ON SAMPLING RATES IN SIMULATION-BASED RECURSIONS

RAGHU PASUPATHY*, PETER GLYNN†, SOUMYADIP GHOSH‡, AND FATEMEH S. HASHEMI§

Abstract. We consider the context of "simulation-based recursions," that is, recursions that involve quantities needing to be estimated using a stochastic simulation. Examples include stochastic adaptations of fixed-point and gradient descent recursions obtained by replacing function and derivative values appearing within the recursion by their Monte Carlo counterparts. The primary motivating settings are Simulation Optimization and Stochastic Root Finding problems, where the low-point and the zero of a function are sought, respectively, with only Monte Carlo estimates of the functions appearing within the problem. We ask how much Monte Carlo sampling needs to be performed within simulation-based recursions in order that the resulting iterates remain consistent, and more importantly, *efficient*, where "efficient" implies convergence at the fastest possible rate. Answering these questions involves trading-off two types of error inherent in the iterates: the deterministic error due to recursion and the "stochastic" error due to sampling. As we demonstrate through a characterization of the relationship between sample sizing and convergence rates, efficiency and consistency are intimately coupled with the speed of the underlying recursion, with faster recursions yielding a wider regime of "optimal" sampling rates. The implications of our results to practical implementation are immediate since they provide specific guidance on optimal simulation expenditure within a variety of stochastic recursions.

1. INTRODUCTION. We consider the question of sampling within algorithmic recursions that involve quantities needing to be estimated using a stochastic simulation. The prototypical example setting is Simulation Optimization (SO) [16, 25], where an optimization problem is to be solved using only a stochastic simulation capable of providing estimates of the objective function and constraints at a requested point. Another closely related example setting is the Stochastic Root Finding Problem (SRFP) [24, 27, 26], where the zero of a vector function is sought, with only simulation-based estimates of the function involved. SO problems and SRFPs, instead of stipulating that the functions involved in the problem statement be known exactly or in analytic form, allow implicit representation of functions through a stochastic simulation, thereby facilitating virtually any level of complexity. Such flexibility has resulted in adoption across widespread application contexts. A few examples are logistics [17, 18, 3], healthcare [1, 13, 11], epidemiology [14], and vehicular-traffic systems [23].

A popular and reasonable solution paradigm for solving SO problems and SRFPs is to simply mimic what a solution algorithm might do within a deterministic context, after estimating any needed function and derivative values using the available stochastic simulation. An example serves to illustrate such a technique best. Consider the basic quasi-Newton recursion

$$x_{k+1} = x_k - \alpha_k \tilde{H}_f^{-1}(x_k)\tilde{\nabla} f(x_k), \tag{1.1}$$

used to find a local minimum of a twice-differentiable real-valued function $f : \mathbb{R}^d \to \mathbb{R}$, where $\tilde{H}_f(x)$ and $\tilde{\nabla} f(x)$ are the Hessian and gradient (deterministic) approximations

*Department of Statistics, Purdue University, West Lafayette, IN 47907, USA (pasupath@purdue.edu).

†Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA (glynn@stanford.edu).

‡T.J. Watson IBM Research, Yorktown Heights, NY 10598, USA (ghosh@us.ibm.com).

§The Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA (fatemeh.s.hashemi@gmail.com).

of the true Hessian $H_f(x)$ and gradient $\nabla f(x)$ of the function $f : \mathbb{R}^d \to \mathbb{R}$ at the point x. (We emphasize that $\tilde{H}_f(x)$ and $\tilde{\nabla} f(x)$ as they appear in (1.1) are "deterministic," and could be, for example, approximations obtained through appropriate finite-differencing of the function f at a set of points around x.) Suppose that the context in consideration is such that only "noisy" simulation-based estimates of f are available, implying that the recursion in (1.1) is not implementable as written. A reasonable adaptation of (1.1) might instead be the recursion

$$X_{k+1} = X_k - \hat{\alpha}_k \hat{H}_f^{-1}(m_k, X_k) \hat{\nabla} f(m_k, X_k), \tag{1.2}$$

where $\hat{\nabla} f(m_k, x), x \in \mathbb{R}^d$ and $\hat{H}_f(m_k, x), x \in \mathbb{R}^d$ are simulation estimators of $\tilde{\nabla} f(x), x \in \mathbb{R}^d$ and $\tilde{H}_f(x), x \in \mathbb{R}^d$ constructed using *estimated* function values, and the "step-length" $\hat{\alpha}_k$ estimates the step-length α_k appearing in the deterministic recursion (1.1). The simulation effort m_k in (1.2) is general and might represent the number of simulation replications in the case of terminating simulations or the simulation run length in the case of non-terminating simulations [20].

While the recursion in (1.2) is intuitively appealing, important questions arise within its context. Since the exact function value $f(x)$ at any point x is unknown and needs to be estimated using stochastic sampling, one might ask how much sampling m_k should be performed during each iteration k. Inadequate sampling can cause non-convergence of (1.2) due to repeated mis-steps from which iterates in (1.2) might fail to recover. Such non-convergence can be avoided through increased sampling, that is, using large m_k values; however, such increased sampling translates to an increase in computational complexity and an associated decreased convergence rate.

The questions we answer in this paper pertain to the (simulation) sampling effort expended within recursions such as (1.2). Our interest is a generalized version of (1.2) that we call Sampling Controlled Stochastic Recursion (SCSR) that will be defined more rigorously in Section 3. Within the context of SCSR, we ask the following questions.

- $Q.1$ What sampling rates in SCSR ensure that the resulting iterates are *strongly consistent*, that is, converge to the correct solution with probability one?
- $Q.2$ What is the convergence rate of the iterates resulting from SCSR, expressed as a function of the sample sizes and the speed of the underlying deterministic recursion?
- $Q.3$ With reference to $Q.2$, are there specific SCSR recursions that guarantee a *canonical rate*, that is, the fastest achievable convergence speed under generic sampling?
- $Q.4$ What do the answers to $Q.1$–$Q.3$ imply for practical implementation?

Questions such as what we ask in this paper have recently been considered [15, 10, 28] but usually within a specific "algorithmic" context. (An exception is [7] which broadly treats the complexity trade-offs stemming from estimation, approximation, and optimization errors within large-scale learning problems.) In [15], for instance, the behavior of the stochastic gradient descent recursion

$$x_{k+1} = x_k - \alpha_k g_k \tag{1.3}$$

is considered for optimizing a smooth function f, where α_k is the step size used during the kth iteration and g_k is an estimate of the gradient $\nabla f(x_k)$. Importantly, g_k is assumed to be estimated such that the error in the estimate $e_k = g_k - \nabla f(x_k)$ satisfies $\mathbb{E}\left[\|e_k\|^2\right] \leq B_k$, where B_k is a "per-iteration" bound that can be seen to be related to the notion of sample size in this paper. The results in [15] detail the functional relationship between the convergence rate of the sequence $\{x_k\}$ in (1.3) and the chosen sequence $\{B_k\}$. Like in [15], the recursion considered in [10] is again (1.3) but [10] considers the question more directly, proposing a dynamic sampling scheme akin to that in [28] that is a result of balancing the variance and the squared bias of the gradient estimate at each step. One of the main results in [10] states that when sample sizes grow geometrically across iterations, the resulting iterates in (1.3) exhibit the fastest achievable convergence rate, something that will be reaffirmed for SCSR recursions considered in this paper.

As already noted, we consider the questions $Q.1 - Q.4$ within a recursive context (SCSR) that is more general than (1.3) or (1.2). Our aim is to characterize the relationship between the errors due to recursion and sampling that naturally arise in SCSR, and their implication to SO and SRFP algorithms. We will demonstrate through our answers that these errors are inextricably linked and fully characterizable. Furthermore, we will show that such characterization naturally leads to sampling regimes which, when combined with a deterministic recursion of a specified speed, result in specific SCSR convergence rates. The implication for implementation seems clear: given the choice of the deterministic recursive structure in use, our error characterization suggests sampling rates that should be employed in order to enjoy the best achievable SCSR convergence rates.

1.1. Summary and Insight from Main Results. The results we present are broadly divided into those concerning the strong consistency of SCSR iterates, and those pertaining to SCSR's efficiency as defined from the standpoint of the total amount of simulation effort. Insight on consistency appears in the form of Theorem 5.2, which relates the estimator quality in SCSR with the minimum sampling rate that will guarantee almost sure convergence. Theorem 5.2 is deliberately generic in that it makes only mild assumptions about the speed of the recursion in use within SCSR and about the simulation estimator quality. Theorem 5.2 also guarantees convergence (to zero) of the mean absolute deviation (or $\mathcal{L}^1$ convergence) of SCSR's iterates to a solution.

Theorems 6.1–6.9 and associated corollaries are devoted to efficiency issues surrounding SCSR. Of these, Theorems 6.6–6.9 are the most important, and characterize the convergence rate of SCSR as a function of the sampling rate and the speed of recursion in use. Specifically, as summarized in Figure 6.1, these results characterize the sampling regimes resulting in predominantly sampling error ("too little sampling") versus those resulting in predominantly recursion error ("too much sampling"), along with identifying the convergence rates for all recursion-sampling combinations. Furthermore, and as illustrated using the shaded region in Figure 6.1, Theorems 6.6–6.9 identify those recursion-sampling combinations yielding the optimal rate, that is, the highest achievable convergence rates with the given simulation estimator at hand. As it turns out, and as implied by Theorems 6.6–6.9, recursions that utilize more structural information afford a wider range of sampling rates that produce the optimal rate. For instance, Theorems 6.6–6.9 imply that recursions such as (1.2) will achieve the optimal rate if the sampling rate is either geometric, or super-exponential up to a certain threshold; sampling rates falling outside this regime yield sub-canonical convergence rates for SCSR. (The notions of optimal rates, sampling rates, and recursion rates will be defined rigorously in short order.) The corresponding regime when using a linearly-converging recursion such as a fixed-point recursion is narrower, and limited to a small band of geometric sampling rates. Interestingly, our results show that sublinearly converging recursions are incapable of yielding optimal rates for SCSR, that is, the sampling regime that produces optimal rates when a sublinearly converging recursion is in use is empty. We also present a result (Theorem 6.10) that provides a complexity bound on the mean absolute error of the SCSR iterates under more restrictive assumptions on the behavior of the recursion in use.

1.2. Paper Organization. The rest of the paper is organized as follows. In the ensuing section, we introduce much of the standing notation and conventions used throughout the paper. This is followed by Section 3 where we present a rigorous problem statement, and by Section 4 where we present specific non-trivial examples of SCSR recursions. Sections 5 and 6 contain the main results of the paper. We provide concluding remarks in Section 7, with a brief commentary on implementation and the use of stochastic sample sizes.

2. NOTATION AND CONVENTION. We will adopt the following notation through the paper. For more details, especially on the convergence of sequences of random variables, see [5]. (i) If $x \in \mathbb{R}^d$ is a vector, then its components are denoted through $x \triangleq (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$. (ii) We use $e_i \in \mathbb{R}^d$ to denote a unit vector whose ith component is 1 and whose every other component is 0, that is, $e_i(i) = 1$ and $e_i(j) = 0$ for $j \neq i$. (iii) For a sequence of random variables $\{Z_n\}$, we say $Z_n \stackrel{\mathrm{p}}{\to} Z$ if $\{Z_n\}$ converges to Z in probability; we say $Z_n \stackrel{\mathrm{d}}{\to} Z$ to mean that $\{Z_n\}$ converges to Z in distribution; we say that $Z_n \stackrel{\mathcal{L}^p}{\to} Z$ if $\mathbb{E}[|Z_n - Z|^p] \to 0$; and finally $Z_n \stackrel{\mathrm{wp1}}{\to} Z$ to mean that $\{Z_n\}$ converges to Z with probability one. When $Z_n \stackrel{\mathrm{wp1}}{\to} z$, where z is a constant, we will say that Z_n is strongly consistent with respect to z. (iv) $\mathbb{Z}^+$ denotes the set of positive integers. (v) $\mathcal{B}_r(x^*) \triangleq \{x : \|x - x^*\| \leq r\}$ denotes the d-dimensional Euclidean ball centered on x^* and having radius r. (vi) $\mathrm{dist}(x, B) = \inf\{\|x - y\| : y \in B\}$ denotes the Euclidean distance between a point $x \in \mathbb{R}^d$ and a set $B \subset \mathbb{R}^d$. (vii) $\mathrm{diam}(B) = \sup\{\|x-y\| : x, y \in B\}$ denotes the diameter of the set $B \subset \mathbb{R}^d$. (viii) For a sequence of real numbers $\{a_n\}$, we say $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$; and $a_n = \mathcal{O}(1)$ if $\{a_n\}$ is bounded, i.e., $\exists c \in (0, \infty)$ with $|a_n| < c$ for large enough n. We say that $a_n = \Theta(1)$ if $0 < \liminf a_n \leq \limsup a_n < \infty$. For positive-valued sequences $\{a_n\}$, $\{b_n\}$, we say $a_n = \mathcal{O}(b_n)$ if $a_n/b_n = \mathcal{O}(1)$ as $n \to \infty$; we say $a_n = \Theta(b_n)$ if $a_n/b_n = \Theta(1)$ as $n \to \infty$. (ix) For a sequence of positive-valued random variables $\{A_n\}$, we say $A_n = o_p(1)$ if $A_n \stackrel{\mathrm{p}}{\to} 0$ as $n \to \infty$; and $A_n = \mathcal{O}_p(1)$ if $\{A_n\}$ is stochastically bounded, that is, for given $\epsilon > 0$ there exists $c(\epsilon) \in (0, \infty)$ with $\mathbb{P}(A_n < c(\epsilon)) > 1 - \epsilon$ for large enough n. If $\{B_n\}$ is another sequence of positive-valued random variables, we say $A_n = \mathcal{O}_p(B_n)$ if $A_n/B_n = \mathcal{O}_p(1)$ as $n \to \infty$; we say $A_n = o_p(B_n)$ if $A_n/B_n = o_p(1)$ as $n \to \infty$. Also, when we say $A_n \leq \mathcal{O}_p(b_n)$, we mean that $A_n \leq B_n$ where $\{B_n\}$ is a random sequence that satisfies $B_n = \mathcal{O}_p(b_n)$. (x) For two sequences of real numbers $\{a_n\}, \{b_n\}$ we say $a_n \sim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$.

