
\[
((n-1)S)^{-1/2} \sqrt{n(n-m)} \bar{e} \sim T_{n-m}(0, I_m, m).
\]

\[1\] If the random variables \( U_1 \sim \chi^2(d_1) \) and \( U_2 \sim \chi^2(d_2) \) are independent, then \( \frac{U_1/d_1}{U_2/d_2} \sim F(d_1, d_2) \).
Similarly to the derivation of (77), we can show that \( (g_2(e))^{-1/2} g_1(e) \) is almost surely continuous in a neighborhood of \( \text{col}\{Y_1, \cdots, Y_n\} \). Then by the continuous mapping theorem [14, Theorem 1.14] and (72), we achieve (53) by the following.

\[
((n - 1)S_k)^{-1/2} \sqrt{n(n - m)}(x_k - x^*)
\]

\[
\Rightarrow \left. (n - 1) \frac{1}{n - 1} \sum_{i=1}^{n} (e_{ik} - \frac{1}{n} \sum_{i=1}^{n} e_{ik}) (e_{ik} - \frac{1}{n} \sum_{i=1}^{n} e_{ik})^T \right]^{1/2} \sqrt{n(n - m)} \sum_{i=1}^{n} e_{ik} \left. \frac{\text{d}}{k \to \infty} \right. \left( n - 1 \right) \frac{1}{n - 1} \sum_{i=1}^{n} \left( Y_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \right) \left( Y_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \right)^T \right]^{1/2} \sqrt{n(n - m)} \sum_{i=1}^{n} Y_i \approx ((n - 1)S)^{-1/2} \sqrt{n(n - m)} e \sim T_{n - m}(0, I_m, m). \]

SUPPLEMENTARY MATERIAL.

The proof of Lemma 2 is motivated by [41, Section 2.2] and [9, Section 3.6.2].

\section{Proof of Lemma 2}

We define \( \phi_k(x) \) and \( p_k \) as follows.

\[
\phi_0(x) = f(x_0) + \frac{\eta}{2} \|x - x_0\|^2, \tag{F.1}
\]

\[
\phi_{k+1}(x) = (1 - \gamma) \phi_k(x) + \gamma \left( f(x_k) + (x - x_k)^T h(x_k) + \frac{\eta}{2} \|x - x_k\|^2 \right), \tag{F.2}
\]

\[
p_{k+1} = (1 - \gamma) p_k + \left( \alpha + \frac{(1 - \gamma) \gamma}{2 \eta} \right) \|w_k, N_k\|^2 + \alpha w_{k,N_k}^T \nabla f(x_k), \quad p_0 = 0, \tag{F.3}
\]

where \( h(x_k) = \frac{x_k - y_{k+1}}{\alpha} = \nabla f(x_k) + w_{k,N_k} \).

We first show by induction that for any \( k \geq 0 \), \( \nabla^2 \phi_k(x) = \eta I_m \). By (F.1) it is seen that \( \nabla^2 \phi_0(x) = \eta I_m \). Suppose \( \nabla^2 \phi_k(x) = \eta I_m \), then by (F.2), we obtain that

\[
\nabla^2 \phi_{k+1}(x) = (1 - \gamma) \nabla^2 \phi_k(x) + \gamma \eta I_m = \eta I_m.
\]

Thus, \( \nabla^2 \phi_k(x) = \eta I_m \) for any \( k \geq 0 \). Because \( \phi_k(x) \) is a quadratic function, \( \phi_k(x) \) can be written as:

\[
\phi_k(x) = \phi_k^* + \frac{\eta}{2} \|x - v_k\|^2 \text{ with } v_k = \arg\min_x \phi_k(x), \quad \forall k \geq 0. \tag{F.4}
\]

We proceed to give a recursive form for \( v_{k+1} \) and \( \phi_{k+1}^* \). Noting from (F.4) that \( \nabla \phi_k(x) = \eta(x - v_k) \).

