Noname manuscript No.
(will be inserted by the editor)

The Adaptive Sampling Gradient Method

Optimizing Smooth Functions with an Inexact Oracle

Fatemeh S. Hashemi · Raghu Pasupathy · Michael R. Taaffe

Received: date / Accepted: date

Abstract Consider settings such as stochastic optimization where a smooth objective function f is unknown but can be estimated with an *inexact oracle* such as quasi-Monte Carlo (QMC) or numerical quadrature. The inexact oracle is assumed to yield function estimates having error that decays with increasing oracle effort. For solving such problems, we present the Adaptive Sampling Gradient Method (ASGM) in two flavors depending on whether the step size used within ASGM is constant or determined through a backtracking line search. ASGM's salient feature is the adaptive manner in which it constructs gradient estimates (henceforth called *gradient approximates*), by exerting just enough oracle effort at each iterate to keep the error in the gradient approximate within a constant factor of the norm of the gradient approximate. ASGM applies to both derivative-based and derivative-free contexts, and generates iterates that globally convergence to a first-order critical point. We also prove two sets of results on ASGM's *work complexity* with respect to the gradient norm: (i) when f is non-convex, ASGM's work complexity is arbitrarily close to $\mathcal{O}(\epsilon^{-2-\frac{1}{\mu(\alpha)}})$, where $\mu(\alpha)$ is the error decay rate of the gradient approximate expressed in terms of the error decay rate α of the objective function approximate; (ii) when f is strongly convex, ASGM's work complexity is arbitrarily close to $\mathcal{O}(\epsilon^{-\frac{1}{\mu(\alpha)}})$. We compare these complexities to those obtained from methods that use traditional random sampling. We also illustrate the calculation of α and $\mu(\alpha)$ for common choices, e.g., QMC with finite-difference derivatives.

Keywords Adaptive Sampling · Stochastic Gradient · Stochastic Optimization

F. Hashemi
The Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA
E-mail: fatemeh.s.hashemi@gmail.com

R. Pasupathy
Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
E-mail: pasupath@purdue.edu

M. R. Taaffe
The Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA
E-mail: taaffe@vt.edu

1 INTRODUCTION

Consider unconstrained optimization problems having the form

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}}\ f(\mathbf{x}) \tag{P}$$

where the function $f : \mathbb{R}^d \to \mathbb{R}$ is bounded from below, that is, $\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$, and f belongs to the class $C_L^{1,1}$ of differentiable functions having a Lipschitz continuous first derivative. An important premise is that we do not have access to exact oracle values $f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$ but instead, we have access to an *inexact oracle* using which an approximate value $f(n;\mathbf{x})$ of $f(\mathbf{x})$ at any chosen point $\mathbf{x} \in \mathbb{R}^d$ can be obtained after expending a chosen amount n of oracle effort.

The inexact oracle is such that the resulting approximator $f(n;\cdot)$ is *consistent*, that is, $|f(n;\mathbf{x}) - f(\mathbf{x})| \to 0$ as $n \to \infty$. Furthermore, we assume that the error $|f(n;\mathbf{x}) - f(\mathbf{x})|$ in the approximator $f(n;\cdot)$ satisfies for $n \geq n_f$,

$$n^{\alpha}|f(n;\mathbf{x}) - f(\mathbf{x})| < \sigma_{0,f} + \sigma_{1,f}\Gamma_f^0(\mathbf{x}), \tag{FE_1}$$

where

1. $\sigma_{0,f}$ and $\sigma_{1,f}$ are real-valued unknown constants that do not depend on $\mathbf{x}$;
2. $\Gamma_f(\cdot) := 1 + \Gamma_f^0(\cdot)$ is a known positive-valued continuous function;
3. $\alpha > 0$ is a known constant called *error decay rate* of the inexact oracle $f(n;\cdot)$.

(Throughout the rest of the paper, we use the function $\Gamma_f(\cdot) := 1 + \Gamma_f^0(\cdot)$ instead of $\Gamma_f^0(\cdot)$ as a matter of convention and convenience.)

As we illustrate in Section 2 using two motivating examples, numerous stochastic optimization settings are naturally subsumed by the above problem setting. Specifically, while the objective function f in such settings is not known in closed-form, f may be suitable for approximation using quasi-Monte Carlo (QMC) [27], numerical quadrature [36, 12], or other low-discrepancy point-set methods [13, Chapter 5], allowing the construction of an approximator $f(n;\cdot)$, a scaling function $\Gamma_f(\cdot)$, and a known error decay α that together satisfy (FE_1).

The problem statement that we have just outlined resembles the now popular simulation optimization (SO) problem [31, 14, 42] and many modern machine learning settings [8] where the objective function f is an expectation. The difference between these settings and the context we consider here pertains primarily to the assumptions on the approximator $f(n;\cdot)$. In SO and machine learning settings, the objective function value $f(\mathbf{x})$ is assumed to be estimated using Monte Carlo or through random draws from a large database, implying that the resulting estimator $f(n;\mathbf{x})$ of $f(\mathbf{x})$ is *random*, and hence any error guarantees analogous to (FE_1) can only be made in a "distributional" or an "expected value sense." In the current setting, the approximate $f(n;\mathbf{x})$ is constructed using QMC or numerical integration, allowing for a deterministic error bound (FE_1), and consequently, sharper complexity results. We say more about this issue in Section 1.2 when we discuss scope and related literature.

(As an aside, there is much ongoing debate [29, 43, 22, 27] on the appropriateness of the use of QMC for high-dimensional integration. Until recently, the general consensus had been that QMC is more efficient than Monte Carlo for constructing estimators of integrals in dimensions less than about 12. However, there is little

doubt that QMC is gaining in popularity and now seems to be used routinely in much higher dimensions [4, 40].)