Also, the following notions will help our exposition and will be used heavily.

Definition 2.1. *(Growth rate of a sequence.) A sequence $\{m_k\}$ is said to exhibit Polynomial(λ_p, p) growth if $m_k = \lambda_p\, k^p, k = 1, 2, \ldots$ for some $\lambda_p, p \in (0, \infty)$; it is said to exhibit Geometric(c) growth if $m_{k+1} = c\, m_k, k = 0, 1, 2, \ldots$ for some $c \in (1, \infty)$; and SupExponential(λ_t, t) growth if $m_{k+1} = \lambda_t m_k^t, k = 0, 1, 2, \ldots$ for some $\lambda_t \in (0, \infty), t \in (1, \infty)$.*

Definition 2.2. *(A sequence increasing faster than another.) Let $\{m_k\}$ and $\{\tilde{m}_k\}$ be two positive-valued increasing sequences that tend to infinity. Then $\{m_k\}$ is* *said to increase faster than $\{\tilde{m}_k\}$ if $m_{k+1}/m_k \geq \tilde{m}_{k+1}/\tilde{m}_k$ for large enough k. In such a case, $\{\tilde{m}_k\}$ is also said to increase slower than $\{m_k\}$.*

According to Definition 2.1 and Definition 2.2, it can be seen that any sequence that is growing as SupExponential(λ_t, t) is faster than any other sequence that is growing as Geometric(c); likewise, any sequence growing as Geometric(c) is faster than any other sequence growing as Polynomial(λ_p, p).

3. PROBLEM SETTING AND ASSUMPTIONS. The general context that we consider is that of unconstrained "sampling-controlled stochastic recursions" (henceforth SCSR), defined through the following recursion:

$$X_{k+1} = X_k + H_k(m_k, X_k), \quad k = 0, 1, 2, \ldots, \qquad \text{(SCSR)}$$

where $X_k \in \mathbb{R}^d$ for all k. The "deterministic analogue" (henceforth DA) of SCSR is

$$x_{k+1} = x_k + h_k(x_k), \quad k = 0, 1, 2, \ldots. \qquad \text{(DA)}$$

The random function $H_k(m_k, x), x \in \mathbb{R}^d$, called the "simulation estimator" should be interpreted as estimating the corresponding deterministic quantity $h_k(x)$ at the point of interest x, after expending m_k amount of simulation effort. We emphasize that the objects $h_k(\cdot)$ and $H_k(m_k, \cdot)$ appearing in (DA) and (SCSR) can be iteration dependent functions. Two illustrative examples are presented in Section 4.

3.1. Assumptions. The following two assumptions are standing assumptions that will be invoked in several of the important results of the paper. Further assumptions will be made as and when required.

ASSUMPTION 3.1. *The recursion (DA) exhibits global convergence to a unique point x^*, that is, the sequence $\{x_k\}$ of iterates generated by (DA) when started with any initial point x_0 satisfies $\lim_{k\to\infty} x_k = x^*$.*

ASSUMPTION 3.2. *Denote the filtration*

$$\mathcal{F}_{k-1} = \sigma\{(X_0, H_0(m_0, X_0)), (X_1, H_1(m_1, X_1)), \ldots, (X_{k-1}, H_{k-1}(m_{k-1}, X_{k-1})), X_k)\}$$

generated by the "history sequence" up to iteration k. (Notice that the random variables that generate the σ-algebra $\mathcal{F}_{k-1}$ include X_k but not $H_k(m_k, X_k)$.) Then the simulation estimator $H_k(m_k, X_k)$ satisfies for $k \geq 0$, with probability one,

$$\mathbb{E}\left[m_k^\alpha \|H_k(m_k, X_k) - h_k(X_k)\| \mid \mathcal{F}_{k-1}\right] \leq \kappa_0 + \kappa_1 \|X_k\|, \qquad (3.1)$$

for some $\alpha > 0$, and where κ_0, κ_1 are some positive constants. We will refer to the constant α as the convergence rate associated with the simulation estimator.

Assumption 3.1 assumes convergence of the deterministic recursion (DA)'s iterates starting from any initial point x_0. Such assumption is needed if we were to expect stochastic iterations in (SCSR) to converge to the correct solution in any reasonable sense. We view the deterministic recursion (DA) to be the "limiting form" of (SCSR), obtained, for example, if the estimator $H_k(m_k, x)$ at hand is a "perfect" estimator of $h_k(x)$, constructed using a hypothetical infinite sample.

Assumption 3.2 is a statement about the behavior of the simulation estimator $H_k(m_k, x), x \in \mathbb{R}^d$, and is analogous to standard assumptions in the literature on stochastic approximation and machine learning, e.g., Assumption A3 in [6] and Assumption 4.3(b),(c) in [8]. In order to develop convergent algorithms for the context we consider in this paper, some sort of restriction on the extent to which a simulation estimator can "mislead" an algorithm is necessary. Assumption 3.2 is a formal codification of such a restriction; it implies that the error in the estimator $H_k(m_k, X_k)$, conditional on the history of the observed random variables up to iteration k, decays with rate α. Furthermore, the manner of such decay can depend on the current iterate X_k. Assumption 3.2 subsumes typical stochastic optimization contexts where the mean squared error of the simulation estimator (with respect to the true objective function value) at any point is bounded by an affine function of the squared $\mathcal{L}_2$-norm of the true gradient at the point, assuming that the gradient function is Lipschitz.

3.2. Work and Efficiency. In the analysis considered throughout this paper, computational effort calculations are limited to simulation effort. Therefore, the total "work done" through k iterations of SCSR is given by

$$W_k = \sum_{i=1}^{k} m_i.$$

Our assessment of any sampling strategy will be based on how fast the error $E_k = \|X_k - x^*\|$ in the kth iterate of SCSR (stochastically) converges to zero *as a function of the total work* W_k. This will usually be achieved by first identifying the convergence rate of E_k with respect to the iteration number k and then translating this rate with respect to the total work W_k.

Under mild conditions, we will demonstrate that E_k cannot converge to zero faster than $W_k^{-\alpha}$ (in a certain rigorous sense), where α is defined through Assumption 3.2. This makes intuitive sense because it seems reasonable to expect that a stochastic recursion's quality is at most as good as the quality of the estimator at hand. We will then deem those recursions having error sequences $\{E_k\}$ that achieve the convergence rate $W_k^{-\alpha}$ as being *efficient*. The convergence rate of E_k with respect to the iteration number k is of little significance.

4. EXAMPLES. In this section, we illustrate SCSR using two popular recursions occurring within the context of SO and SRFPs. For each example, we show the explicit form of the SCSR and the DA recursions through their corresponding functions $H_k(m_k, \cdot)$ and $h_k(\cdot)$. We also identify the estimator convergence rate α in each case.

4.1. Sampling Controlled Gradient Method with Fixed Step. Consider the context of solving an unconstrained optimization problem using the gradient method [9, Section 9.3], usually written as

$$x_{k+1} = x_k + t(-\nabla f(x_k)), \quad k = 0, 1, \ldots \tag{4.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is the real-valued function being optimized, $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is its gradient function, and $t > 0$ is an appropriately chosen constant. (Instead of a fixed stepsize t in (4.1), one might use a diminishing step size sequence $\{t_k\}$ chosen to satisfy $t_k \to 0, \sum_{k=1}^{\infty} t_k = \infty$ [4, Chapter 1].) Owing to its simplicity, the recursion in (4.1) has recently become popular in large-scale SO contexts [8].

Let us now suppose that the gradient function $g(\cdot) \triangleq \nabla f(\cdot)$ in (4.1) is unobservable but we have access to iid observations $G_i(x), i = 1, 2, \ldots$ satisfying $\mathbb{E}[G_i(x)] = g(x)$ for any $x \in \mathbb{R}^d$. The sampling controlled version of the gradient method then takes the form

$$X_{k+1} = X_k + t(-m_k^{-1} \sum_{i=1}^{m_k} G_i(X_k)), \quad k = 0, 1, \ldots \tag{4.2}$$

thus implying the (SCSR) and (DA) recursive objects $H_k(m_k, x) \triangleq t(-m_k^{-1} \sum_{i=1}^{m_k} G_i(x))$, $h_k(x) \triangleq -t\nabla f(x)$ for all $x \in \mathbb{R}^d$. Using standard arguments [22] it can be shown that when f is convex and differentiable with a gradient that satisfies $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$, $L < \infty$, and the step size $t \leq L^{-1}$, the iterates in (4.1) exhibit $\mathcal{O}(k^{-1/2})$ convergence to a zero of ∇f. Furthermore, elementary probabilistic arguments show that Assumption 3.2 is satisfied with rate constant $\alpha = 1/2$.

4.2. Sampling Controlled Kiefer-Wolfowitz Iteration. Consider unconstrained simulation optimization on a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ that is estimated using $F(m, x) = m^{-1} \sum_{i=1}^{m} F_i(x)$, where $F_i(x), x \in \mathbb{R}^d; i = 1, 2, \ldots$ are iid copies of an unbiased function estimator $F(x), x \in \mathbb{R}^d$ of f. Assume that we do not have direct stochastic observations of the gradient function $\nabla f(x)$ so that the current context differs from that in Section 4.1. We thus choose the SCSR iteration to be a modified Kiefer-Wolfowitz [19] iteration constructed using a finite difference approximation of the stochastic function observations. Specifically, recalling the notation $G \triangleq (G^{(1)}, G^{(2)}, \ldots, G^{(d)})$, suppose

$$X_{k+1} = X_k - tG(m_k, X_k), k = 0, 1, \ldots \tag{4.3}$$

where

$$G^{(i)}(m_k, X_k) = \frac{F(m_k, X_k + s_k^{(i)}) - F(m_k, X_k - s_k^{(i)})}{2s_k^{(i)}} \tag{4.4}$$

estimates the ith partial derivative of f at X_k, $s_k \triangleq (s_k^{(1)}, s_k^{(2)}, \ldots, s_k^{(d)})$ is the vector step, and t is an appropriately chosen constant. Assume, for simplicity, that the function observations generated at $X_k - s_k$ are independent of those generated at $X_k + s_k$.

In the notation of (SCSR) and (DA), the simulation estimator $H_k(m_k, x) \triangleq -tG(m_k, x)$ and $h_k(x) \triangleq -t\nabla f(x)$ for all $x \in \mathbb{R}^d$ assuming that s_k is chosen so that $s_k^{(i)} \to 0$ and $\sqrt{m_k} s_k^{(i)} \to \infty, i = 1, 2, \ldots, d$. Furthermore, if s_k is chosen as $s_k^{(i)} = cm_k^{-1/6}$ and f has a bounded third derivative, then Assumption 3.2 is satisfied with $\alpha = 1/3$ [2, Proposition 1.1]. Also, the deterministic recursion (DA) corresponding to (4.3) is the same as that in Section 4.1 and the iteration complexity discussed there applies here as well.

Remark 4.1. *In (4.4), derivative estimators with faster convergence rates can be constructed by estimating higher order derivatives of f. For instance, by observing $G^{(i)}(m_k, x + u_j^{(i)}), j = 1, 2, \ldots, n$ at n strategically located design points $x + u_1, x + u_2, \ldots, x + u_n$, the error $\mathbb{E}[\|H_k(m_k, x) - h_k(x)\|] = \mathcal{O}(m_k^{-n/2n+1})$, that is, the error in the estimator can be made arbitrarily close to the Monte Carlo canonical rate $\mathcal{O}(m_k^{-1/2})$ [2, Chapter VII, Section 1a].*

5. CONSISTENCY. In this section, we present a result that clarifies the conditions on the sampling rates to ensure that the iterates produced by SCSR exhibit almost sure convergence to the solution x^*. We will rely on the following elegant result that appears in slightly more specific form as Lemma 11 on page 50 of [29] .

Lemma 5.1. *Let $\{V_k\}$ be a sequence of nonnegative random variables, where $\mathbb{E}[V_0] < \infty$, and let $\{r_k\}$ and $\{q_k\}$ be deterministic scalar sequences such that*

$$\mathbb{E}[V_{k+1} | V_0, V_1, \ldots, V_k] \leq (1 - r_k)V_k + q_k \textit{ almost surely for } k \geq k_0,$$

where k_0 is fixed, $0 \leq r_k \leq 1, q_k \geq 0$, $\sum_{k=0}^{\infty} r_k = \infty, \sum_{k=0}^{\infty} q_k < \infty, \lim_{k\to\infty} r_k^{-1} q_k = 0$. Then, $\lim_{k\to\infty} V_k = 0$ almost surely and $\lim_{k\to\infty} \mathbb{E}[V_k] = 0$.

We now state the main consistency result for (SCSR) when the corresponding deterministic DA recursion exhibits Sub-Linear(s) or Linear(ℓ) convergence.

THEOREM 5.2. *Let Assumption 3.1 and Assumption 3.2 hold. Let the sample size sequence $\{m_k\}$ satisfy $m_k^{-1} = \mathcal{O}(k^{-\frac{1}{\alpha}-\delta})$ for some $\delta > 0$. (The constant α is the convergence rate of the simulation estimator appearing in Assumption 3.2.)*

(i) Suppose the recursion (DA) guarantees a Sub-Linear(s) decrease at each k, that is, for every x, k and some $s \in (0, 1)$,

$$\|x + h_k(x) - x^*\| \leq (1 - \frac{s}{k})\|x - x^*\|. \tag{5.1}$$

Then $\|X_k - x^*\| \overset{wp1}{\to} 0$ and $\mathbb{E}[\|X_k - x^*\|] \to 0$.