Then by using the first-order optimality condition \( \nabla \phi_{k+1}(x) = 0 \) of the unconstrained convex optimization \( \min_x \phi_{k+1}(x) \), and the definition of \( \phi_{k+1}(x) \) in (F.2), we obtain that

\[
\nabla \phi_{k+1}(x) = (1 - \gamma) \eta(x - v_k) + \gamma h(x_k) + \gamma \eta(x - x_k) = 0.
\]
This implies that
\[ v_{k+1} = (1 - \gamma)v_k + \gamma x_k - \gamma h(x_k)/\eta. \] (F.5)

By using (F.2) and (F.4), evaluating \( \phi_{k+1}(x) \) at \( x = x_k \) we obtain that
\[
\phi^*_{k+1} = \phi_{k+1}(x_k) - \frac{\eta}{2} \|x_k - v_{k+1}\|^2 = (1 - \gamma)\phi_k(x_k) + \gamma f(x_k) - \frac{\eta}{2} \|x_k - v_{k+1}\|^2
\]
\[
= (1 - \gamma)\phi^*_k + \frac{\eta(1 - \gamma)}{2} \|x_k - v_k\|^2 + \gamma f(x_k) - \frac{\eta}{2} \|v_{k+1} - x_k\|^2. \] (F.6)

Note by (F.5) that
\[
\|v_{k+1} - x_k\|^2 = \|(1 - \gamma)(v_k - x_k) - \gamma h(x_k)/\eta\|^2
\]
\[
= (1 - \gamma)^2 \|v_k - x_k\|^2 + \frac{\gamma^2}{\eta^2} \|h(x_k)\|^2 - \frac{2\gamma(1 - \gamma)}{\eta} (v_k - x_k)^T h(x_k). \]

This together with (F.6) leads to
\[
\phi^*_{k+1} = (1 - \gamma)\phi^*_k + \gamma f(x_k) + \frac{\eta\gamma(1 - \gamma)}{2} \|x_k - v_k\|^2 - \frac{\gamma^2}{\eta} \|h(x_k)\|^2 + \gamma(1 - \gamma)(v_k - x_k)^T h(x_k). \] (F.7)

We then show by induction that the following holds for any \( k \geq 0 \).
\[
v_k - x_k = \frac{1}{\gamma}(x_k - y_k). \] (F.8)

From (F.1) it is seen that the optimal solution of \( \phi_0(x) \) is \( v_0 = x_0 \). Then by the initial condition \( x_0 = y_0 \), we see that (F.8) holds for \( k = 0 \). We inductively assume that (F.8) holds for \( k \), and proceed to prove that (F.8) holds for \( k + 1 \). By substituting \( v_k = x_k + (x_k - y_k)/\gamma \) into (F.5), one obtains
\[
v_{k+1} - x_{k+1} = (1 - \gamma)(x_k + (x_k - y_k)/\gamma) + \gamma x_k - \gamma h(x_k)/\eta - x_{k+1}
\]
\[
= \frac{1}{\gamma} (x_k - \gamma^2 h(x_k)/\eta) - \left( \frac{1}{\gamma} - 1 \right) y_k - x_{k+1}.
\]

This together with \( h(x_k) = \frac{x_k - y_{k+1}}{\alpha} \) and \( \gamma = \sqrt{\alpha\eta} \) produces
\[
v_{k+1} - x_{k+1} = \frac{1}{\gamma} (x_k - \alpha h(x_k)) - \left( \frac{1}{\gamma} - 1 \right) y_k - x_{k+1}
\]
\[
= \frac{1}{\gamma} y_{k+1} - \left( \frac{1}{\gamma} - 1 \right) \frac{(1 + \beta)y_{k+1} - x_{k+1}}{\beta} - x_{k+1} = \frac{1}{\gamma} (x_{k+1} - y_{k+1}),
\]
where the last equality holds by \( \beta = \frac{1 - \gamma}{1 + \gamma} \). Thus, we have shown that (F.8) holds for any \( k \geq 0 \).