We emphasize that neither the error decay rate α nor the scaling function $\Gamma_f(\cdot)$ appearing in (FE_1) are unique, that is, numerous choices of $\Gamma_f(\cdot)$ and α may satisfy (FE_1). However, such non-uniqueness will not concern us, and we make no assumptions about any optimality properties of $\Gamma_f(\cdot)$ and α. Instead, we only assume that, in addition to the inexact oracle approximator $f(n;\cdot)$, we have at our disposal one possible scaling function $\Gamma_f(\cdot)$ and one possible error decay rate α that together satisfy the stipulation (FE_1). All our results will accordingly be expressed in terms of $\Gamma_f(\cdot)$ and α.

1.1 **Terminology, Notation, and Convention**

We emphasize that, while our primary motivating context is *stochastic* optimization, the approximator $f(n;\cdot)$ of the objective function $f(\cdot)$ is indeed *deterministic*. Since this issue tends to cause confusion especially among readers steeped in stochastic and simulation optimization, we have refrained from using upper case notation for function and gradient approximators throughout the paper. For the same reason, we have also limited our use of words such as "estimator" and "estimate," instead preferring the non-standard terms "approximator" and "approximate." An exception, of course, is the name of the proposed procedure "Adaptive Sampling Method" where we have used the word "sampling" with no connotation to statistical sampling.

The algorithms we present are *derivative free* by which we mean that we do not assume that the inexact oracle provides direct observations of the gradient approximate of f. Instead, when seeking an approximate for the gradient $\nabla f(\mathbf{x})$ at the point $\mathbf{x}$, we can resort to "finite differencing" of values from the ineact oracle.

We use bold font for vectors, script font for sets, lower case font for real numbers and upper case font for random variables. Hence $\{\mathbf{X}_k\}$ denotes a sequence of random vectors in $\mathbb{R}^d$ and $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ denotes a d-dimensional vector of real numbers. We use $\mathbf{e}_i \in \mathbb{R}^d$ to denote a unit vector whose ith component is 1 and whose every other component is 0, that is, $e_{ii} = 1$ and $e_{ij} = 0$ for $j \neq i$. The set $\mathcal{B}(\mathbf{x}; r) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| \leq r\}$ is the closed ball of radius $r > 0$ with center $\mathbf{x}$. For a point $\mathbf{y} := (y_1, y_2, \ldots, y_q)$, the norm $\|\mathbf{y}\|$ denotes the q-dimensional Euclidean norm $\|\mathbf{y}\| = (\sum_{i=1}^{q} y_i^2)^{\frac{1}{2}}$. We denote $(a)_+ = \max(a, 0)$ to refer to the maximum of a number $a \in \mathbb{R}$ and zero.

A sequence $\{\mathbf{y}_n\}, \mathbf{y} \in \mathbb{R}^q$ is said to *consistently* approximate $\mathbf{y} \in \mathbb{R}^q$ if $\|\mathbf{y}_n - \mathbf{y}\| \to 0$ as $n \to \infty$. For a sequence of real numbers $\{a_k\}$, we say $a_k = o(1)$ if $\lim_{k\to\infty} a_k = 0$; we say $a_k = o^{-1}(b_k)$ if $b_k = o(a_k)$. We say $a_k = \mathcal{O}(1)$ if $\{a_k\}$ is bounded, that is, there exists a constant $M > 0$ such that $|a_k| < M$ for large enough k. For sequences of positive real numbers $\{a_k\}$, $\{b_k\}$, we say that $a_k \sim b_k$ if $\lim_{k\to\infty} a_k/b_k = 1$.

We say $f \in C_L^{1,1}$ if $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and has derivative $\nabla f(\cdot)$ such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We say $f \in \mathcal{F}_L^{1,1}$ if f is convex and $f \in C_L^{1,1}$. We say $f \in \mathcal{S}_{\lambda,L}^{1,1}$ if f is twice differentiable, $f \in \mathcal{F}_L^{1,1}$, and $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq \lambda\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

1.2 **Related Literature**

As was remarked earlier, the problem we consider is closely related to what has been called the SO problem [31, 14, 42], and more recently, to optimization settings in machine learning [8, 9]. While SO appears in many flavors [14], a predominant version involves solving the optimization problem

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize:}}\ f(\mathbf{x}) := \mathbb{E}[Y(\mathbf{x})], \tag{1}$$

where the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is a mathematical expectation that cannot be observed directly but can be estimated using a Monte Carlo (MC) based inexact oracle. Specifically, $f(\mathbf{x})$ in SO is usually assumed to be estimable as the sample mean $F(m, \mathbf{x}) = m^{-1}\sum_{j=1}^{m} Y_j(\mathbf{x})$ of independent and identically distributed (iid) copies $Y_1(\mathbf{x}), Y_2(\mathbf{x}), \ldots, Y_m(\mathbf{x})$ of a random variable $Y(\mathbf{x})$ that is unbiased with respect to $f(\mathbf{x})$, and has finite second moment. Depending on the context, iid (unbiased) observations $W_1(\mathbf{x}), W_2(\mathbf{x}), \ldots$ on the gradient $\nabla f(\mathbf{x})$ may also be available.

SO has its roots in the seminal paper of Robbins and Monro [37] on the stochastic root-finding problem [37], and its immediate successor on derivative-free optimization by Kiefer and Wolfowitz [17]. The class of recursions proposed in [37] and [17], called *stochastic approximation*, has since seen six decades of development pertaining to numerous variations based on the nature of the feasible region, e.g., continuous versus integer, nature of needed solution, e.g., global versus local, incorporation of constraints, and the nature of sampling, e.g., correlated versus iid. We will not attempt a summary of this development but instead refer the interested reader to surveys [18, 19, 7, 31, 14, 2, 3, 42] that might serve as suitable "entry points" into the literature. In the more recent machine learning stream of literature that is related to the context of the current paper, the reader is especially directed to [8] for a remarkable and organized account of the vast material that has rapidly accumulated.