(ii) Suppose the recursion (DA) guarantees a Linear(ℓ) decrease at each k, that is, for every x, k, the recursion (DA) satisfies

$$\|x + h_k(x) - x^*\| \leq \ell\|x - x^*\|. \tag{5.2}$$

Then $\|X_k - x^*\| \overset{wp1}{\to} 0$ and $\mathbb{E}[\|X_k - x^*\|] \to 0$.

Proof. Let us first prove the assertion in (i). Using (SCSR) and recalling the unique solution x^* to the recursion (DA), we can write

$$X_{k+1} - x^* = X_k + h_k(X_k) - x^* + H_k(m_k, X_k) - h_k(X_k), \quad k = 0, 1, 2, \ldots. \tag{5.3}$$

Denoting $E_k = \|X_k - x^*\|$, (5.3) gives

$$E_{k+1} \leq (1 - \frac{s}{k})E_k + \|H_k(m_k, X_k) - h_k(X_k)\|, \quad k = 0, 1, 2, \ldots. \tag{5.4}$$

Now conditioning on $\mathcal{F}_{k-1}$ and then taking expectation on both sides of (5.4), we get

$$\begin{aligned}
\mathbb{E}\left[E_{k+1}|\mathcal{F}_{k-1}\right] &\leq (1 - \frac{s}{k})E_k + \mathbb{E}[\|H_k(m_k, X_k) - h_k(X_k)\||\mathcal{F}_{k-1}] \\
&\leq (1 - \frac{s}{k})E_k + \frac{\kappa_0}{m_k^\alpha} + \frac{\kappa_1\|X_k\|}{m_k^\alpha} \\
&\leq (1 - \frac{s}{k} + \frac{\kappa_1}{m_k^\alpha})E_k + \frac{\kappa_0 + \kappa_1\|x^*\|}{m_k^\alpha} \\
&\leq (1 - \frac{(s - \kappa_1 m_k^{-\alpha} k)}{k})E_k + \frac{\kappa_0 + \kappa_1\|x^*\|}{m_k^\alpha}.
\end{aligned} \tag{5.5}$$

If the sequence $\{m_k\}$ is chosen so that $m_k^{-1} = \mathcal{O}\left(k^{-\frac{1}{\alpha}-\delta}\right)$ for some $\delta > 0$ (as has been postulated by the theorem), then for any given ϵ, we see that $\kappa_1 m_k^{-\alpha} k < \epsilon$ for large enough k. Therefore, after "integrating out" the random variables $H_i(m_i, X_i), i = 0, 1, \ldots, k-1$ in (5.5), we can write for any given $\epsilon \in (0, s)$ and k large enough that

$$\mathbb{E}\left[E_{k+1}|E_0, E_1, \ldots, E_k\right] \leq (1 - \frac{(s - \epsilon)}{k})E_k + \frac{\kappa_0 + \kappa_1\|x^*\|}{m_k^\alpha}. \tag{5.6}$$

Now, if we apply Lemma 5.1 to (5.6) with $r_k \triangleq (s - \epsilon)k^{-1}$ and $q_k \triangleq \beta m_k^{-\alpha}$ for $\beta = \kappa_0 + \kappa_1\|x^*\|$, then $\sum_{k=1}^{\infty} r_k = \sum_{k=1}^{\infty}(s - \epsilon)k^{-1} = \infty$, $\sum_{k=1}^{\infty} q_k = \sum_{k=1}^{\infty} \beta m_k^{-\alpha} =$

$\mathcal{O}(\sum_{k=1}^{\infty} k^{-1-\alpha\delta}) < \infty$, and $\limsup r_k^{-1} q_k \leq \limsup \beta(s-\epsilon)^{-1} k^{-\alpha\delta} = 0$. We thus see that the postulates of Lemma 5.1 hold implying that $E_k \overset{\text{wp1}}{\to} 0$ and $\mathbb{E}[E_k] \to 0$.

Next, suppose the recursion (DA) exhibits Linear(ℓ) convergence. The inequality analogous to (5.6) is then

$$\mathbb{E}\left[E_{k+1}|E_0, E_1, \ldots, E_k\right] \leq (1-(1-\ell-\kappa_1 m_k^{-\alpha}))E_k + \frac{\kappa_0 + \kappa_1\|x^*\|}{m_k^{\alpha}}. \tag{5.7}$$

Since $\{m_k\} \to \infty$, we see that for any given $\epsilon \in (0, 1-\ell)$, for large enough k

$$\mathbb{E}\left[E_{k+1}|E_0, E_1, \ldots, E_k\right] \leq (1-(1-\ell-\epsilon))E_k + \frac{\kappa_0 + \kappa_1\|x^*\|}{m_k^{\alpha}}. \tag{5.8}$$

Now, apply Lemma 5.1 to (5.8) with $r_k \triangleq 1-\ell-\epsilon$ and $q_k \triangleq \beta m_k^{-\alpha}$ for $\beta = \kappa_0+\kappa_1\|x^*\|$. If the sequence $\{m_k\}$ is chosen so that $m_k^{-1} = \mathcal{O}\left(k^{-\frac{1}{\alpha}-\delta}\right)$ for some $\delta > 0$, then $\sum_{k=1}^{\infty} r_k = \sum_{k=1}^{\infty} \ell - \epsilon = \infty$, $\sum_{k=1}^{\infty} q_k = \sum_{k=1}^{\infty} \beta m_k^{-\alpha} = \mathcal{O}(\sum_{k=1}^{\infty} k^{-1-\alpha\delta}) < \infty$ and $\limsup r_k^{-1} q_k \geq \limsup \beta(\ell-\epsilon)^{-1} k^{-1-\alpha\delta} = 0$. We thus see that the postulates of Lemma 5.1 hold implying that $E_k \overset{\text{wp1}}{\to} 0$ and $\mathbb{E}[E_k] \to 0$.
□

It is important to note that the assumed decrease condition, (5.1) or (5.2), is on the (hypothetical) deterministic recursion (DA) and not the stochastic recursion (SCSR). The motivating setting here is unconstrained convex minimization where a decrease such as (5.1) or (5.2) can usually be guaranteed. The theorem can be relaxed to more general settings where the decrease condition (5.2) holds only when X_k is close enough to x^* but as we show later when we characterize convergence rates, we will still need a weak decrease condition such as (5.1) to hold for all X_k. For this reason, part (i) in Theorem 5.2 should be seen as the main result on the strong consistency of SCSR.

The stipulation $m_k^{-1} = \mathcal{O}(k^{-\frac{1}{\alpha}-\delta})$ for some $\delta > 0$ in Theorem 5.2 amounts to a weak stipulation on the sample size increase rate for guaranteeing strong consistency and $\mathcal{L}^1$ convergence. That the minimum stipulated sample size increase depends on the quality (as encoded by the convergence rate α) of the simulation estimator is to be expected. However, part (ii) of Theorem 5.2 implies that the minimum stipulated sample size increase does not depend on the speed of the underlying deterministic recursion as long as it exceeds a sub-linear rate. So, when a linear decrease (5.2) as in part (ii) of Theorem 5.2 is assured, the sample size stipulation $m_k^{-1} = \mathcal{O}(k^{-\frac{1}{\alpha}-\delta})$ needed for strong consistency remains the same. This, as we shall see in greater detail in ensuing sections, is because sampling error dominates the error due to recursion and is hence decisive in determining whether the iterates converge.

6. CONVERGENCE RATES AND EFFICIENCY. In this section, we present results that shed light on the convergence rate and the efficiency of SCSR under different sampling and recursion contexts. Specifically, we derive the convergence rates associated with using various combinations of sample size increases (polynomial, geometric, super-exponential) and the speed of convergence of the DA recursion (sublinear, linear, superlinear). This information is then used to identify what sample size growth rates may be best, that is, *efficient*, for various combinations of recursive structures and simulation estimators. (See Figure 6.1 for a concise and intuitive summary of the results in this section.)

In what follows, convergence rates are first expressed as a function of the iteration k and the various constants associated with sampling and recursion. These obtained rates are then related to the total work done through k iterations of SCSR given by $W_k = \sum_{i=1}^{k} m_i$, in order to obtain a sense of the efficiency.

As we show next, the quantity $W_k^{-\alpha}$ is a stochastic lower bound on the error E_k in the SCSR iterates; thus, loosely speaking, α is an upper bound on the *convergence rate* of the error in SCSR iterates. It is in this sense that we say SCSR's iterates are *efficient* whenever they attain the rate $W_k^{-\alpha}$.

THEOREM 6.1. *Let the postulates of Theorem 5.2 hold with a non-decreasing sample size sequence $\{m_k\}$, and let the recursion (DA) satisfy postulate (i) in Theorem 5.2. Furthermore, suppose there exist $\delta', \epsilon > 0$, and a set $\mathcal{B}_{\delta''}(x^*)$ such that, for large enough k,*

$$\inf_{\{(x,u)\,:x\in\mathcal{B}_{\delta''}(x^*);\|u\|=1\}} \mathbb{P}(m_k^{\alpha}(H_k(m_k,x)-h_k(x))^T u \geq \delta') > \epsilon. \tag{6.1}$$

Then the recursion SCSR cannot converge faster than $W_k^{-\alpha}$, that is, there exists $\tilde{\epsilon} > 0$ such that for any sequence of sample sizes $\{m_k\}$, $\liminf_k \mathbb{P}(W_k^{\alpha} E_k > \delta') > \tilde{\epsilon}$.

Proof. Since the postulates of Theorem 5.2 are satisfied, we are guaranteed that $\|X_k - x^*\| \overset{\text{wp1}}{\to} 0$ and hence that $\|X_k - x^*\| \overset{\text{p}}{\to} 0$. For proving the theorem, we will show that for large enough k, $\mathbb{P}(m_k^{\alpha} E_k \geq \delta') > \tilde{\epsilon}$, where $\tilde{\epsilon} > 0$. Since $W_k = \sum_{j=1}^{k} m_j \geq m_k$, the assertion of Theorem 6.1 will then hold.

Choose $\delta''' = \min(\delta', \delta'')$, where δ'' is the constant appearing in (6.1). Since $\{X_k\} \overset{\text{p}}{\to} x^*$, for large enough k, we have

$$\mathbb{P}(X_k \in \mathcal{B}_{\delta'''}(x^*)) \geq 1 - \epsilon. \tag{6.2}$$

Denoting $U_k(X_k) := X_k + h_k(X_k) - x^*$, we can write for large enough k,

$$\mathbb{P}(m_{k+1}^{\alpha} E_{k+1} \geq \delta') \geq \mathbb{P}(m_k^{\alpha} E_{k+1} \geq \delta') = \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2), \tag{6.3}$$

where the events $\mathcal{A}_1$ and $\mathcal{A}_2$ in (6.3) are defined as follows.

$$\begin{aligned} \mathcal{A}_1 &:= (m_k^{\alpha} E_{k+1} \geq \delta') \cap (U_k(X_k) \neq 0)\,; \\ \mathcal{A}_2 &:= (m_k^{\alpha} E_{k+1} \geq \delta') \cap (U_k(X_k) = 0)\,. \end{aligned} \tag{6.4}$$

We also define the following two other events.

$$\begin{aligned} \mathcal{C}_1 &:= \left(m_k^{\alpha}(H_k(m_k,X_k)-h_k(X_k))^T\, U_k(X_k) \geq \delta'\|U_k(X_k)\|\right) \cap (U_k(X_k) \neq 0)\,; \\ \mathcal{C}_2 &:= (m_k^{\alpha}\|H_k(m_k,X_k)-h_k(X_k)\| \geq \delta') \cap (U_k(X_k) = 0)\,. \end{aligned} \tag{6.5}$$

Since $E_{k+1} = \|X_k + H_k(m_k,X_k) - x^*\|$, we notice that

$$E_{k+1}^2 \geq 2(H_k(m_k,X_k)-h_k(X_k))^T U_k(X_k) + \|H_k(m_k,X_k)-h_k(X_k)\|^2. \tag{6.6}$$

Due to (6.6) and the Cauchy-Schwarz inequality [5], we see that $\mathcal{A}_1 \supseteq \mathcal{C}_1$; due to (6.6), we also see that $\mathcal{A}_2 \supseteq \mathcal{C}_2$. Hence

$$\mathbb{P}(\mathcal{A}_1) \geq \mathbb{P}(\mathcal{C}_1) \text{ and } \mathbb{P}(\mathcal{A}_2) \geq \mathbb{P}(\mathcal{C}_2). \tag{6.7}$$

Define $\mathcal{R}_k := \{x : U_k(x) = 0\}$. Then, due to the assumption in (6.1), we see that for any $x \in \mathcal{B}_{\delta'''}(x^*) \cap \mathcal{R}_k^c$,

$$\mathbb{P}(\mathcal{C}_1 \mid X_k = x) > \epsilon. \tag{6.8}$$

And, since Cauchy-Schwarz inequality [5] implies that $m_k^\alpha \|H_k(m_k, X_k) - h_k(X_k)\| \geq m_k^\alpha (H_k(m_k, X_k) - h_k(X_k))^T u$ for any unit vector u, we again see from the assumption in (6.1) that for any $x \in \mathcal{B}_{\delta'''}(x^*) \cap \mathcal{R}_k$,

$$\mathbb{P}(\mathcal{C}_2 \mid X_k = x) > \epsilon. \tag{6.9}$$

Next, letting F_{X_k} denote the distribution function of X_k, we write

$$\begin{aligned}\mathbb{P}(\mathcal{C}_1) &= \int \mathbb{P}(\mathcal{C}_1 \mid X_k = x)\, dF_{X_k}(x) \\ &\geq \int \mathbb{P}(\mathcal{C}_1 \mid X_k = x) \mathbb{I}\{x \in \mathcal{B}_{\delta'''}(x^*) \cap \mathcal{R}_k^c\}\, dF_{X_k}(x) \\ &\geq \epsilon\, \mathbb{P}(X_k \in \mathcal{B}_{\delta'''}(x^*) \cap \mathcal{R}_k^c),\end{aligned} \tag{6.10}$$

where the second inequality in (6.10) follows from (6.8). Similarly,

$$\mathbb{P}(\mathcal{C}_2) \geq \epsilon\, \mathbb{P}(X_k \in \mathcal{B}_{\delta'''}(x^*) \cap \mathcal{R}_k). \tag{6.11}$$

Combining (6.10), (6.11) and (6.2), we see that for large enough k,

$$\mathbb{P}(\mathcal{C}_1) + \mathbb{P}(\mathcal{C}_2) \geq \epsilon\,(1 - \epsilon). \tag{6.12}$$

Finally, from (6.12), (6.7), and (6.3), we conclude that for large enough k,

$$\mathbb{P}(m_k^\alpha E_k \geq \delta') > \tilde{\epsilon} := \epsilon(1 - \epsilon), \tag{6.13}$$

and the theorem is proved.
□

Theorem 6.1 is important in that it provides a benchmark for efficiency. Specifically, Theorem 6.1 implies that sampling and recursion choices that result in errors achieving the rate $W_k^{-\alpha}$ are efficient. Theorem 6.1 relies on the assumption in (6.1) which puts an upper bound on the quality (convergence rate) of the simulation estimator $H_k(m_k, x)$; the reader will recall that Assumption 3.2 puts a lower bound on the quality of $H_k(m_k, x)$. The condition in (6.1) is weak especially since $H_k(m_k, x)$, being a simulation estimator, is routinely governed by a central limit theorem (CLT) [5] of the form $m_k^\alpha (H_k(m_k, x) - h_k(x)) \xrightarrow{\text{d}} N(0, \Sigma(x))$ as $k \to \infty$, where $N(0, \Sigma(x))$ is a normal random variable with zero mean and covariance $\Sigma(x)$.