Then by substituting (F.8) into (F.7), and using \( h(x_k) = \nabla f(x_k) + w_kN_k \), we obtain that
\[
\phi^*_{k+1} = (1 - \gamma)\phi^*_k + \gamma f(x_k) + \frac{\eta(1 - \gamma)}{2\gamma} \|x_k - y_k\|^2 - \frac{\gamma^2}{2\eta} \|h(x_k)\|^2 + (1 - \gamma)(x_k - y_k)^T h(x_k)
\]

42
\[
= (1 - \gamma) \phi_k^* + \gamma f(x_k) - \frac{\gamma^2}{2\eta} \|h(x_k)\|^2 + (1 - \gamma)(x_k - y_k)^T \nabla f(x_k) + \frac{\eta(1 - \gamma)}{2\gamma} \|x_k - y_k\|^2 + (1 - \gamma)(x_k - y_k)^T w_{k,N_k} \\
\geq (1 - \gamma) \phi_k^* + \gamma f(x_k) - \frac{\gamma^2}{2\eta} \|h(x_k)\|^2 + (1 - \gamma)(x_k - y_k)^T \nabla f(x_k) - \frac{(1 - \gamma)\gamma}{2\eta} \|w_{k,N_k}\|^2,
\]

where the last inequality follows by \( \|a\|^2 + 2a^T b \geq -\|b\|^2 \).

We proceed to show that \( \phi_k^* \geq f(y_k) - p_k \) for any \( k \geq 0 \). By the definitions (F.1) and (F.3), we see that \( \phi_0^* = f(x_0) = f(y_0) \) and \( p_0 = 0 \). Hence \( \phi_k^* \geq f(y_k) - p_k \) holds for \( k = 0 \). We inductively assume that \( f(y_k) \leq \phi_k^* + p_k \), and aim to show that \( f(y_{k+1}) \leq \phi_{k+1}^* \leq p_{k+1} \). Since \( f(x) \) is \( L \)-smooth, by using \( h(x_k) = \frac{x_k - y_{k+1}}{\alpha} = \nabla f(x_k) + w_{k,N_k} \), we obtain that

\[
f(y_{k+1}) \leq f(x_k) + (y_{k+1} - x_k)^T \nabla f(x_k) + \frac{L}{2} \|y_{k+1} - x_k\|^2 \\
\leq f(x_k) - \alpha h(x_k)^T (h(x_k) - w_{k,N_k}) + \frac{L\alpha^2}{2} \|h(x_k)\|^2 \\
\leq f(x_k) + \left( \frac{L\alpha^2}{2} - \alpha \right) \|h(x_k)\|^2 + \alpha \|w_{k,N_k}\|^2 + \alpha w_{k,N_k}^T \nabla f(x_k),
\]

where the last inequality holds because \( h(x_k)^T w_{k,N_k} = \|w_{k,N_k}\|^2 + w_{k,N_k}^T \nabla f(x_k) \). By the induction assumption \( f(y_k) \leq \phi_k^* + p_k \) and the convexity of \( f(\cdot) \), we obtain that

\[
f(x_k) = (1 - \gamma)f(y_k) + (1 - \gamma)(f(x_k) - f(y_k)) + \gamma f(x_k) \\
\leq (1 - \gamma)\phi_k^* + (1 - \gamma)p_k + (1 - \gamma)(x_k - y_k)^T \nabla f(x_k) + \gamma f(x_k).
\]

The above bound combined with (F.10) produces

\[
f(y_{k+1}) \leq (1 - \gamma)\phi_k^* + \gamma f(x_k) + (1 - \gamma)(x_k - y_k)^T \nabla f(x_k) + (1 - \gamma)p_k \\
+ \left( \frac{L\alpha^2}{2} - \alpha \right) \|h(x_k)\|^2 + \alpha \|w_{k,N_k}\|^2 + \alpha w_{k,N_k}^T \nabla f(x_k).
\]

It incorporated with (F.9) leads to the following relation:

\[
f(y_{k+1}) - \phi_{k+1}^* \leq \left( \frac{L\alpha^2}{2} - \alpha + \frac{\gamma^2}{2\eta} \right) \|h(x_k)\|^2 + (1 - \gamma)p_k + \alpha w_{k,N_k}^T \nabla f(x_k) \\
+ \left( \alpha + \frac{(1 - \gamma)\gamma}{2\eta} \right) \|w_{k,N_k}\|^2 \leq p_{k+1},
\]

where the last inequality holds by the definition of \( p_{k+1} \) in (F.3) and \( \frac{L\alpha^2}{2} - \alpha + \frac{\gamma^2}{2\eta} = \frac{\alpha(\alpha L - 1)}{2} \leq 0 \) from \( \alpha \leq 1/L \) and \( \gamma^2 = \alpha \eta < 1 \). Therefore, we conclude that \( f(y_k) \leq \phi_k^* + p_k \) for any \( k \geq 0 \).