The error stipulation in (FE_1) and its relationship to corresponding assumptions in the SO and machine learning literature are worthy of discussion. The stipulation in (FE_1) embodies two elements that seem essential to solving problems of the type considered in this paper: (i) access to an inexact oracle that can be used to construct an approximator of the objective (or gradient) function; and (ii) an explicit or implicit characterization of the error suffered by the approximator across the feasible region. The approximator in (i) is "deterministic" when the inexact oracle is based on QMC or numerical integration as in the current context, or "stochastic" as in SO and other machine learning contexts where the inexact oracle is based on Monte Carlo or random draws from a large database. In either case, there tends to be some stipulation on the structure of the error across the feasible region. For example, in typical SO contexts, the variance $\mathbb{V}\text{ar}(Y_1(\mathbf{x})) = \mathbb{E}[\|Y_1(\mathbf{x})\|^2] - \mathbb{E}^2[\|Y_1(\mathbf{x})\|]$ is subject to a regularity condition such as

$$\mathbb{V}\text{ar}(Y_1(\mathbf{x})) < \kappa_0 + \kappa_1\|\mathbf{x}\|^2, \tag{2}$$

where κ_0, κ_1 are some positive constants. The seminal paper [17] and most of its derivative-free successors assume a condition such as (2); Robbins and Monro [37], and all its recent important successors [35, 24, 15], assume something similar but

the stipulations are on the directly observed (unbiased) gradient estimates. Also, the objective function estimate $F(m, \mathbf{x})$ in SO, being the simple sample mean of iid observations $Y_1(\mathbf{x}), Y_2(\mathbf{x}), \ldots, Y_m(\mathbf{x})$, is governed by the central limit theorem [5]:

$$\lim_{m\to\infty} m^{1/2}(F(m,\mathbf{x}) - f(\mathbf{x})) \xrightarrow{\mathrm{d}} \sqrt{\mathbb{V}\mathrm{ar}(Y_1(\mathbf{x}))}Z(0,1), \tag{3}$$

where $Z(0,1)$ is the standard normal random variable and "$\xrightarrow{\mathrm{d}}$" refers to convergence in distribution. The conditions (3) and (2) taken together are analogous to the general error bound (FE_1) assumed in the current paper.

Structural assumptions akin to (2) are also made in much of the machine learning literature. For example, one of the standing assumptions (Assumption 4.3(c)) in [8] is that

$$\mathbb{V}\mathrm{ar}(G(\mathbf{x},\xi)) := \mathbb{E}[\|G(\mathbf{x},\xi)\|^2] - \mathbb{E}^2[\|G(\mathbf{x},\xi)\|] \leq M + M_v\|\nabla f(\mathbf{x})\|^2, \tag{4}$$

where $G(\mathbf{x},\xi)$ is constructed from directly observed estimates of $\nabla f(\mathbf{x})$ satisfying certain regularity properties, and M, M_v are finite constants. Similarly, the assumptions in [9, Chapter 6] are subsumed by (4) since they amount to assuming that the constant M_v in (4) satisfies $M_v = 0$. Again, the error structure stipulation in (4) is seen as being analogous to (FE_1), but for the fixed sample size context.

It will become evident that the error structure characterization in (FE_1) is a crucial enabler of *adaptive sampling* within ASGM, by allowing judicious oracle effort determination during algorithm evolution. Such judicious oracle effort determination, along with the use of a faster estimation procedure such as QMC, is responsible for ASGM's faster convergence rates (see Table 1) compared to stochastic approximation. Efforts on constructing similar adaptive sampling algorithms but for the Monte Carlo context are currently underway [10, 34], but sharp characterizations of complexity have thus far been elusive because of the mathematical difficulties associated with analyzing sequential sampling algorithms.

1.3 **Summary of the Proposed Algorithm and Main Results**

ASGM is a *gradient search* procedure in that, during the kth iteration, a step of size β_k is taken along the direction of the approximated negative gradient, $-g(n(\mathbf{x}_k);\mathbf{x}_k)$, to yield the subsequent iterate $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k g(n(\mathbf{x}_k);\mathbf{x}_k)$. The step size β_k is either chosen to be a constant, or calculated using a backtracking line search, resulting in the *fixed step ASGM* and *backtracking line search ASGM* flavors, respectively. ASGM can accommodate both derivative-based and "derivative free" settings and no assumptions are made about the availability of direct gradient observations of the gradient function $\nabla f(\cdot)$. In the derivative-free context, for instance, ASGM accommodates the construction of the gradient approximate $g(n(\mathbf{x}_k);\mathbf{x}_k)$ at the point $\mathbf{x}_k$ by "finite differencing" the function approximator $f(n(\mathbf{x}_k),\cdot)$ at appropriately chosen points near the point $\mathbf{x}_k$.

The salient feature of ASGM is its adaptive oracle effort determination. During each iteration k, ASGM expends an amount $n(\mathbf{x}_k)$ of oracle effort for determining the step size β_k and the gradient approximate $g(n(\mathbf{x}_k);\mathbf{x}_k)$ used in constructing the subsequent iterate $\mathbf{x}_{k+1}$. How should $n(\mathbf{x}_k)$ be determined? Large values of $n(\mathbf{x}_k)$ will yield better approximates $g(n(\mathbf{x}_k);\mathbf{x}_k)$, leading to a higher quality of the

Table 1: The first two columns list upper bounds on ASGM's work complexity for function classes $C_L^{1,1}, \mathcal{F}_L^{1,1}$ and $\mathcal{S}_{\lambda,L}^{1,1}$. The quantity $\mu(\alpha)$ is the error decay rate of the gradient approximator used within ASGM, e.g., for forward-difference (FD) gradients $\mu(\alpha) = \alpha/2$, for central-difference (CD) gradients $\mu(\alpha) = 2\alpha/3$, and for direct-gradients (DG), $\mu(\alpha) = \alpha$. When QMC is used, α is arbitrarily close to 1. Also, $\mu_A(\alpha) := \min(\mu(\alpha), \frac{\alpha}{2})$, and the constant $\delta > 0$ is arbitrary. The third column lists corresponding rates for an exact oracle, and the last three columns list the complexities when random sampling is used, e.g., as in stochastic gradient descent (SGD). See [25, 21] to get a sense of the results appearing in the last three columns.