We emphasize that Theorem 6.1 only says that α is an upper bound for the convergence rate of SCSR, and says nothing about whether this rate is in fact achievable. We will now work towards a general lower bound on the sampling rates that achieve efficiency. We will need the following lemma for proving such a lower bound.

Lemma 6.2. *Let $\{a_k\}$ be any positive-valued sequence. Then*

(i) $a_k = \Theta(\sum_{j=1}^k a_j)$ if $\{a_k\}$ Geometric(c) or faster;

(ii) $a_k = o(\sum_{j=1}^k a_j)$ if $\{a_k\}$ is Polynomial(λ_p, p) or slower.

Proof. Proof of (i). If $\{a_k\}$ is Geometric(c) or faster, we know that $a_{k+1}/a_k \geq c > 1$ for large enough k. Hence, for some k_0 and all $k \geq j \geq k_0$, $a_j/a_k \leq c^{j-k}$. This implies that for $k \geq k_0$,

$$a_k^{-1} \sum_{j=1}^{k} a_j = a_k^{-1} \sum_{j=1}^{k_0} a_j + a_k^{-1} \sum_{j=k_0+1}^{k} a_j$$
$$\leq a_k^{-1} \sum_{j=1}^{k_0} a_j + \sum_{j=k_0+1}^{k} c^{j-k}$$
$$= a_k^{-1} \sum_{j=1}^{k_0} a_j + \sum_{j=0}^{k-k_0-1} c^{-j} \to \frac{1}{1-c}. \tag{6.14}$$

Using (6.14) and since $a_k \leq \sum_{j=1}^{k} a_j$, we conclude that the assertion holds.

Proof of (ii). Let $p > 0$ be such that $\{a_k\}$ is Polyomial(λ_p, p) or slower. We then know that for some $k_0 > 0$ and all $k \geq j \geq k_0$, $a_j/a_k \geq j^p/k^p$. This implies that

$$a_k^{-1} \sum_{j=1}^{k} a_j = a_k^{-1} \sum_{j=1}^{k_0} a_j + a_k^{-1} \sum_{j=k_0+1}^{k} a_j \geq a_k^{-1} \sum_{j=1}^{k_0} a_j + k^{-p} \sum_{j=k_0+1}^{k} j^p. \tag{6.15}$$

Now notice that the term $k^{-p} \sum_{j=k_0+1}^{k} j^p$ appearing on the right-hand side of (6.15) diverges as $k \to \infty$ to conclude that the assertion in (ii) holds. □

We are now ready to present a lower bound on the rate at which sample sizes should be increased in order to ensure optimal convergence rates.

Theorem 6.3. *Let the postulates of Theorem 6.1 hold.*

(i) Suppose $m_k = o(W_k)$. Then the sequence of solutions $\{X_k\}$ is such that $W_k^\alpha E_k$ is not $\mathcal{O}_p(1)$, that is, there exists $\tilde{\epsilon} > 0$ and a subsequence $\{k_n\}$ such that $\mathbb{P}(W_{k_n}^\alpha E_{k_n} \geq n) > \tilde{\epsilon}$.

(ii) If $\{m_k\}$ grows as Polynomial(λ_p, p), then $W_k^\alpha E_k$ is not $\mathcal{O}_p(1)$.

Proof. Proof of (i). The postulates of Theorem 6.1 hold and hence we know from (6.13) in the proof of Theorem 6.1 that there exists $K_1(\delta', \epsilon)$ such that for $k \geq K_1(\delta', \epsilon)$,

$$\mathbb{P}(m_k^\alpha E_k \geq \delta') > (1-\epsilon)\epsilon, \tag{6.16}$$

where the constants ϵ, δ' are positive constants that satisfy the assumption in (6.1).

Since $m_k = o(W_k)$ and $\alpha > 0$, we see that $m_k^\alpha / W_k^\alpha = o(1)$ as $k \to \infty$. Therefore, for any $n > 0$, there exists $K_2(n)$ such that for $k \geq K_2(n)$,

$$\delta' \geq n\, m_k^\alpha / W_k^\alpha. \tag{6.17}$$

Combining (6.16), (6.17), we see that for any $n > 0$, if $k \geq \max(K_1(\delta', \epsilon), K_2(n))$, then

$$\mathbb{P}(m_k^\alpha E_k \geq n \frac{m_k^\alpha}{W_k^\alpha}) > (1-\epsilon)\epsilon, \tag{6.18}$$

and hence, for $k \geq \max(K_1(\delta', \epsilon), K_2(n))$,

$$\mathbb{P}(W_k^\alpha E_k \geq n) > (1-\epsilon)\epsilon. \tag{6.19}$$

Proof of (ii). The assertion is seen to be true from the assertion in (i) and upon noticing that if $\{m_k\}$ grows as Polynomial(λ_p, p), then $m_k = o(W_k)$. □

Theorem 6.3 is important since it's assertions imply that for SCSR to have any chance of efficiency, sample sizes should be increased at least geometrically. This is irrespective of the speed of the recursion DA. Of course, since this is only a lower bound, increasing the sample size at least geometrically does not guarantee efficiency, which, as we shall see, depends on the speed of the DA recursion. Before we present such an efficiency result for linearly converging DA recursions, we need two more lemmas.

LEMMA 6.4. *Let* $\{a_j(k)\}_{j=1}^k, k \geq 1$ *be a triangular array of positive-valued real numbers. Assume that the following hold.*

(i) There exists j^* *and* $\beta > 1$ *such that* $\dfrac{a_{j+1}(k)}{a_j(k)} \geq \beta$ *for all* $j \in [j^*, k-1]$ *and all* $k \geq 1$.

(ii) $\limsup_k \dfrac{a_j(k)}{a_k(k)} = \ell_j < \infty$ *for each* $j \in [1, j^*-1]$.

Then $S_k = \sum_{i=1}^k a_i(k) = \mathcal{O}(a_k(k))$.

Proof. We have, for large enough k and any $\epsilon \in (0, \infty)$,

$$\begin{aligned} S_k &= a_k(k)\left(\sum_{j=0}^{j^*-1}\frac{a_j(k)}{a_k(k)} + \sum_{j=j^*}^{k}\frac{a_j(k)}{a_k(k)}\right) \\ &\leq a_k(k)\left(j^*\epsilon + \sum_{j=0}^{j^*-1}\ell_j + \sum_{j=j^*}^{k}\beta^{j-k}\right), \end{aligned} \tag{6.20}$$

where the inequality follows from assumptions (i) and (ii). Since $\beta > 1, j^* < \infty$, and $\ell_j < \infty$, the term within parenthesis on the right-hand side of (6.20) is finite and the assertion holds. ∎

LEMMA 6.5. *Let* $\{S_n\}$ *be a non-negative sequence of random variables,* N_0 *a well-defined random variable, and* $\{a_n\}, \{b_n\}$ *positive-valued deterministic sequences.*

(i) If $\mathbb{E}[S_n] = \mathcal{O}(a_n)$, *then* $S_n = \mathcal{O}_p(a_n)$.

(ii) If $S_n \leq \mathcal{O}_p(b_n)$ *for* $n \geq N_0$, *then* $S_n = \mathcal{O}_p(b_n)$.

Proof. Suppose the first assertion is false. Then there exists $\epsilon > 0$ and a subsequence $\{n_j\} \to \infty$ such that $\mathbb{P}\left(\frac{S_{n_j}}{a_{n_j}} \geq j\right) \geq \epsilon$ for all $j \geq 1$. This, however, implies that $\mathbb{E}\left[\frac{S_{n_j}}{a_{n_j}}\right] \geq j\epsilon$ for all $j \geq 1$, contradicting the postulate $\mathbb{E}[S_n] = \mathcal{O}(a_n)$. The first assertion of the lemma is thus proved.

For proving the second assertion, we first note that the postulate $S_n \leq \mathcal{O}_p(b_n)$ for $n \geq N_0$ means that $S_n \leq B_n$ for $n \geq N_0$, where $\{B_n\}$ is a sequence of random variables satisfying $B_n = \mathcal{O}_p(b_n)$. Now, since $B_n = \mathcal{O}_p(b_n)$, given $\epsilon > 0$, we can choose $b(\epsilon), n_1(\epsilon)$ so that $\mathbb{P}(B_n/b_n \geq b(\epsilon)) \leq \epsilon/2$ for all $n \geq n_1(\epsilon)$. Also, since N_0 is a well-defined random variable we can find $n_2(\epsilon)$ such that for all $n \geq n_2(\epsilon)$, $\mathbb{P}(N_0 > n) \leq \epsilon/2$. We can then write for $n \geq \max(n_1(\epsilon), n_2(\epsilon))$ that

$$\begin{aligned} \mathbb{P}(S_n/b_n \geq b(\epsilon)) &\leq \mathbb{P}((B_n/b_n \geq b(\epsilon)) \cap (N_0 \leq n)) + \mathbb{P}((B_n/b_n \geq b(\epsilon)) \cap (N_0 > n)) \\ &\leq \mathbb{P}(B_n/b_n \geq b(\epsilon)) + \mathbb{P}(N_0 > n) \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned} \tag{6.21}$$

thus proving the assertion in (ii).

∎

We are now ready to prove the main result on the convergence rate and efficiency of SCSR when the DA recursion exhibits linear convergence. Theorem 6.6 presents the convergence rate in terms of the iteration number k first, and then in terms of the total simulation work W_k.

THEOREM 6.6. *(Linearly Converging DA) Let Assumption 3.1 and Assumption 3.2 hold. Also, suppose the following two assumptions hold.*

A.1 The deterministic recursion (DA) exhibits Linear(ℓ) convergence in a neighborhood around x^*, that is, there exists a neighborhood $\mathcal{V}$ of x^* such that whenever $x \in \mathcal{V}$, and for all k,

$$\|x + h_k(x) - x^*\| \leq \ell \|x - x^*\|, \ \ell \in (0,1).$$

A.2 For all x, k,

$$\|x + h_k(x) - x^*\| \leq (1 - \frac{s}{k})\|x - x^*\|, \ s \in (0,1).$$

Then, recalling that $E_k \triangleq \|X_k - x^*\|$, as $k \to \infty$, the following hold.