Because \( f(x) \) is \( \eta \)-strongly convex, from (F.2) and \( \mathbb{E}[h(x_k)|F_k] = \nabla f(x_k) \) it follows that

\[
\mathbb{E}[\phi_{k+1}|F_k] = (1 - \gamma)\phi_k(x) + \gamma f(x_k) + (x_k - x_k)^T \nabla f(x_k) + \frac{\eta}{2} \|x_k - x_k\|^2 \\
\leq (1 - \gamma)\phi_k(x) + \gamma f(x), \quad \forall x \in \mathbb{R}^m.
\]
By taking unconditional expectations, we obtain that $\mathbb{E}[\phi_{k+1}(x)] \leq (1 - \gamma)\mathbb{E}[\phi_k(x)] + \gamma f(x)$ for any $x \in \mathbb{R}^m$. Therefore, by rearranging terms and setting $x = x^*$ in the above inequality, we have

$$\mathbb{E}[\phi_{k+1}(x^*)] - f(x^*) \leq (1 - \gamma)(\mathbb{E}[\phi_k(x^*)] - f(x^*)) \leq (1 - \gamma)^{k+1}(\mathbb{E}[\phi_0(x^*)] - f(x^*))$$

Then by the fact that $f(y_k) \leq \phi_k^* + p_k$, there holds

$$\mathbb{E}[f(y_k)] - f(x^*) \leq \mathbb{E}[\phi_k(x^*)] - f(x^*) + \mathbb{E}[p_k] \leq (1 - \gamma)^k(\mathbb{E}[\phi_0(x^*)] - f(x^*)) + \mathbb{E}[p_k] \leq \frac{\eta + L}{2}(1 - \gamma)^k\mathbb{E}[\|x_0 - x^*\|^2] + \mathbb{E}[p_k],$$

which is (21). By taking expectations on both sides of (F.3), using $p_0 = 0$ and (5), we obtain that

$$\mathbb{E}[p_k] = (1 - \gamma)\mathbb{E}[p_{k-1}] + \left(\alpha + \frac{(1 - \gamma)\gamma}{2\eta}\right)\mathbb{E}[\|w_{k-1,N_{k-1}}\|^2]$$

$$= \left(\alpha + \frac{(1 - \gamma)\gamma}{2\eta}\right)\sum_{i=0}^{k-1}(1 - \gamma)^i\mathbb{E}[\|w_{k-i-1,N_{k-i-1}}\|^2] \leq \nu^2\left(\alpha + \frac{(1 - \gamma)\gamma}{2\eta}\right)\sum_{i=0}^{k-1}(1 - \gamma)^i/N_{k-1-i}.$$ 

This together with (F.12) proves Lemma 2.

\[\square\]

### G Proof of Proposition 2

By using $\rho_2 \in (1 - \gamma, 1)$ and substituting $N_k = \lceil \rho_2^{-k+1} \rceil$ into (11), we obtain that

$$\mathbb{E}[f(y_k)] - f^* \leq \rho_2^k \frac{\eta + L}{2} \mathbb{E}[\|x_0 - x^*\|^2] + \nu^2 \left(\alpha + \frac{(1 - \gamma)\gamma}{2\eta}\right) \rho_2^k \sum_{i=0}^{k-1} \left(\frac{1 - \gamma}{\rho_2}\right)^i,$$

which leads to (21) by using the bound $\sum_{i=0}^{k-1} \left(\frac{1 - \gamma}{\rho_2}\right)^i \leq \frac{1}{1 - \frac{\gamma}{\rho_2}}$.

Because $f(x)$ is $\eta$-strongly convex and $\nabla f(x^*) = 0$, we have $f(x) - f(x^*) \geq \frac{\eta}{2}||x - x^*||^2$. Thus, $y_k$ generated by Algorithm 2 satisfies $||y_k - x^*||^2 \leq \frac{2}{\eta}(f(y_k) - f^*)$. Then from (21) it follows that

$$\mathbb{E}[\|y_k - x^*\|^2] \leq c\rho_2^k$$

with $c \equiv \frac{2\rho_2^\nu^2}{\eta} \left(\frac{\eta + L}{2} \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\rho_2^\nu^2}{\rho_2 - (1 - \gamma)} \left(\alpha + \frac{(1 - \gamma)\gamma}{2\eta}\right)\right)$.