	ASGM (fixed step)	ASGM (back tracking)	Exact Oracle	SGD with Monte Carlo		
				FD	CD	DG
$C_L^{1,1}$	$\mathcal{O}(\epsilon^{-2-\frac{1}{\mu(\alpha)-\delta}})$	$\mathcal{O}(\epsilon^{-2-\frac{1}{\mu_A(\alpha)-\delta}})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-5})$	$\mathcal{O}(\epsilon^{-4})$
$\mathcal{F}_L^{1,1}$	$\mathcal{O}(\epsilon^{-2-\frac{1}{\mu(\alpha)-\delta}})$	$\mathcal{O}(\epsilon^{-2-\frac{1}{\mu_A(\alpha)-\delta}})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-5})$	$\mathcal{O}(\epsilon^{-4})$
$S_{\lambda,L}^{1,1}$	$\mathcal{O}(\epsilon^{-\frac{1}{\mu(\alpha)-\delta}})$	$\mathcal{O}(\epsilon^{-\frac{1}{\mu_A(\alpha)-\delta}})$	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$

subsequent iterate $\mathbf{x}_{k+1}$. However, large values of $n(\mathbf{x}_k)$ also translate to higher overall computational effort, leading to poorer complexity rates. ASGM hedges this trade-off by striving to exert just enough oracle effort to ensure that the error $\|g(n(\mathbf{x}_k); \mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|$ in the approximated gradient is within a fixed proportion of true gradient $\|\nabla f(\mathbf{x}_k)\|$. (Ensuring that the error $\|g(n(\mathbf{x}_k); \mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|$ is a fixed proportion of the true gradient $\|\nabla f(\mathbf{x}_k)\|$ turns out to be relatively easy in fixed step ASGM but more involved in backtracking line search ASGM.) Such adaptive oracle effort determination within ASGM also ensures that ASGM's moves result in descent steps for large enough k, that is, the difference in objective function values at the successive steps $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ is strictly decreasing when the iteration k is large enough.

Our results pertain to the performance of fixed step ASGM and backtracking line search ASGM on the three function classes $C_L^{1,1}$, $\mathcal{F}_L^{1,1}$ and $\mathcal{S}_{\lambda,L}^{1,1}$ — see Section 1.1 for a formal definition of $C_L^{1,1}$, $\mathcal{F}_L^{1,1}$, and $\mathcal{S}_{\lambda,L}^{1,1}$. For all three function classes, we demonstrate that when f is also bounded below, the iterates $\{\mathbf{x}_k\}$ generated by both flavors of ASGM are such that $\|\nabla f(\mathbf{x}_k^*)\| \to 0$. (The point $\mathbf{x}_k^*$ is the "returned solution," that is, that iterate amongst $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ where the smallest gradient norm was observed. We formally define $\mathbf{x}_k^*$ in Section 4.1.)

For characterizing ASGM's efficiency, we prove two sets of results, as summarized through Table 1. When $f \in C_L^{1,1}$ and bounded below, we demonstrate that the fixed step ASGM's total oracle effort $w_k = \sum_{j=1}^{k} n(\mathbf{x}_k) \to \infty$ satisfies $\sup_k w_k \|\nabla f(\mathbf{x}_k^*)\|^{2+\frac{1}{\mu(\alpha)-\delta}} < \infty$ for any $\delta > 0$, where $\mu(\alpha)$ is the error decay rate of the gradient approximates constructed by ASGM and expressed as a function of the error decay rate α appearing in (FE_1). The corresponding result for backtracking line search ASGM is $\sup_k w_k \|\nabla f(\mathbf{x}_k^*)\|^{2+\frac{1}{\mu_A(\alpha)-\delta}} < \infty$ for any $\delta > 0$ and $\mu_A(\alpha) := \min(\mu(\alpha), \frac{\alpha}{2})$. Loosely, this means that upper bounds on the "work complexity" of fixed step ASGM and backtracking line search ASGM are arbitrarily close to $\mathcal{O}(\epsilon^{-2-\frac{1}{\mu(\alpha)}})$ and $\mathcal{O}(\epsilon^{-2-\frac{1}{\mu_A(\alpha)}})$ respectively. Likewise, when $f \in \mathcal{S}_{\lambda,L}^{1,1}$ and bounded below, we demonstrate that fixed step ASGM satisfies $\sup_k w_k \|\nabla f(\mathbf{x}_k)\|^{\frac{1}{\mu(\alpha)-\delta}} < \infty$ and backtracking line search ASGM satisfies

$\sup_k w_k \|\nabla f(\mathbf{x}_k)\|^{\frac{1}{\mu_A(\alpha)-\delta}} < \infty$ for any $\delta > 0$, indicating, loosely, that upper bounds on ASGM's work complexity for the two flavors are arbitrarily close to $\mathcal{O}(\epsilon^{-\frac{1}{\mu(\alpha)}})$ and $\mathcal{O}(\epsilon^{-\frac{1}{\mu_A(\alpha)-\delta}})$ respectively.

Comparing the first two columns of Table 1 with the third column gives a sense of the "price" that is paid due to the non-availability of an exact oracle. Since we have found that the gradient error decay rate $\mu(\alpha)$ is invariably smaller than the function error decay rate α, the resulting upper bound on the complexity of backtracking line Search ASGM coincides with the upper bound on the complexity of fixed step ASGM. Also notable are the corresponding complexities (listed in the last three columns of Table 1) that are obtained when Monte Carlo is used to construct the approximator $f(n;\cdot)$. We will see later that α is arbitrarily close to 1 when QMC is used, and that $\mu(\alpha) = 2\alpha/3$ when a central-difference approximation is used for gradient estimation. Considering these numbers, we see from the first, second and fourth columns that the work complexity of the proposed methods are superior to a corresponding method involving Monte Carlo whenever the use of QMC is viable.