(i)

$$E_k = \begin{cases} \mathcal{O}_p(k^{-p\alpha}), & \text{if } \{m_k\} \text{ grows as } Polynomial(\lambda_p, p), p\alpha > 1; \\ \mathcal{O}_p\left(c^{-k\alpha}\right), & \text{if } \{m_k\} \text{ grows as } Geometric(c) \text{ with } c \in (1, \ell^{-1/\alpha}); \\ \mathcal{O}_p\left(\ell^{k}\right), & \text{if } \{m_k\} \text{ grows as } Geometric(c) \text{ with } c \geq \ell^{-1/\alpha}; \\ \mathcal{O}_p(\ell^k), & \text{if } \{m_k\} \text{ grows as } SupExponential(\lambda_t, t). \end{cases}$$

(ii)

$$\begin{aligned} W_k^{\alpha\frac{p}{p+1}} E_k &= \mathcal{O}_p(1), && \text{if } \{m_k\} \text{ grows as } Polynomial(\lambda_p, p), p\alpha > 1; \\ W_k^{\alpha} E_k &= \mathcal{O}_p\left(1\right), && \text{if } \{m_k\} \text{ grows as } Geometric(c) \text{ with } c \in (1, \ell^{-1/\alpha}); \\ \left(c^{-\alpha}\ell^{-1}\right)^k W_k^{\alpha} E_k &= \mathcal{O}_p\left(1\right), && \text{if } \{m_k\} \text{ grows as } Geometric(c) \text{ with } c \geq \ell^{-1/\alpha}; \\ \left(\log W_k\right)^{\log_t(1/\ell)} E_k &= \mathcal{O}_p(1), && \text{if } \{m_k\} \text{ grows as } SupExponential(\lambda_t, t). \end{aligned}$$

Proof. First we see that Assumption 3.1, Assumption 3.2, and A.2 hold, and that all sample size sequences $\{m_k\}$ considered in (i) and (ii) satisfy $m_k = \mathcal{O}(k^{-\frac{1}{\alpha}-\delta})$ for some $\delta > 0$. We thus see that the postulates of Theorem 5.2 hold, implying that $E_k = \|X_k - x^*\| \overset{\text{wp1}}{\to} 0$. Therefore, excluding a set of measure zero, for any given $\Delta > 0$ there exists a well-defined random variable $K_0 = K_0(\Delta)$ such that $\|X_k - x^*\| \leq \Delta$ for $k \geq K_0$. Now choose Δ such that the ball $\mathcal{B}_\Delta(x^*) \subset \mathcal{V}$, where $\mathcal{V}$ is the neighborhood appearing in Assumption A.1. Since $X_{k+1} = X_k + H_k(m_k, X_k)$ we can write $X_{K_0+k+1} - x^* = X_{K_0+k} - x^* + h_{K_0+k}(X_{K_0+k}) + H_{K_0+k}(m_{K_0+k}, X_{K_0+k}) - h_{K_0+k}(X_{K_0+k})$ and hence

$$\|X_{K_0+k+1} - x^*\| \leq \ell \|X_{K_0+k} - x^*\| + \|H_{K_0+k}(m_{K_0+k}, X_{K_0+k}) - h_{K_0+k}(X_{K_0+k})\|. \tag{6.22}$$

Recursing (6.22) backwards and recalling the notation $E_k \triangleq \|X_k - x^*\|$, we have for

$k \geq 0$

$$
\begin{aligned}
E_{K_0+k+1} &\leq \ell^{k+1} E_{K_0} + \sum_{j=0}^{k} \ell^{k-j} \| H_{K_0+j}(m_{K_0+j}, X_{K_0+j}) - h_{K_0+j}(X_{K_0+j}) \| \\
&\leq \ell^{k+1} \Delta + \sum_{j=0}^{k} \ell^{k-j} \| H_{K_0+j}(m_{K_0+j}, X_{K_0+j}) - h_{K_0+j}(X_{K_0+j}) \| \\
&= \ell^{k+1} \Delta + \sum_{j=K_0}^{k+K_0} \ell^{k+K_0-j} \| H_j(m_j, X_j) - h_j(X_j) \| \\
&\leq \ell^{k+1} \Delta + \sum_{j=0}^{k+K_0} \ell^{k+K_0-j} \| H_j(m_j, X_j) - h_j(X_j) \| \qquad (6.23)
\end{aligned}
$$

where the second inequality above follows from the definition of K_0, the equality follows after relabeling $j \equiv j + K_0$, and the third inequality follows from the addition of some positive terms to the summation.

Relabeling $k \equiv K_0 + k$ in (6.23) and denoting $\zeta_j := H_j(m_j, X_j) - h(j(X_j)$, we can write for $k \geq K_0$,

$$
E_{k+1} \leq \ell^{k-K_0+1} \Delta + \sum_{j=0}^{k} \ell^{k-j} \| \zeta_j \|. \qquad (6.24)
$$

Recalling the filtration $\mathcal{F}_{k-1}$ generated by the history sequence, we notice that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{j=0}^{k} \ell^{k-j} \| \zeta_j \| \right] &= \sum_{j=0}^{k} \ell^{k-j} \mathbb{E}\left[\| \zeta_j \| \right] \\
&= \sum_{j=0}^{k} \ell^{k-j} \mathbb{E}\left[\mathbb{E}\left[\| \zeta_j \| \mid \mathcal{F}_{j-1} \right] \right] \\
&\leq \sum_{j=0}^{k} \ell^{k-j} \mathbb{E}\left[m_j^{-\alpha} (\kappa_0 + \kappa_1 \| X_j \|) \right] \\
&\leq \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha} (\kappa_0 + \kappa_1 \| x^* \| + \kappa_1 \mathbb{E}\left[E_j \right]), \qquad (6.25)
\end{aligned}
$$

where the first inequality in (6.25) is due to Assumption 3.2.

Due to Theorem 5.2, we know that for a given $\epsilon > 0$, there exists $k_0(\epsilon)$ such that for all $k \geq k(\epsilon)$, $\mathbb{E}[E_j] \leq \epsilon$. We use this in (6.25) and write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=0}^{k} \ell^{k-j}\|\zeta_j\|\right] \\
&\quad \leq \sum_{j=0}^{k_0(\epsilon)} \ell^{k-j} m_j^{-\alpha}(\kappa_0 + \kappa_1\|x^*\| + \kappa_1\mathbb{E}[E_j]) + \sum_{j=k_0(\epsilon)+1}^{k} \ell^{k-j} m_j^{-\alpha}(\kappa_0 + \kappa_1\|x^*\| + \kappa_1\epsilon) \\
&\quad \leq \ell^{k+1}\left(\sum_{j=0}^{k_0(\epsilon)} \ell^{-j-1} m_j^{-\alpha}(\kappa_0 + \kappa_1\|x^*\| + \kappa_1\mathbb{E}[E_j])\right) + (\kappa_0 + \kappa_1\|x^*\| + \kappa_1\epsilon)\sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}.
\end{aligned} \tag{6.26}$$

Since $k_0(\epsilon)$ is finite and $\mathbb{E}[E_j] < \infty$ for all $k \leq k_0(\epsilon)$, the inequality in (6.26) implies that

$$\mathbb{E}\left[\sum_{j=0}^{k} \ell^{k-j}\|\zeta_j\|\right] = \mathcal{O}\left(\ell^{k+1} + \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}\right). \tag{6.27}$$

From part (i) of Lemma 6.5, we know that if a positive random sequence $\{S_k\}$ satisfies $\mathbb{E}[S_k] = \mathcal{O}(a_k)$, where $\{a_k\}$ is a positive-valued deterministic sequence, then $S_k/a_k = \mathcal{O}_p(1)$. Therefore we see from (6.27) (after setting S_k to be $\sum_{j=0}^{k} \ell^{k-j}\|\zeta_j\|$ and a_k to be $\ell^{k+1} + \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}$) that

$$\sum_{j=0}^{k} \ell^{k-j}\|\zeta_j\| = \mathcal{O}_p\left(\ell^{k+1} + \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}\right). \tag{6.28}$$

Use (6.24) and (6.28) to write, for $k \geq K_0$,

$$\begin{aligned}
E_{k+1} &\leq \ell^{k+1}\ell^{-K_0}\Delta + \mathcal{O}_p\left(\ell^{k+1} + \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}\right) \\
&= \mathcal{O}_p\left(\ell^{k+1} + \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}\right).
\end{aligned} \tag{6.29}$$

Now use part (ii) of Lemma 6.5 on (6.29) to conclude that

$$E_{k+1} = \mathcal{O}_p\left(\ell^{k+1} + \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}\right). \tag{6.30}$$

We will now show that the first equality in assertion (i) of Theorem 6.6 holds by showing that the two assumptions of Lemma 6.4 hold for $\sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}$ appearing in (6.30). Set the summand of $\sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}$ to $a_j(k)$ and since $m_j = \lambda_p j^p$, we have $\dfrac{a_{j+1}(k)}{a_j(k)} = \dfrac{1}{\ell}(\dfrac{j}{j+1})^{p\alpha}$. Choosing β such that $\beta > 1$ and $\ell\beta < 1$, and setting $j^* = \mathrm{Max}(1, \dfrac{1}{1-(\ell\beta)^{1/p\alpha}} - 1)$, we see that the first assumption of Lemma 6.4 is satisfied. The second assumption of Lemma 6.4 is also satisfied since for any fixed $j^* > 0$, $\limsup_k \dfrac{a_j(k)}{a_k(k)} = \limsup_k \ell^{k-j}(\dfrac{k}{j})^{p\alpha} = 0$ for all $j \in [1, j^*]$.

To prove the second and third equalities in assertion (i) of Theorem 6.6, suppose $\{m_k\}$ grows as Geometric(c) with $c < \ell^{-1/\alpha}$, that is, $c^{-\alpha} > \ell$. Then, noticing that $m_j = m_0 c^j$, we write

$$\sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha} = m_0^{-\alpha} \sum_{j=0}^{k} \ell^{k-j} c^{-j\alpha} = m_0^{-\alpha} c^{-k\alpha} \frac{(1-(\dfrac{\ell}{c^{-\alpha}})^{k+1})}{1-\dfrac{\ell}{c^{-\alpha}}}$$
$$= \Theta(c^{-k\alpha}), \tag{6.31}$$

and use (6.31) in (6.30). If $\{m_k\}$ grows as Geometric(c) with $c > \ell^{-1/\alpha}$, that is, $c^{-\alpha} < \ell$, then notice that (6.31) becomes

$$\sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha} = m_0^{-\alpha} \sum_{j=0}^{k} \ell^{k-j} c^{-j\alpha} = m_0^{-\alpha} (\frac{c^{-\alpha}}{\ell}) \ell^k \frac{(1-(\dfrac{c^{-\alpha}}{\ell})^{k+1})}{1-\dfrac{c^{-\alpha}}{\ell}}$$
$$= \Theta(\ell^k). \tag{6.32}$$

Now use (6.32) in (6.30).

To see that the fourth equality in assertion (i) of Theorem 6.6 holds, we notice that a sample size sequence $\{m_k\}$ that grows as SupExponential(λ_t, t) is faster (as defined in Definition 2.2) than a sample size sequence $\{m_k\}$ that grows as Geometric(c).

Proof of (ii). To prove the assertion in (ii), we notice that since $W_k = \sum_{j=1}^{k} m_j$, and we have

$$\begin{aligned} W_k &= \Theta(k^{p+1}) \text{ if } \{m_k\} \text{ grows as Polynomial}(\lambda_p, p); \\ &= \Theta(c^k) \text{ if } \{m_k\} \text{ grows as Geometric}(c); \\ &= \Theta\left((\lambda_t^{\frac{1}{t-1}} m_0)^{t^k}\right) \text{ if } \{m_k\} \text{ grows as SupExponential}(\lambda_t, t). \end{aligned} \tag{6.33}$$

Now use (6.33) in assertion (i) to obtain the assertion in (ii). □

Theorem 6.6 provides various insights about the behavior of the error in SCSR iterates. For instance, the error structures detailed in (i) of Theorem 6.6 suggest two well-defined sampling regimes where only one of the two error types, sampling error or recursion error, is dominant. Specifically, note that $E_k = \mathcal{O}_p(k^{-p\alpha})$ when the sampling rate is Polyonimal(λ_p, p). This implies that when DA exhibits Linear(ℓ) convergence, polynomial sampling is "too little" in the sense that SCSR's convergence rate is dictated purely by sampling error since the constant ℓ corresponding to DA's convergence is absent in the expression for E_k. The corresponding reduction in efficiency can be seen in (ii) where E_k is shown to converge as $\mathcal{O}_p(W_k^{-\alpha\frac{p}{1+p}})$. (Recall that efficiency amounts to $\{E_k\}$ achieving a convergence rate $\mathcal{O}_p(W_k^{-\alpha})$.)

The case that is diametrically opposite to polynomial sampling is super-exponential sampling, where the sampling is "too much" in the sense that the convergence rate $E_k = \mathcal{O}_p(\ell^k)$ is dominated by recursion error. There is a corresponding reduction in efficiency as can be seen in the expression provided in (ii) of Theorem 6.6.

The assertion (ii) in Theorem 6.6 also implies that the only sampling regime that achieves efficiency for linearly converging DA recursions is a Geometric(c) sampling rate with $c \in (1, \ell^{-1/\alpha})$. Values of c on or above the threshold $\ell^{-1/\alpha}$ result in "too much" sampling in the sense of a dominating recursion error and a corresponding reduction in efficiency, as quantified in (i) and (ii) of Theorem 6.6.

Before we state a result that is analogous to Theorem 6.6 for the context of superlinearly converging DA recursions, we state and prove a lemma that will be useful.

LEMMA 6.7. *Suppose $\{a_j\}$ is a positive-valued sequence satisfying $a_j^{1/q^j} \to 1$ as $j \to \infty$, where $q > 1$. If $\Lambda \in (0,1)$ is a well-defined random variable, then $a_j \Lambda^{q^j} \xrightarrow{p} 0$.*

Proof. Since $\Lambda \in (0,1)$, for any given $\epsilon \in (0,1)$, there exists $\delta_1(\epsilon) > 0$ such that $\mathbb{P}(\Lambda > (1+\delta_1(\epsilon))^{-1}) \leq \epsilon$. Also, since $a_j^{1/q^j} \to 1$ as $j \to \infty$, we can find N_1 such that for all $j \geq N_1$, $a_j^{1/q^j} > 1 - \delta_1^2(\epsilon)$. Therefore we see that for any given $\epsilon > 0$, there exists $\delta_1(\epsilon) > 0$ such that for $j \geq N_1$,

$$\mathbb{P}(a_j^{1/q^j} \Lambda > (1-\delta_1^2(\epsilon))(1+\delta_1(\epsilon))^{-1}) = \mathbb{P}(a_j^{1/q^j} \Lambda > 1 - \delta_1(\epsilon)) \leq \epsilon. \tag{6.34}$$

Also notice that, for any given $\delta_0 > 0$, we can choose N_2 such that $(1-\delta_1(\epsilon))^{q^j} \leq \delta_0$ for $j \geq N_2$. Using this observation with (6.34), we see that, for any given $\delta_0 > 0$ and $\epsilon > 0$, $\mathbb{P}(a_j \Lambda^{q^j} > \delta_0) \leq \epsilon$ for $j \geq \max(N_1, N_2)$, and the assertion of the lemma holds.□

THEOREM 6.8. *(SuperLinearly Converging DA) Let Assumption 3.1, Assumption A.2 from Theorem 6.6, and the following assumption on superlinear decrease hold.*

A.3 The deterministic recursion (DA) exhibits SuperLinear(λ_q, q) convergence in a neighborhood around x^*, that is, there exists a neighborhood $\mathcal{V}$ of x^* and constants $\lambda_q > 0, q > 1$ such that whenever $x \in \mathcal{V}$, and for all k,

$$\|x + h_k(x) - x^*\| \leq \lambda_q \|x - x^*\|^q.$$

Also, suppose the simulation estimator $H_k(m_k, X_k)$ satisfies for all $k \geq 0, n \geq 0$, with probability one,

$$\mathbb{E}[m_k^{\alpha n} \|H_k(m_k, X_k) - h_k(X_k)\|^n \mid \mathcal{F}_{k-1}] \leq \kappa_0^n + \kappa_1^n \|X_k\|, \tag{6.35}$$

for some $\alpha > 0$, and where κ_0 and $\kappa_1 > 0$ are constants. Then, as $k \to \infty$, the following hold.