Note by (10b) that $\|x_k - x^*\| \leq \|(1 + \beta)(y_k - x^*)\| + ||\beta(y_{k-1} - x^*)||$, and hence

$$\mathbb{E}[\|x_k - x^*\|^2] \leq 2(1 + \beta)^2\mathbb{E}[\|y_k - x^*\|^2] + 2\beta^2\mathbb{E}[\|y_{k-1} - x^*\|^2]$$

$$\leq c \left(2(1 + \beta)^2 + 2\beta^2\rho_2^{-1}\right) \rho_2^k.$$ 

Therefore, sequences $\{x_k\}$ and $\{y_k\}$ converge to the optimal solution $x^*$ at a geometric rate $O(\rho_2^k)$ in the mean-squared sense.

Suppose we set $\alpha \equiv \frac{1}{L}$. Then $\gamma = \sqrt{\frac{\eta}{L}} = \frac{1}{\sqrt{\kappa}}$ and $\beta = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$. Select $\rho_2 \equiv 1 - \frac{1}{2\sqrt{\kappa}}$ such that $\rho_2 > 1 - \gamma$. We can show that the number of iterations required to obtain an $\epsilon$-optimal solution in a mean-squared sense is $O\left(\frac{\ln(1/\epsilon)}{\ln(1/\rho_2)}\right) = O\left(\sqrt{\kappa} \ln(1/\epsilon)\right)$ since $\ln\left(\frac{1}{1 - 1/(2\sqrt{\kappa})}\right) \approx \frac{1}{2\sqrt{\kappa}}$ for large $\kappa$, and hence the oracle complexity $\sum_{k=0}^{K(\epsilon)-1} N_k = O\left(\sqrt{\kappa}/\epsilon\right)$.

\[\square\]
H Proof of Proposition 3

By substituting \( \alpha = \frac{4}{(\sqrt{\eta}+\epsilon)L^2} \) into \( \beta = \max\{1-\sqrt{\alpha\eta}, 1-\sqrt{\alpha L}\}^2 \), there holds \( \beta = \left(1 - \frac{2}{\sqrt{\kappa+1}}\right)^2 < 1. \) Then Lemma 3 holds. Therefore, by using (13), \( N_k = \lceil \rho_3^{-(k+1)} \rceil, x_{-1} = x_0, \) and \( \rho_3 \in (\beta, 1), \) we obtain that

\[
E \left[ \left\| \left( x_{k+1} - x^* \right) \right\| \left( x_k - x^* \right) \right] \leq \beta E \left[ \left\| \left( x_k - x^* \right) \right\| \left( x_{k-1} - x^* \right) \right] + \alpha^2 \nu^2 \rho_3^{k+1}
\]

\[
\leq 2 \beta^{k+1} E \left[ \| x_0 - x^* \| \right] + \alpha^2 \nu^2 \sum_{t=0}^{k} \beta^t \rho_3^{k+1-t}, \quad \forall k \geq 0.
\]

This together with \( \sum_{t=0}^{k} \beta^t \rho_3^{k+1-t} = \rho_3^{k+1} \sum_{t=0}^{k} (\beta/\rho_3)^t \leq \rho_3^{k+1} \) proves (22).

By (22), \( E \left[ \| x_k - x^* \| \right] \leq c \rho_3^{k} \) for some constant \( c > 0. \) Suppose \( \rho_3 = \left(1 - \frac{4}{\sqrt{\kappa+1}}\right)^2 > \beta. \) Akin to the proof of Proposition 1(ii), we can show that the number of iterations required to obtain an \( \epsilon \)-optimal solution satisfying \( E \left[ \| x_k - x^* \| \right] = O \left( \ln(1/\epsilon) \right) = O \left( \sqrt{\kappa} \ln(1/\epsilon) \right) \) for large \( \kappa, \) and the oracle complexity \( \sum_{k=0}^{K(\epsilon)-1} N_k = O \left( \sqrt{\kappa}/\epsilon \right). \)