1.4 **Paper Organization**

The subsequent sections of the paper are organized as follows. In Section 2 we present two application contexts that have motivated this study. Section 3 characterizes the quality of gradient approximates that might be used within ASGM. This is followed by Section 4 which contains the description, analysis, and main results pertaining to the convergence and convergence rate calculations for ASGM. We end the paper with some concluding remarks in Section 5.

2 Motivating Contexts

The predominant settings that motivate problems of the form Problem P are stochastic optimization [39, 41, 42, 6] and stochastic root-finding [32, 30, 33] problems, both large problem classes that have recently generated much attention. In what follows, we present two instances of Problem P chosen as an oversimplified version of an "actual problem" that one or more of the authors have recently encountered.

2.1 **Optimal Scheduling in a Queue**

Consider a time-varying service system, e.g. a healthcare facility, operating in $[0, T]$. Suppose that the facility services two types of patients: random arrivals corresponding to patients having semi-emergent circumstances and following a general probability law, e.g., the non-homogeneous Poisson process [38] with specified rate function $\rho(t), t \geq 0$, and scheduled arrivals pertaining to non-emergent patients who arrive at times $\mathbf{x} = (x_1, x_2, \ldots, x_d)$. The problem is to determine the times $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ of the scheduled arrivals in such a way that the functioning of the facility is optimal in the following sense. Suppose $h(\xi; \mathbf{x})$ is the expected virtual waiting time at time $\xi \in [0, T]$, that is, the expected amount

of time a "virtual customer" arriving at time ξ will have to wait before being served, given the deterministic arrivals have been scheduled at $\mathbf{x}$. The objective is to identify $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ such that the time average of the function $h(\cdot; \mathbf{x})$ is minimized. Formally, we want to solve the following optimization problem.

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) := T^{-1} \int_0^T h(\xi, \mathbf{x})\, d\xi, \\ &\text{subject to } \mathbf{x} \in [0, T]^d, \end{aligned} \tag{5}$$

where the function $h(\cdot;\cdot)$ in (5) can be determined through an oracle that solves the Kolmogorov forward equations [16] either exactly, or using closure approximations [23].

Let's perform the variable transformation $x_i = T\phi(\tilde{x}_i), i = 1, 2, \ldots, d$ where $\phi : \mathbb{R} \to [0, 1]$ is any strictly increasing smooth transformation with $\lim_{y\to-\infty} \phi(y) = 0$ and $\lim_{y\to\infty} \phi(y) = 1$. Let $\phi(\mathbf{x}) := (\phi(\tilde{x}_1), \phi(\tilde{x}_2), \ldots, \phi(\tilde{x}_d))$ and rewrite the problem in (5) as the unconstrained optimization problem:

$$\underset{\tilde{\mathbf{x}}\in\mathbb{R}^d}{\text{minimize}}\ f(\tilde{\mathbf{x}}) := T^{-1} \int_0^T h(\xi; T\phi(\tilde{\mathbf{x}}))\, d\xi. \tag{6}$$

To approximate the objective function in (6) at any given point $\tilde{\mathbf{x}}$, we could use one of a variety of methods, apart from naïve Monte Carlo. For example, using quasi-Monte Carlo (QMC) [27] with a low-dispersion sequence $\{u_n\}$, e.g., Sobol, Halton, Faure, we can estimate $f(\mathbf{x})$ as

$$f(n; \tilde{\mathbf{x}}) := T^{-1} \sum_{j=1}^{n} h(Tu_j, \phi(\tilde{\mathbf{x}})).$$

Then, the Koksma-Hlawka [27] bound guarantees that for all n, there exists $\sigma < \infty$ such that

$$|f(n; \tilde{\mathbf{x}}) - f(\tilde{\mathbf{x}})| < \sigma\, n^{-1} \log n. \tag{7}$$

The constant σ in (7) will depend, amongst other things, on the total variation of $h(\mathbf{x}, \cdot)$ on the interval $[0, T]$ and the particular QMC sequence that is used in constructing $f(n; \mathbf{x})$. Also, the term $\log n$ in (7) becomes $(\log n)^d$ when ξ is a d-dimensional random vector. To avoid going far afield, we will go into no further detail but direct the reader to [27].

The optimization problem in (5) and the QMC error bound in (7) assure us that the example just described is subsumed by Problem P with scaling function $\Gamma_f(\cdot) \equiv 1$, and the function error decay rate $\alpha = 1 - \epsilon$ for any choice $\epsilon > 0$.

In the above example, the function approximate $f(n; \cdot)$ can also be constructed using one of the standard numerical integration techniques [36, 12] in place of QMC. For example, we can employ the trapezoidal rule [36] by observing $h(\cdot; \tilde{\mathbf{x}})$ at n "evenly spaced" points $\xi_0, \xi_1, \ldots, \xi_{n-1}$ in the interval $[0, T]$ and obtain the function approximate

$$f_T(n; \tilde{\mathbf{x}}) = \frac{1}{2}h(\xi_0; \tilde{\mathbf{x}}) + h(\xi_1; \tilde{\mathbf{x}}) + h(\xi_2; \tilde{\mathbf{x}}) + \cdots + h(\xi_{n-2}; \tilde{\mathbf{x}}) + \frac{1}{2}h(\xi_{n-1}; \tilde{\mathbf{x}}).$$

The resulting error is well-studied; for example, since $h(\cdot; \tilde{\mathbf{x}})$ is continuous in $[0, T]$, we can apply a well-known crude bound from [11] to see that there exists $\sigma < \infty$

such that the resulting approximator satisfies $\sup_{\tilde{\mathbf{x}}\in\mathbb{R}^d} \frac{|f_T(n;\tilde{\mathbf{x}})-f(\tilde{\mathbf{x}})|}{\left(\frac{T^2}{8}\right)n^{-1}} < \sigma$ suggesting the scaling function $\Gamma_f(\cdot) = T^2/8$ and $\alpha = 1$. The use of Simpson's rule [36] will yield a similar error bound but with the constant $5T^2/72$ instead of $T^2/8$. (See [11] for other bounds on estimation error based on different levels of smoothness of the integrand when using numerical integration.)