(i)

$$E_k = \begin{cases} \mathcal{O}_p(k^{-\alpha p}), & \text{if } \{m_k\} \text{ grows as Polynomial}(\lambda_p, p), p\alpha > 1; \\ \mathcal{O}_p\left(c^{-\alpha k}\right), & \text{if } \{m_k\} \text{ grows as Geometric}(c); \\ \mathcal{O}_p(c_1^{-\alpha t^k}), & \text{if } \{m_k\} \text{ grows as SupExponential}(\lambda_t, t), t \in (1,q); \\ \mathcal{O}_p(\Lambda^{q^k} + c_2^{-\alpha q^k}), & \text{if } \{m_k\} \text{ grows as SupExponential}(\lambda_t, t), t \geq q, \end{cases}$$

where $c_1 = m_0 \lambda_t^{1/(t-1)}$, $c_2 = \kappa^{-\frac{1}{\alpha}} (2\lambda_q^{\frac{1}{q}})^{-\frac{q}{\alpha(q-1)}} m_0$, $\kappa = \max(\kappa_0, \kappa_1)$, and Λ is a random variable that satisfies $\Lambda \in (0,1)$.

(ii)

$$\begin{aligned} W_k^{\alpha\frac{p}{p+1}} E_k &= \mathcal{O}_p(1), && \text{if } \{m_k\} \text{ grows as Polynomial}(\lambda_p, p), p\alpha > 1; \\ W_k^{\alpha} E_k &= \mathcal{O}_p(1), && \text{if } \{m_k\} \text{ grows as Geometric}(c); \\ W_k^{\alpha} E_k &= \mathcal{O}_p(1), && \text{if } \{m_k\} \text{ grows as SupExponential}(\lambda_t, t), t \in (1,q); \\ \left(\Lambda^{-\tilde{W}_k} + c_2^{\alpha \tilde{W}_k}\right) E_k &= \mathcal{O}_p(1), && \text{if } \{m_k\} \text{ grows as SupExponential}(\lambda_t, t), t \geq q, \end{aligned}$$

where $\tilde{W}_k = \left(\log_{c_1} W_k\right)^{\log_t q}$.

Proof. Repeating arguments leading to (6.22) in the proof of Theorem 6.6, we write

$$E_{K_0+k+1} \le \lambda_q E^q_{K_0+k} + \|\zeta_{K_0+k}\|, \tag{6.36}$$

where $\zeta_{K_0+k} = \|h_{K_0+k}(X_{K_0+k}) - H_{K_0+k}(m_{K_0+k}, X_{K_0+k})\|$, and, as in the proof of Theorem 6.6, K_0 is a random variable such that, except for a set of measure zero, $\|X_k - x^*\| \le \Delta$ for $k \ge K_0$; the constant Δ is chosen such that the ball $\mathcal{B}_\Delta(x^*) \subset \mathcal{V}$ and $\Delta < (2\lambda_q)^{\frac{-1}{q-1}}$, where the set $\mathcal{V}$ is the neighborhood appearing in A.3.

Denote $s(n) := 1 + q + \cdots + q^{n-1} = (q^n - 1)/(q-1), n \ge 1$ and recurse (6.36) to obtain for $k \ge 0$,

$$\begin{aligned}
E_{K_0+k+1} &\le \lambda_q E^q_{K_0+k} + \|\zeta_{K_0+k}\| \\
&\le \lambda_q \left(\lambda_q E^q_{K_0+k-1} + \|\zeta_{K_0+k-1}\|\right)^q + \|\zeta_{K_0+k}\| \\
&\le \lambda_q \left(2^q \lambda_q^q E^{q^2}_{K_0+k-1} + 2^q \|\zeta_{K_0+k-1}\|^q\right) + \|\zeta_{K_0+k}\| \\
&= 2^q \lambda_q^{1+q} E^{q^2}_{K_0+k-1} + 2^q \lambda_q \|\zeta_{K_0+k-1}\|^q + \|\zeta_{K_0+k}\| \\
&\le 2^q \lambda_q^{1+q} \left(\lambda_q E^q_{K_0+k-2} + \|\zeta_{K_0+k-2}\|\right)^{q^2} + 2^q \lambda_q \|\zeta_{K_0+k-1}\|^q + \|\zeta_{K_0+k}\| \\
&\vdots \\
&\le 2^{s(k+1)-1} \lambda_q^{s(k+1)} E^{q^{k+1}}_{K_0} + \sum_{j=0}^{k} \|\zeta_{K_0+j}\|^{q^{k-j}} \lambda_q^{s(k-j)} 2^{s(k-j+1)-1} \\
&\le 2^{s(k+1)-1} \lambda_q^{s(k+1)} E^{q^{k+1}}_{K_0} + \sum_{j=K_0}^{k+K_0} \|\zeta_j\|^{q^{k+K_0-j}} \lambda_q^{s(k+K_0-j)} 2^{s(k+K_0-j+1)-1} \\
&\le 2^{s(k+1)-1} \lambda_q^{s(k+1)} E^{q^{k+1}}_{K_0} + \sum_{j=0}^{k+K_0} \|\zeta_j\|^{q^{k+K_0-j}} \lambda_q^{s(k+K_0-j)} 2^{s(k+K_0-j+1)-1},
\end{aligned} \tag{6.37}$$

where the second to last inequality in (6.37) is after relabeling $j \equiv j + K_0$ in the inequality immediately preceding it, and the last inequality is obtained by adding some positive terms to the right-hand side of (6.37).

Now relabel $k \equiv k + K_0$ in (6.37) and notice that $E_{K_0} \le \Delta$ by the definition of K_0 to get, for $k \ge K_0$,

$$\begin{aligned}
E_{k+1} &\le 2^{s(k-K_0+1)-1} \lambda_q^{s(k-K_0+1)} \Delta^{q^{k-K_0+1}} + \sum_{j=0}^{k} \|\zeta_j\|^{q^{k-j}} \lambda_q^{s(k-j)} 2^{s(k-j+1)-1} \\
&\le (2\lambda_q)^{s(k-K_0+1)} \Delta^{q^{k-K_0+1}} + (2\lambda_q^{\frac{1}{q}})^{-1} \sum_{j=0}^{k} \|\zeta_j\|^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)}.
\end{aligned} \tag{6.38}$$

Now, we see from (6.35) (after taking expectation with respect to $\mathcal{F}_{j-1}$) that

$$\begin{aligned}
\mathbb{E}[\|\zeta_j\|^{q^{k-j}}] &\le \kappa^{q^{k-j}} m_j^{-\alpha q^{k-j}} (1 + \mathbb{E}[\|X_j\|]) \\
&\le \kappa^{q^{k-j}} m_j^{-\alpha q^{k-j}} (1 + \|x^*\| + \mathbb{E}[E_j]),
\end{aligned} \tag{6.39}$$

where $\kappa = \max(\kappa_0, \kappa_1)$. As in the proof of Theorem 6.6, due to Theorem 5.2, for given $\epsilon > 0$ there exists $k_0(\epsilon)$ such that for $j \geq k_0(\epsilon)$, $\mathbb{E}[E_j] \leq \epsilon$. This and (6.39) imply that

$$\mathbb{E}\left[\sum_{j=0}^{k} \|\zeta_j\|^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)}\right]$$
$$\leq \sum_{j=0}^{k_0(\epsilon)} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} (1 + \|x^*\| + \mathbb{E}[E_j]) +$$
$$\sum_{j=k_0(\epsilon)+1}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} (1 + \|x^*\| + \epsilon). \quad (6.40)$$

Since $\mathbb{E}[E_j] < \infty$ for $j \leq k_0(\epsilon)$, $e^* = \max(\max\{\mathbb{E}[E_j] : j = 1, 2, \ldots, k_0(\epsilon)\}, \epsilon.) < \infty$. The inequality in (6.40) then implies that

$$\mathbb{E}\left[\sum_{j=0}^{k} \|\zeta_j\|^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)}\right] \leq (1 + \|x^*\| + e^*) \sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}. \quad (6.41)$$

Again, as in the proof of Theorem 6.6, we know from part (i) of Lemma 6.5 that if a positive random sequence $\{S_n\}$ satisfies $\mathbb{E}[S_n] = \mathcal{O}(a_n)$, where $\{a_n\}$ is a deterministic positive-valued sequence, then $S_n = \mathcal{O}_p(a_n)$. Therefore we see from (6.41) that

$$\sum_{j=0}^{k} \|\zeta_j\|^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} = \mathcal{O}_p\left(\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}\right). \quad (6.42)$$

Use (6.42) and (6.38) to write, for $k \geq K_0$,

$$E_{k+1} \leq 2^{s(k-K_0+1)-1} \lambda_q^{s(k-K_0+1)} \Delta^{q^{k-K_0+1}} + \mathcal{O}_p\left(\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}\right)$$
$$\leq (2\lambda_q)^{-\frac{1}{q-1}-1} \left(\Lambda(K_0)\right)^{q^k} + \mathcal{O}_p\left(\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}\right), \quad (6.43)$$

where, after some algebra, the random variable $\Lambda(K_0)$ in (6.43) can be seen to be

$$\Lambda(K_0) = \left((2\lambda_q)^{\frac{1}{q-1}} \Delta\right)^{q^{-K_0+1}}. \quad (6.44)$$

(The constant $\Lambda(K_0) \in (0, 1)$ because Δ has been chosen so that $\Delta < (2\lambda_q)^{\frac{-1}{q-1}}$.)
Proof of (i). In what follows, the assertions in (i) will be proved using conclusions from three parts named Part A, Part B, and Part C that follow. In Part A, we will analyze the behavior of the summation $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ appearing in (6.43) when the sample size sequence $\{m_j\}$ is Polynomial(λ_p, p), or Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$. In Part B, we will analyze the behavior of the

summation $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ appearing in (6.43) when the sample size sequence $\{m_j\}$ is SupExponential(λ_t, t) with $t \geq q$. In Part C, we will analyze the behavior of the term $(2\lambda_q)^{-\frac{1}{q-1}-1} \left(\Lambda(K_0)\right)^{q^k}$ appearing in (6.43). The conclusions from Part A, Part B, and Part C will be "put together" in Part D to prove the assertion in (i).

Part A. When the sample size sequence $\{m_j\}$ is Polynomial(λ_p, p), or Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$, we will show that the two postulates of Lemma 6.4 hold for the sum $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$, thereby proving that $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ is of the same order as the last summand $2\kappa\lambda_q^{\frac{1}{q}} m_k^{-\alpha}$, that is, $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} = \mathcal{O}\left(2\kappa\lambda_q^{\frac{1}{q}} m_k^{-\alpha}\right)$. Towards this, set $a_j(k) = \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ and we have

$$\frac{a_{j+1}(k)}{a_j(k)} = \left(\kappa^{\frac{1}{q}-1} (2\lambda_q^{\frac{1}{q}})^{-1} \left(\frac{m_j^q}{m_{j+1}}\right)^{\frac{\alpha}{q}}\right)^{q^{k-j-1}}. \tag{6.45}$$

If $\{m_j\}$ grows as Polynomial(λ_p, p), Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$, some algebra yields that $a_{j+1}(k)/a_j(k) > \beta$ for any $\beta \in (0, \infty)$ and large-enough j. Thus, the first postulate of Lemma 6.4 is satisfied when $\{m_j\}$ is Polynomial(λ_p, p), Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$. Also, since

$$\frac{a_j(k)}{a_k(k)} = \left(\kappa \, (2\lambda_q^{\frac{1}{q}})^{\frac{q}{q-1}}\right)^{q^{k-j}-1} \left(\frac{m_k}{m_j^{q^{k-j}}}\right)^{\alpha},$$

some algebra again yields that $\limsup_k a_j(k)/a_k(k) \to 0$ for j lying in any fixed interval for the three cases Polynomial(λ_p, p), Geometric(c), and SupExponential(λ_t, t) with $t \in (1, q)$, and hence the second postulate of Lemma 6.4 is satisfied as well.