I Proof of Proposition 4

By substituting \( N_k = \lceil (k+1)^v \rceil \) into (7), using \( q = 1 - \frac{2\alpha \eta L}{\eta + L} \) and Lemma 7(i), one obtains

\[
E \left[ \| x_k - x^* \| \right] \leq qE \left[ \| x_{k-1} - x^* \| \right] + \alpha^2 \nu^2 k^{-v} = q^k E \left[ \| x_0 - x^* \| \right] + \alpha^2 \nu^2 \sum_{m=1}^{k} q^{k-m} m^{-v}
\]

\[
= q^k E \left[ \| x_0 - x^* \| \right] + \alpha^2 \nu^2 \left( q^k \frac{\epsilon^{2v+1} - 1}{1-q} + \frac{k^{-v}}{\ln(1/q)} \right),
\]

which together with Lemma 7(ii) proves (45). Then \( E \left[ \| x_k - x^* \| \right] \leq \epsilon \) for any \( k \geq K(\epsilon) = \lceil \frac{C_v}{\epsilon} \rceil^{1/v}. \) By noting that \( C_v = O(\epsilon^v v^2) \), the iteration complexity is \( O(v(1/\epsilon)^{1/v}). \) Therefore, the number of sampled gradients required to obtain an \( \epsilon-\)NE is bounded by

\[
\sum_{k=0}^{K(\epsilon)-1} \lceil (k+1)^v \rceil \leq K(\epsilon) + (K(\epsilon))^v + \sum_{k=1}^{K(\epsilon)-1} K(\epsilon)^v \leq (\frac{C_v}{\epsilon})^{1/v} + \frac{C_v}{\epsilon} + \int_1^{K(\epsilon)} t^v dt
\]

\[
= (\frac{C_v}{\epsilon})^{1/v} + \frac{C_v}{\epsilon} + (v + 1)^{-1} \left( \frac{C_v}{\epsilon} \right)^{1+v} - (v + 1)^{-1}.
\]

Therefore, the oracle complexity is \( O \left( \epsilon^v v^2 (1/\epsilon)^{1+1/v} \right). \)

J Proof of Proposition 5

It is noticed by \( t = k - i \) and Lemma 7(i),

\[
\sum_{i=0}^{k-1} (1 - \gamma)^i (k-i)^{-v} = \sum_{t=1}^{k} (1 - \gamma)^{k-t} t^{-v} \leq \frac{(1-\gamma)^k (\epsilon^{2v(1-\gamma)-1})}{\gamma} + \frac{2k^{-v}}{(1-\gamma) \ln(1/(1-\gamma))}. \quad (J.1)
\]
By substituting $N_k = \lceil (k + 1)^v \rceil$ into (11), we obtain that

$$
\mathbb{E}[f(y_k)] - f^* \leq \frac{(1-\gamma)^h(v+L)}{2} \mathbb{E}[\|x_0 - x^*\|^2] + \nu^2 \left( \alpha + \frac{(1-\gamma)\gamma}{2\eta} \right) \sum_{i=0}^{k-1} (1 - \gamma)^i (k - i)^{-v}.
$$

This together with Lemma 7(ii) and (J.1) proves the required result. \hfill \Box

### K Proof of Proposition 6

By substituting $N_k = \lceil (k + 1)^v \rceil$ into (13) and using Lemma 7(i), we obtain that

$$
\mathbb{E} \left[ \left\| \frac{x_{k+1} - x^*}{x_k - x^*} \right\|^2 \right] \leq \beta \mathbb{E} \left[ \left\| \frac{x_k - x^*}{x_{k-1} - x^*} \right\|^2 \right] + \alpha^2 \nu^2 (k+1)^{-v}
$$

$$
\leq 2\beta^{k+1} \mathbb{E}[\|x_0 - x^*\|^2] + \alpha^2 \nu^2 \sum_{m=1}^{k+1} \beta^{k+1-m} m^{-v}.
$$

$$
\leq 2\beta^{k+1} \mathbb{E}[\|x_0 - x^*\|^2] + \alpha \nu \left( \beta^{k+1} \frac{e^2 \nu \beta^{-1} - 1}{1 - \beta} + \frac{2(k+1)^{-v}}{\beta \ln(1/\beta)} \right),
$$

which together with Lemma 7(ii) proves the lemma. \hfill \Box