2.2 **Classical Markowitz Portfolio Optimization**

The classical Markowitz portfolio optimization problem [20] seeks to identify weights $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ that minimize the function $f(\mathbf{x}) = \mu^T\mathbf{x} - \eta\,\mathbf{x}^T\,\Sigma\,\mathbf{x}$ subject to $\sum_{i=1}^d x_i = 1, x_i \geq 0$, where $\mu = \mathbb{E}[\mathbf{X}]$ and $\Sigma = \mathbb{V}\text{ar}(\mathbf{X})$ are the mean and covariance of a random vector $\mathbf{X} \in [0,\infty)^d$ representing the returns from a portfolio of d invested assets, and $\eta > 0$ is the risk-aversion parameter.

Since the returns $\mathbf{X}$ are usually the result of detailed models of evolution of an asset over time (e.g., see Chapter 3 in [13]), the quantities μ and Σ are not known in closed-form but approximators μ_n, Σ_n of μ and Σ, respectively, can be constructed using a method such as QMC. The resulting approximator $f(n;\cdot)$ of the objective function $f(\cdot)$ can then be constructed as

$$f(n;\mathbf{x}) = \mu_n^T\mathbf{x} - \eta\,\mathbf{x}^T\,\Sigma_n\,\mathbf{x}. \tag{8}$$

Furthermore, if QMC is used to approximate μ_n and Σ_n, as usual the approximators μ_n and Σ_n satisfy error bounds stemming from the Koksma-Hlawka bound under mild conditions. Then, similar to the example discussed in Section 2.1, there exist constants $\sigma_1 < \infty$ and $\sigma_2 < \infty$ such that for all $n \geq 1$,

$$\frac{\|\mu_n - \mu\|}{n^{-1}(\log n)^d} < \sigma_1 \text{ and } \frac{\|\Sigma_n - \Sigma\|}{n^{-1}(\log n)^d} < \sigma_2. \tag{9}$$

Combining (8) and (9), and after some algebra, we see that the resulting approximator satisfies for all $n \geq 1$ and some $\sigma < \infty$

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \frac{|\,f(n;\mathbf{x}) - f\,(\mathbf{x})\,|}{(\|\mathbf{x}\| + \eta\|\mathbf{x}\|^2)\,n^{-1}(\log n)^d} < \sigma. \tag{10}$$

We see then that the described example falls within the scope of Problem P with the approximator having the scaling function $\Gamma_f(\mathbf{x}) = \|\mathbf{x}\| + \eta\|\mathbf{x}\|^2$ and $\alpha = 1 - \epsilon$ for any $\epsilon > 0$.

3 The Gradient Approximator

As we will see shortly, ASGM is a gradient method that relies on the repeated construction of an approximator of the gradient function $\nabla f(\cdot)$. Given the approximator $f(n;\cdot)$ of the objective function $f(\cdot)$, a gradient approximator of $\nabla f(\cdot)$ can be constructed in a number of ways, the most prominent of which is by "finite differencing" observations of $f(n;\cdot)$ at carefully selected points. Two popular finite-differencing schemes are the forward and central difference approximations having the following partial derivative approximators at the point $\mathbf{x}$.

$$g_{fi}(\frac{n}{d+1};\mathbf{x}) := \zeta_n^{-1}\left(f(\frac{n}{d+1};\mathbf{x}+\zeta_n\mathbf{e}_i) - f(\frac{n}{d+1};\mathbf{x})\right), \quad i=1,2,\ldots,d; \tag{FD}$$

$$g_{ci}(\frac{n}{2d};\mathbf{x}) := \frac{\zeta_n^{-1}}{2}\left(f(\frac{n}{2d};\mathbf{x}+\zeta_n\mathbf{e}_i) - f(\frac{n}{2d};\mathbf{x}-\zeta_n\mathbf{e}_i)\right), \quad i=1,2,\ldots,d \tag{CD}$$

where $\zeta_n > 0$ is the step size for finite-difference, and $\mathbf{e}_i$ is a unit vector defined in Section 1.1. (In the expressions (FD) and (CD), we have ignored the possible non-integrality of $n/(d+1)$ and $n/(2d)$.) The forward-difference derivative approximate $g_f(n;\mathbf{x})$ is then obtained as $g_f(n;\mathbf{x}) = (g_{f1}(\frac{n}{d+1};\mathbf{x}), g_{f2}(\frac{n}{d+1};\mathbf{x}), \ldots, g_{fd}(\frac{n}{d+1};\mathbf{x}))$ by accumulating the partial derivatives in (FD); likewise, the central difference approximator $g_c(n;\mathbf{x}) = (g_{c1}(\frac{n}{2d};\mathbf{x}), g_{c2}(\frac{n}{2d};\mathbf{x}), \ldots, g_{cd}(\frac{n}{2d};\mathbf{x}))$ is obtained by accumulating the partial derivative approximators in (CD).