We thus conclude that when the sample size sequence $\{m_j\}$ is Polynomial(λ_p, p), or Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$,

$$\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} = \mathcal{O}\left(2\kappa\lambda_q^{\frac{1}{q}} m_k^{-\alpha}\right). \tag{6.46}$$

Part B. When the sample size sequence $\{m_j\}$ is SupExponential(λ_t, t) with $t \geq q$, the sum $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ appearing in (6.43) turns out to be of the order of the first summand $\kappa^{q^k} (2\lambda_q^{\frac{1}{q}})^{s(k+1)} m_0^{-\alpha q^k}$. To prove this, we write $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ in reverse order as

$$\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} = \sum_{j=0}^{k} \kappa^{q^j} (2\lambda_q^{\frac{1}{q}})^{s(j+1)} m_{k-j}^{-\alpha q^j} \tag{6.47}$$

and again apply Lemma 6.4. Set the jth summand on the right-hand side of (6.47) to $a_j(k)$ and obtain

$$\frac{a_{j+1}(k)}{a_j(k)} = \left(\kappa^{q-1} (2\lambda_q^{\frac{1}{q}})^{q} \left(\frac{m_{k-j-1}^q}{m_{k-j}}\right)^{\alpha}\right)^{q^j}. \tag{6.48}$$

Then some algebra as in (6.45) yields that if $\{m_j\}$ grows as SupExponential(λ_t, t) with $t \geq q$, then $a_{j+1}(k)/a_j(k) > \beta$ for any $\beta \in (0,\infty)$ and large enough j. Thus, the first postulate of Lemma 6.4 is satisfied when $\{m_j\}$ is SupExponential(λ_t, t) with $t \geq q$. Also, since

$$\frac{a_j(k)}{a_k(k)} = \left(\kappa(2\lambda_q^{\frac{1}{q}})^{\frac{q}{q-1}}\right)^{q^j - q^k} (\frac{m_0^{q^k}}{m_{k-j}^{q^j}})^{\alpha},$$

some algebra again yields that $\limsup_k a_j(k)/a_k(k) \to 0$ for j lying in any fixed interval when $\{m_k\}$ is SupExponential(λ_t, t) with $t \geq q$, and hence the second postulate of Lemma 6.4 is satisfied as well. We thus see that the sum $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ appearing in (6.43) satisfies $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} = \mathcal{O}(\kappa^{q^k} (2\lambda_q^{\frac{1}{q}})^{\frac{q^{k+1}-1}{q-1}} m_0^{-\alpha q^k}) = \mathcal{O}\left(c_2^{-\alpha q^k}\right)$, where $c_2 = \kappa^{-\frac{1}{\alpha}} (2\lambda_q^{\frac{1}{q}})^{-\frac{q}{\alpha(q-1)}} m_0$.

We thus conclude that when the sample size sequence $\{m_j\}$ is SupExponential(λ_t, t) with $t \geq q$ the sum $\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}}$ appearing in (6.43) satisfies

$$\sum_{j=0}^{k} \kappa^{q^{k-j}} (2\lambda_q^{\frac{1}{q}})^{s(k-j+1)} m_j^{-\alpha q^{k-j}} = \mathcal{O}(\kappa^{q^k} (2\lambda_q^{\frac{1}{q}})^{\frac{q^{k+1}-1}{q-1}} m_0^{-\alpha q^k}) = \mathcal{O}\left(c_2^{-\alpha q^k}\right). \quad (6.49)$$

Part C. Let's analyze the behavior of the term $(2\lambda_q)^{-\frac{1}{q-1}-1} (\Lambda(K_0))^{q^k}$ appearing in (6.43). We recall that if $\{m_j\}$ is Polynomial(λ_p, p), then $m_j = \lambda_p j^p$ where $\lambda_p, p \in (0,\infty)$; if $\{m_j\}$ is Geometric(c), then $m_j = m_0 c^j$ where $c \in (1,\infty)$; and if $\{m_j\}$ is SupExponential(λ_t, t), then $m_j = \lambda_t^{\frac{t^j-1}{t-1}} m_0^{t^j}$ where $\lambda_t, t \in (1,\infty)$. (See Definition 2.1.) We thus see that if the sample size sequence $\{m_j\}$ is Polynomial(λ_p, p), or Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$, then $m_j^{\alpha/q^j} \to 1$ as $j \to \infty$. Thus, when $\{m_j\}$ is Polynomial(λ_p, p), or Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$, we can invoke Lemma 6.7 with $a_j \equiv m_j^{\alpha}$ and $\Lambda \equiv \Lambda(K_0) \in (0, 1)$ to see that the term $(\Lambda(K_0))^{q^k}$ appearing in (6.43) satisfies, as $k \to \infty$,

$$m_k^{\alpha} (\Lambda(K_0))^{q^k} \xrightarrow{\text{p}} 0. \quad (6.50)$$

(It is important that the assertion in (6.50) is not true when $\{m_j\}$ is SupExponential(λ_t, t) with $t \geq q$.)

Part D. Finally, we use (6.46) and (6.50) in (6.43) to see that when the sample size sequence $\{m_j\}$ is Polynomial(λ_p, p), or Geometric(c), or SupExponential(λ_t, t) with $t \in (1, q)$, for $k \geq K_0$,

$$E_{k+1} \leq \mathcal{O}_p\left(2\kappa\lambda_q^{\frac{1}{q}} m_k^{-\alpha}\right). \quad (6.51)$$

Now use part (ii) of Lemma 6.5 and the expressions for m_k to conclude that the first three assertions in (i) hold.

Similarly, use (6.49) in (6.43) when $\{m_j\}$ is SupExponential(λ_t, t) with $t \geq q$, to see that for $k \geq K_0$,

$$E_{k+1} \leq \mathcal{O}_p\left(\Lambda^{q^k} + c_2^{-\alpha\, q^k}\right). \quad (6.52)$$

Now use part (ii) of Lemma 6.5 to conclude that the last assertion in (i) holds as well.

Proof of (ii). To prove the assertion in (ii), we recall that since $W_k = \sum_{j=1}^{k} m_j$, and the expressions in (6.33) hold here as well. Now use (6.33) in assertion (i) to obtain the assertion in (ii).
□

Theorem 6.8 is the analogue to Theorem 6.6 but for superlinearly converging DA recursions. Like Theorem 6.6, Theorem 6.8 demonstrates that there are two well-defined sampling regimes corresponding to predominant sampling and recursion errors. Assertion (i) in Theorem 6.8 implies that all of polynomial sampling, all of geometric sampling, and part of super-exponential sampling result in predominant sampling error. This makes intuitive sense because the rapidly decaying recursion error demands (or allows) additional sampling to drive the sampling error down at the same rate as the recursion error. Correspondingly, as (ii) in Theorem 6.8 implies, there is a larger range of sampling rates that result in efficiency; in some sense, this is the advantage of using a faster DA recursion. Specifically, (ii) in Theorem 6.8 implies that any Geometric(c) ($c \in (1, \infty)$) sampling is efficient, and, SupExponential(λ_t, t) sampling results in efficiency as long as $t \in (1, q)$.

We next prove a result analogous to Theorems 6.6 and 6.8 but for the case where the DA recursion exhibits SubLinear(s) convergence.

THEOREM 6.9. *(SubLinearly Converging DA) Let Assumption 3.1 and Assumption 3.2 hold. Also, suppose the deterministic recursion (DA) satisfies the sub-linear decrease condition for each k, that is, for all x, k and some $s \in (0, 1)$,*

$$\|x + h_k(x) - x^*\| \leq (1 - \frac{s}{k})\|x - x^*\|.$$

Then, as $k \to \infty$, the following hold.
(i)

$$E_k = \begin{cases} \mathcal{O}_p(k^{-p\alpha+1}), & \textit{if } \{m_k\} \textit{ grows as Polynomial}(\lambda_p, p), p\alpha \in (1, s+1); \\ \mathcal{O}_p(k^{-s}), & \textit{if } \{m_k\} \textit{ grows as Polynomial}(\lambda_p, p), p\alpha \geq s+1; \\ \mathcal{O}_p(k^{-s}), & \textit{if } \{m_k\} \textit{ grows as Geometric}(c); \\ \mathcal{O}_p(k^{-s}), & \textit{if } \{m_k\} \textit{ grows as SupExponential}(\lambda_t, t). \end{cases}$$

(ii)

$$\begin{aligned} W_k^{\alpha\frac{1-1/(p\alpha)}{1+1/p}} E_k &= \mathcal{O}_p(1), && \textit{if } \{m_k\} \textit{ grows as Polynomial}(\lambda_p, p), p\alpha \in (1, s+1); \\ W_k^{\alpha\frac{s}{p\alpha+\alpha}} E_k &= \mathcal{O}_p(1), && \textit{if } \{m_k\} \textit{ grows as Polynomial}(\lambda_p, p), p\alpha \geq s+1; \\ (\log_c W_k)^s E_k &= \mathcal{O}_p(1), && \textit{if } \{m_k\} \textit{ grows as Geometric}(c); \\ \left(\log_t(\log_{c_1} W_k)\right)^s E_k &= \mathcal{O}_p(1), && \textit{if } \{m_k\} \textit{ grows as SupExponential}(\lambda_t, t); \end{aligned}$$

where $c_1 = m_0 \lambda_t^{1/(t-1)}$.

Proof. Recalling the notation $\zeta_k = H_k(m_k, X_k) - h_k(X_k)$, we write

$$\begin{aligned} E_{k+1} &\leq \|X_k - x^* + h(X_k)\| + \|\zeta_k\| \\ &\leq E_k(1 - \frac{s}{k}) + \|\zeta_k\|. \end{aligned} \tag{6.53}$$

We then recurse (6.53) to obtain

$$E_{k+1} \leq E_1 \left(\prod_{j=1}^{k} (1 - \frac{s}{j}) \right) + \sum_{j=1}^{k-1} \|\zeta_j\| \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + \|\zeta_k\|. \tag{6.54}$$

Following the same argument we used in the proofs of Theorem 6.6 and Theorem 6.8, we see that due to Theorem 5.2, there exists $k_0(\epsilon)$ such that for all $k \geq k_0(\epsilon)$, $\mathbb{E}[E_k] \leq \epsilon$. Combining this with Assumption 3.2 implies that for $j \geq k_0(\epsilon)$,

$$\begin{aligned} \mathbb{E}[\|\zeta_j\|] &= \mathbb{E}\left[\mathbb{E}\left[\|\zeta_j\| \mid \mathcal{F}_{j-1}\right]\right] \\ &\leq m_j^{-\alpha}(\kappa_0 + \kappa_1 \mathbb{E}[\|X_j\|]) \\ &\leq m_j^{-\alpha}(\kappa_0 + \kappa_1 \|x^*\| + \kappa_1\, \epsilon). \end{aligned} \tag{6.55}$$

We can then write for $k > k_0(\epsilon)$,

$$\begin{aligned} &\mathbb{E}\left[\sum_{j=1}^{k-1} \|\zeta_j\| \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + \|\zeta_k\| \right] \\ &\quad = \sum_{j=1}^{k_0(\epsilon)} \mathbb{E}[\|\zeta_j\|] \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + \sum_{j=k_0(\epsilon)+1}^{k-1} \mathbb{E}[\|\zeta_j\|] \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + \mathbb{E}[\|\zeta_k\|] \\ &\quad \leq \left(\sum_{j=1}^{k_0(\epsilon)} \mathbb{E}[\|\zeta_j\|] \right) \left(\prod_{i=k_0(\epsilon)+1}^{k} (1 - \frac{s}{i}) \right) + \sum_{j=k_0(\epsilon)+1}^{k-1} \mathbb{E}[\|\zeta_j\|] \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + \mathbb{E}[\|\zeta_k\|] \\ &\quad \leq \left(\sum_{j=1}^{k_0(\epsilon)} \mathbb{E}[\|\zeta_j\|] \right) \left(\prod_{i=k_0(\epsilon)+1}^{k} (1 - \frac{s}{i}) \right) + \\ &\qquad\qquad (\kappa_0 + \kappa_1 \|x^*\| + \kappa_1 \epsilon) \left(\sum_{j=k_0(\epsilon)+1}^{k-1} m_j^{-\alpha} \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + m_k^{-\alpha} \right) \\ &\quad = \mathcal{O}\left(\sum_{j=1}^{k-1} m_j^{-\alpha} \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + m_k^{-\alpha} \right), \end{aligned} \tag{6.56}$$

where the first inequality in (6.56) follows since $(1 - s/i) \in (0, 1)$ for all $i \geq 1$, the second inequality follows from the application of (6.55), and the last equality follows since $k_0(\epsilon) < \infty$ and $\mathbb{E}[\|\zeta_j\|] < \infty$ for $j \leq k_0(\epsilon)$.

As in the proof of Theorem 6.6, we know from part (i) of Lemma 6.5 that if a positive random sequence $\{S_n\}$ satisfies $\mathbb{E}[S_n] = \mathcal{O}(a_n)$, where $\{a_n\}$ is a deterministic positive-valued sequence, then $S_n = \mathcal{O}_p(a_n)$. Therefore we see from (6.56) that

$$\sum_{j=1}^{k-1} \|\zeta_j\| \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + \|\zeta_k\| = \mathcal{O}_p\left(\sum_{j=1}^{k-1} m_j^{-\alpha} \prod_{i=j+1}^{k} (1 - \frac{s}{i}) + m_k^{-\alpha} \right). \tag{6.57}$$

Combining (6.54) and (6.57), we can write

$$E_{k+1} \leq E_1 \left(\prod_{j=1}^{k} (1 - \frac{s}{j}) \right) + \mathcal{O}_p \left(\sum_{j=1}^{k-1} m_j^{-\alpha} \prod_{i=j+1}^{k} (1 - \frac{s}{i}) \right)$$
$$= \mathcal{O}_p(k^{-s}) + k^{-s} \mathcal{O}_p \left(\sum_{j=1}^{k} m_j^{-\alpha} (j+1)^s \right), \tag{6.58}$$

where the equality in (6.58) follows upon noticing that $\prod_{i=j+1}^{k}(1 - \frac{s}{i})) = \Theta((\frac{j+1}{k})^s)$.