We now characterize the behavior of $g_f(n,\cdot)$ and $g_c(n,\cdot)$ as approximators of the gradient function $\nabla f(\cdot)$. Lemma 1 that follows asserts that both $g_f(n,\cdot)$ and $g_c(n,\cdot)$ are consistent approximators of the gradient function $\nabla f(\cdot)$ as long as the step size ζ_n used in finite-differencing decays to zero slower than $n^{-\alpha}$. Furthermore, for a particular choice of ζ_n, Lemma 1 characterizes the convergence rate of $g_f(n,\cdot)$ and $g_c(n,\cdot)$ expressed as a function of the error decay rate α of the function approximator $f(n,\cdot)$.

Lemma 1 *(Forward-Difference and Central-Difference Approximator Quality)*

(i) Suppose $f \in C_L^{1,1}$. Then, the forward-difference derivative approximator $g_f(n;\cdot)$ is a consistent approximator of $\nabla f(\cdot)$ if the step size ζ_n satisfies $\zeta_n = o^{-1}(n^{-\alpha})$ and $\zeta_n = o(1)$. Moreover, if the step size is chosen as $\zeta_n = cn^{-\alpha/2}, c \in (0,\infty)$, then for large enough n,

$$n^{\alpha/2}\|g_f(n;\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \sigma_{0,g_f} + \sigma_{1,g_f}\Gamma_f(\mathbf{x}), \tag{11}$$

where $\sigma_{0,g_f} = \sqrt{d}\left((d+1)^{\alpha}(\frac{2}{c})(\sigma_{0,f}+\sigma_{1,f}) + \frac{1}{2}Lc\right)$ and $\sigma_{1,g_f} = (d+1)^{\alpha}\sqrt{d}(\frac{2}{c})$.

(ii) Suppose $f \in C_\nu^{2,2}$. Then, the central-difference derivative approximator $g_c(n;\cdot)$ is a consistent approximator of $\nabla f(\cdot)$ if the step size ζ_n satisfies $\zeta_n = o^{-1}(n^{-\alpha})$ and $\zeta_n = o(1)$. Moreover, if the step size is chosen as $\zeta_n = cn^{-\alpha/3}, c \in (0,\infty)$, then for large enough n

$$n^{2\alpha/3}\|g_c(n;\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \sigma_{0,g_c} + \sigma_{1,g_c}\Gamma_f(\mathbf{x}), \tag{12}$$

where $\sigma_{0,g_c} = \sqrt{d}\left((2d)^{\alpha}c^{-1}(\sigma_{0,f}+\sigma_{1,f}) + \frac{1}{6}\nu c^2\right)$ and $\sigma_{1,g_c} = (2d)^{\alpha}\sqrt{d}c^{-1}$.

Proof In what follows, we prove only the assertions in (ii). A proof of the assertions in (i) follow in an identical fashion.

We see that

$$\begin{aligned} g_{ci}(\frac{n}{2d};\mathbf{x}) &= \frac{\zeta_n^{-1}}{2}\left(f(\frac{n}{2d},\mathbf{x}+\zeta_n\mathbf{e}_i) - f(\frac{n}{2d},\mathbf{x}-\zeta_n\mathbf{e}_i)\right) \\ &= \frac{\zeta_n^{-1}}{2}\left(f(\frac{n}{2d},\mathbf{x}+\zeta_n\mathbf{e}_i) - f(\mathbf{x}+\zeta_n\mathbf{e}_i)\right) - \frac{\zeta_n^{-1}}{2}\left(f(\frac{n}{2d},\mathbf{x}-\zeta_n\mathbf{e}_i) - f(\mathbf{x}-\zeta_n\mathbf{e}_i)\right) \\ &\qquad + \frac{\zeta_n^{-1}}{2}\left(f(\mathbf{x}+\zeta_n\mathbf{e}_i) - f(\mathbf{x}-\zeta_n\mathbf{e}_i)\right). \end{aligned} \tag{13}$$

Since we have assumed that the function $f \in C_{\nu}^{2,2}$ we can write

$$|\frac{\zeta_n^{-1}}{2}\left(f(\mathbf{x}+\zeta_n\mathbf{e}_i)-f(\mathbf{x}-\zeta_n\mathbf{e}_i)\right)-\frac{\partial}{\partial x_i}f(\mathbf{x})| \leq \frac{\nu}{6}\zeta_n^2. \tag{14}$$

Denoting $\Delta(n;\mathbf{y}) := |f(n;\mathbf{y})-f(\mathbf{y})|$, and using (13), (14), we write

$$|g_{ci}(\frac{n}{2d};\mathbf{x})-\frac{\partial}{\partial x_i}f(\mathbf{x})| \leq \frac{\zeta_n^{-1}}{2}\left(\Delta(\frac{n}{2d},\mathbf{x}+\zeta_n\mathbf{e}_i)+\Delta(\frac{n}{2d},\mathbf{x}-\zeta_n\mathbf{e}_i)\right)+\frac{\nu}{6}\zeta_n^2. \tag{15}$$

Since $f(n,\cdot)$ satisfies (FE_1), we know that for $n \geq n_f$, $\Delta(n;\mathbf{x}+\zeta_n\mathbf{e}_i) \leq \frac{n}{2d}^{-\alpha}(\sigma_{0,f}+\sigma_{1,f}\Gamma(\mathbf{x}+\zeta_n\mathbf{e}_i))$. Since $\Gamma_f(\cdot)$ is a continuous function, for large enough n,

$$\Delta(n;\mathbf{x}+\zeta_n\mathbf{e}_i)) \leq (\frac{n}{2d})^{-\alpha}(\sigma_{0,f}+\sigma_{1,f}\left(1+\Gamma_f(\mathbf{x})\right). \tag{16}$$

Similar arguments imply that for large enough n,

$$\Delta(n;\mathbf{x}-\zeta_n\mathbf{e}_i)) \leq (\frac{n}{2d})^{-\alpha}(\sigma_{0,f}+\sigma_{1,f}\left(1+\Gamma_f(\mathbf{x})\right). \tag{17}$$

The inequality in (15), along with (16) and (17), implies that $g_c(n;\mathbf{x})$ is a consistent approximator of $\nabla f(\cdot)$ if $\zeta_n = o^{-1}(n^{-\alpha})$ and $\zeta_n = o(1)$, proving the first asseertion of part (ii).