If $\{m_k\}$ grows as Polynomial(λ_p, p), as $k \to \infty$,

$$\sum_{j=1}^{k} m_j^{-\alpha}(j+1)^s = \sum_{j=1}^{k} \lambda_p^{-\alpha} (\frac{j}{j+1})^{-p\alpha} (j+1)^{-p\alpha+s} = \Theta(k^{-p\alpha+s+1}). \tag{6.59}$$

If $\{m_k\}$ grows as Geometric(c), as $k \to \infty$,

$$\sum_{j=1}^{k} m_j^{-\alpha}(j+1)^s = \sum_{j=1}^{k} m_0 c^{-j\alpha}(j+1)^s = o(1). \tag{6.60}$$

If $\{m_k\}$ grows as SupExponential(λ_t, t), as $k \to \infty$,

$$\sum_{j=1}^{k} m_j^{-\alpha}(j+1)^s = \sum_{j=1}^{k} m_0^{-\alpha t^j} \lambda_t^{\alpha/(t-1)} \lambda_t^{-\alpha t^j/(t-1)} (j+1)^s = o(1). \tag{6.61}$$

Now use (6.59), (6.60), and (6.61) in (6.58) to see that the assertion in (i) holds.
Proof of (ii). To prove the assertion in (ii), we again notice that since $W_{k+1} = \sum_{j=1}^{k} m_k$, and the expressions in (6.33). Now use (6.33) in assertion (i) to obtain the assertion in (ii). □

The context of Theorem 6.9, sublinearly converging DA recursions, is diametrically opposite to that of Theorem 6.8 in the sense that most sampling regimes result in predominantly recursion error. Specifically, (i) in Theorem 6.9 implies that all geometric, all super-exponential, and a portion ($p\alpha \geq s+1$) of polynomial sampling result in dominant recursion error. Perhaps more importantly, (ii) in Theorem 6.9 implies that there is no sampling regime that results in efficiency when the DA recursion exhibits sublinear convergence. The best achievable rate under these conditions is $W_k^{-\alpha\eta^*}$ where $\eta^* = s/(s+1+\alpha)$, obtained through Polynomial(λ_p, p) sampling for $p = (s+1)\alpha^{-1}$.

We will end this section with a finite-time bound on the mean absolute deviation of SCSR's iterates from the solution, assuming that the DA recursion exhibits a linear contraction at every step, and not just asymptotically. This context occurs when optimizing a strongly convex function using a linearly converging DA recursion. The utility of Theorem 6.10 is clear — to identify the number of steps of the algorithm required to achieve a mean absolute deviation that is below a specified threshold. Analogous crude bounds for stochastic approximation have appeared recently [21].

THEOREM 6.10. *(Finite-time bounds.) Let Assumption 3.1 and Assumption 3.2 hold. Also, let the deterministic recursion (DA) be such that*

$$\|x_{k+1} - x^*\| \leq \ell \|x_k - x^*\|, \quad k = 0, 1, \dots \tag{6.62}$$

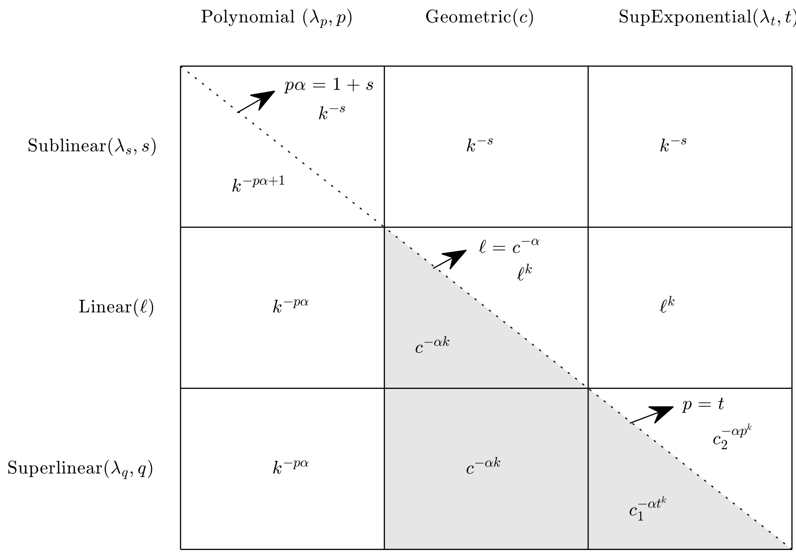

FIG. 6.1. *A summary of the error rates achieved by various combinations of recursion quality and sampling rates. Each row corresponds to a recursion rate while each column corresponds to a sampling rate, and k denotes the iteration number. The entries refer to the error decay rates resulting from corresponding choices of the sampling and recursion rates. For example, when using a linearly converging recursion with Polynomial(λ_p, p) sampling, the table shows that the resulting error $E_k = \|X_k - x^*\|$ in the iterates satisfies $E_k = \mathcal{O}_p(k^{-p\alpha})$. The combinations lying below the dotted line have dominant sampling error since the corresponding entries involve only sampling related constants. Likewise, those above dotted line have dominant recursive error due to the presence of only recursion related constants. The combinations in the shaded region are efficient in the sense that they result in the fastest possible convergence rates as measured in terms of the total simulation work done. Notice that faster recursions afford a wider range of sampling rates that are efficient. For example, it can be seen that all geometric sampling rates yield efficiency when using a superlinear DA recursion. By contrast, only geometric sampling rates with $c < \ell^{-1/\alpha}$ yield efficiency when using a linear DA recursion. Sublinearly converging recursions yield no sampling regimes that are efficient.*

for some $\ell \in (0,1)$. Then, for some constant $\sigma > 0$,

$$\mathbb{E}[\|X_{k+1} - x^*\|] \leq \ell^{k+1}\mathbb{E}[\|X_0 - x^*\|] + \sigma \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}. \tag{6.63}$$

Furthermore, suppose the sample size sequence $\{m_k\}$ increases as Geometric(c). Then, in the efficient regime, that is, when $\ell c^{\alpha} < 1$, the following bound on the mean absolute deviation of the solution X_{k+1} holds for $k = 0, 1, \ldots$.

$$\mathbb{E}[\|X_{k+1} - x^*\|] \leq \ell^{k+1}\left(\mathbb{E}[\|X_0 - x^*\|] + \frac{\sigma}{\ell m_0^{\alpha}}\left(\frac{1 - (\ell c^{\alpha})^{k+1}}{1 - (\ell c^{\alpha})}\right)\right). \tag{6.64}$$

Proof. Since Theorem 5.2 holds, there exists $k_0(\epsilon) < \infty$ such that for all $k \geq k_0(\epsilon)$, $\mathbb{E}[E_k] \leq \epsilon$. Also, $\mathbb{E}[E_j] < \infty$ for $j \leq k_0(\epsilon) < \infty$. Thefore, we see that for all k,

$$\mathbb{E}[E_k] \leq e^* := \max(\max\{\mathbb{E}[E_j] : j = 1, 2, \ldots, k_0(\epsilon)\}, \epsilon.) < \infty.$$

Using this and Assumption 3.2, we see that $\zeta_k = H_k(m_k, X_k) - h_k(X_k)$ satisfies

$$
\begin{aligned}
\mathbb{E}\left[\|\zeta_k\|\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\zeta_k\| \mid \mathcal{F}_{k-1}\right]\right] \\
&\le m_k^{-\alpha}(\kappa_0 + \kappa_1 \mathbb{E}[\|X_k\|]) \\
&\le m_k^{-\alpha}(\kappa_0 + \kappa_1 \|x^*\| + \kappa_1 \mathbb{E}[E_k]) \\
&\le m_k^{-\alpha}(\kappa_0 + \kappa_1 \|x^*\| + \kappa_1 e^*). \qquad (6.65)
\end{aligned}
$$

Setting $\sigma := \kappa_0 + \kappa_1\|x^*\| + \kappa_1 e^*$, and using (6.65) along with (6.62), we get for $k \ge 0$,

$$
\begin{aligned}
\mathbb{E}[\|X_{k+1} - x^*\|] &= \mathbb{E}[\|X_k - x^* + h_k(X_k)\|] + \mathbb{E}[\|\zeta_k\|] \\
&\le \ell\, \mathbb{E}[\|X_k - x^*\|] + \frac{\sigma}{m_k^\alpha} \\
&\le \ell^2 \mathbb{E}[\|X_{k-1} - x^*\|] + \ell \frac{\sigma}{m_{k-1}^\alpha} + \frac{\sigma}{m_k^\alpha} \\
&\;\;\vdots \\
&= \ell^{k+1}\|X_0 - x^*\| + \sigma \sum_{j=0}^{k} \ell^{k-j} m_j^{-\alpha}. \qquad (6.66)
\end{aligned}
$$

The first assertion (6.63) of the theorem is thus proved.

To prove the assertion in (6.64), use $m_{k+1} = c\, m_k, k = 0, 1, \ldots$ in (6.63).

□

The utility of Theorem 6.10 is limited due to the fact that we have assumed that a linear decrease happens at every step. Results analogous to Theorem 6.10 but for other types of DA recursions should be obtainable in a similar fashion.

7. Concluding Remarks. The use of simulation-based estimators within well-established algorithmic recursions is becoming an attractive paradigm to solve optimization and root-finding problems in contexts where the underlying functions can only be estimated. In such contexts, the question of how much to simulate (towards estimating function and derivative values at any point) becomes important particularly when the available simulations are computationally expensive. In this paper, we have argued that there is an interplay between the structural error inherent in the recursion in use and the sampling error inherent in the simulation estimator. Our characterization of this interplay provides guidance (see Figure 6.1) on how much sampling should be undertaken under various recursive contexts in order to ensure that the resulting iterates are provably efficient.

A few other comments relating to the results we have presented and about ongoing research are now in order.

1. All of the results we have presented assume that the sequence of sample sizes $\{m_k\}$ used within SCSR is deterministic. To this extent, our results provide guidance on only the *rate* at which sampling should be performed in order that SCSR's iterates remain efficient. We envision an implementable algorithm dynamically choosing sample sizes as a function of the observed trajectory of the algorithm while ensuring that the increase rates prescribed by our results are followed. Our ongoing research attempts such a strategy within the context of derivative-free optimization [12, 10].
2. None of the results we have presented are of the "central limit theorem" type; they are cruder and of the $O_p(\cdot)$ type. This is because, when the sampling

and recursion choices lie off the diagonal in Figure 6.1, either the recursion error or the sampling error are dominant and consequently lead to a situation where the contribution to the error in the SCSR iterates is due only to a few terms. When the sampling and recursion choices lie on the diagonal, a central-limit theorem will likely hold but characterizing such a fine result will involve further detailed assumptions on the convergence characteristics of the deterministic recursion.

3. Another interesting question that we have not treated here is that of iterate averaging [30] to increase efficiency. Recall that our results suggest that efficiency cannot be achieved for sampling regimes slower than geometric. it is possible that iterate averaging might be useful in such sub-geometric low-sampling regimes, e.g., polynomial. It also seems that such averaging is of less value in high sampling regimes for the same reason that CLT-type results do not take hold due to the Lindberg-Feller condition [31] failing on constituent sums.

Acknowledgements. The first author thanks the support provided by Office of Naval Research contracts N000141110419, N00014-11-1-0065 and National Science Foundation grant CMMI 478709. He is also grateful for the financial and logistics support provided by IBM Research, Yorktown Heights, NY where he spent his sabbatical year 2012–2013. We thank Bruce Schmeiser, Purdue University for interesting discussions that have influenced the ideas appearing in this paper.

References.

[1] O. Alagoz, A. J. Schaefer, and M. S. Roberts. Optimization in organ allocation. In P. Pardalos and E. Romeijn, editors, *Handbook of Optimization in Medicine*. Kluwer Academic Publishers, 2009.

[2] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York, NY., 2007.

[3] J. Atlason, M. A. Epelman, and S. G. Henderson. Optimizing call center staffing using simulation and analytic center cutting plane methods. *Management Science*, 54(2):295–309, 2008.

[4] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Nashua, NH., 1995.

[5] P. Billingsley. *Probability and Measure*. Wiley, New York, NY., 1995.

[6] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, Cambridge, UK, 2008.

[7] L. Bottou and O. Bosquet. The trade-offs of large scale learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

[8] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. 2016.

[9] S. Boyd. *Convex Optimization*. Cambridge University Press, Cambridge, U.K., 2004.

[10] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample size selection for optimization methods for machine learning. *Mathematical Programming, Series B*, 134:127–155, 2012.

[11] M. Chu, Y. Zinchenko, S. G. Henderson, and M. B. Sharpe. Robust optimization for intensity modulated radiation therapy treatment planning under uncertainty. *Physics in Medicine and Biology*, 50:5463–5477, 2005.

[12] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, Philadelphia, PA, 2009.

[13] G. Deng and M. C. Ferris. Adaptation of the UOBQYA algorithm for noisy functions. In L.F. Perrone, F.P. Wieland, J. Liu, B.G. Lawson, D.M. Nicol, and R.M. Fujimoto, editors, *Proceedings of the 2006 Winter Simulation Conference*, pages 312–319. Institute of Electrical and Electronics Engineers: Piscataway, New Jersey, 2006.

[14] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

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

[16] S. G. Henderson and B. L. Nelson, editors. volume 13 of *Handbooks in Operations Research and Management Science: Simulation*. Elsevier, 2006.

[17] Y. T. Herer, , M. Tzur, and E. Yucesan. The multilocation transshipment problem. *IIE Transactions*, 38:185–200, 2006.

[18] T. Homem-de-Mello, A. Shapiro, and M. L. Spearman. Finding optimal release times using simulation based optimization. *Management Science*, 45:86–102, 1999.

[19] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

[20] A. M. Law. *Simulation Modeling and Analysis*. McGraw-Hill, New York, NY., 2007.

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

[22] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science + Business Media, LLC, New York, NY, 2004.

[23] C. Osorio and M. Bierlaire. A simulation-based optimization problem for urban transportation networks. *Operations Research*, 2014. To appear.

[24] R. Pasupathy. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58:889–901, 2010.

[25] R. Pasupathy and S. Ghosh. Simulation optimization: A concise overview and implementation guide. INFORMS TutORials. INFORMS, 2013.

[26] R. Pasupathy and S. Kim. The stochastic root-finding problem: overview, solutions, and open questions. *ACM TOMACS*, 21(3), 2011.

[27] R. Pasupathy and B. W. Schmeiser. Retrospective-approximation algorithms for multidimensional stochastic root-finding problems. *ACM TOMACS*, 19(2):5:1–5:36, 2009.

[28] R. Pasupathy and B. W. Schmeiser. DARTS — dynamic adaptive random target shooting. In B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, editors, *Proceedings of the 2010 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers: Piscataway, New Jersey, 2010.

[29] B. Polyak. *Introduction to Optimization*. Optimization Software In, New York, NY, 1987.

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

[31] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., New York, NY., 1980.