To prove the second assertion of part (ii), use (15), (16) and (17) to write, for large enough n,

$$\begin{aligned} n^{2\alpha/3}\|g_c(n;\mathbf{x})-\nabla f(\mathbf{x})\| &\leq n^{2\alpha/3}\left(\sqrt{d}\zeta_n^{-1}\frac{n}{2d}^{-\alpha}(\sigma_{0,f}+\sigma_{1,f}\left(1+\Gamma_f(\mathbf{x})\right)+\frac{\sqrt{d}}{6}\nu\zeta_n^2\right) \\ &= \sqrt{d}\left((2d)^{\alpha}c^{-1}(\sigma_{0,f}+\sigma_{1,f})+\frac{1}{6}\nu c^2\right)+\sqrt{d}c^{-1}(2d)^{\alpha}\Gamma_f(\mathbf{x}), \end{aligned} \tag{18}$$

where the inequality in (18) follows since we have chosen the step size $\zeta_n = cn^{-\alpha/3}$. We conclude from (18) that the second part of assertion (ii) holds.

Lemma 1 characterizes the error decay rate of the forward and central difference gradient approximators to be $\alpha/2$ and $2\alpha/3$ respectively, where α is the error decay rate associated with the function approximator $f(n;\cdot)$. Most importantly, the assertions of Lemma 1 imply that whenever the approximator $f(n;\cdot)$ satisfies the error structure (FE_1), so do the finite-difference gradient approximators, as described by (11) and (12), but with an altered scaling function and an altered error decay rate.

The optimization method we propose in the subsequent section assumes that we have at our disposal a gradient approximator $g(n;\cdot)$ constructed through one of various available means, e.g., finite differencing of the function approximator $f(n,\cdot)$ as in (FD) or (CD) or using even more number of design points [1]. Towards providing a general analysis that subsumes foward-difference, central-difference, and potentially other gradient estimation techniques, we assume going forward that the available gradient approximator $g(n,\cdot)$ satisfies the following stipulation

on accuracy. There exist $\sigma_{0,g}, \sigma_{1,g} < \infty$ and a threshold $m_g < \infty$ such that for $n \geq m_g$,

$$n^{\mu(\alpha)} \| g(n;\mathbf{x}) - \nabla f(\mathbf{x}) \| < \sigma_{0,g} + \sigma_{1,g} \Gamma_g^0(\mathbf{x}), \qquad (GE_1)$$

where $\Gamma_g(\cdot) := 1 + \Gamma_g^0(\cdot)$ is a known positive-valued continuous function, $\alpha > 0$ is the error decay rate of the approximator $f(n,\cdot)$, and $\mu(\alpha) > 0$ is the error decay rate of the gradient approximator $g(n,\cdot)$. (Throughout the rest of the paper, we use the function $\Gamma_g(\cdot) := 1 + \Gamma_g^0(\cdot)$ instead of $\Gamma_g^0(\cdot)$ as a matter of convention and convenience.)

Comparing (GE_1) to (11) and (12) appearing in the assertion of Lemma 1, we notice that when we use the forward-difference gradient approximator, that is, when $g(n;\mathbf{x}) := g_f(n;\mathbf{x})$, we have $\Gamma_g(\mathbf{x}) := \Gamma_f(\mathbf{x})$ and $\mu(\alpha) := \alpha/2$; likewise for the central-difference gradient approximator, that is, when $g(n;\mathbf{x}) := g_c(n;\mathbf{x})$, we have $\Gamma_g(\mathbf{x}) := \Gamma_f(\mathbf{x})$ and $\mu(\alpha) := 2\alpha/3$. (In particular, note that the error decay rate $\mu(\alpha)$ of the gradient approximator is a function of the error decay rate α of the function approximator.)

4 The Adaptive Sampling Gradient Method (ASGM)

Recall that Problem P seeks to minimize the function $f(\cdot)$ using only the approximates $f(n;\mathbf{x})$ assumed to satisfy the stipulation (FE_1). We have also assumed that a gradient approximate $g(n;\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$ satisfying the error stipulation (GE_1) can be constructed using appropriate finite-differencing of the function approximator $f(n;\cdot)$, as detailed in Section 3. And, importantly, observing $f(n;\cdot)$ at any point $\mathbf{x} \in \mathbb{R}^d$ entails oracle effort n, which is the sole unit of computational effort in this paper.

ASGM has a simple iterative structure: at each iteration k, take a step along the direction of the approximated negative gradient $-g(n(\mathbf{x}_k);\mathbf{x}_k)$, where $g(n(\mathbf{x}_k);\mathbf{x}_k)$ is constructed after exerting an amount of oracle effort $n(\mathbf{x}_k)$.

The ASGM recursion is as follows:

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k g(n(\mathbf{x}_k);\mathbf{x}_k), \quad k = 1, 2, \ldots.} \qquad (ASGM)$$

Different flavors of ASGM result from different ways of choosing the step size β_k and the sampling effort $n(\mathbf{x}_k)$ in (ASGM). We propose the following two flavors of ASGM based on whether or not the Lipschitz constant L associated with $f \in C_L^{1,1}$ is known.

(i) In *fixed step ASGM*, the step size β_k is chosen as $\beta_k = (1-\theta)/L$, where L satisfies $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. The oracle effort $n(\mathbf{x}_k)$ is chosen as

$$n(\mathbf{x}_k) = \min\{n \geq \eta_k : \Gamma_g(\mathbf{x}_k) n^{-\mu(\alpha)+\delta} \leq \theta \|g(n;\mathbf{x}_k)\|\}, \qquad (SS_c)$$

where $\{\eta_k\}$ is lower-bound sequence satisfying $\eta_k \to \infty$ as $k \to \infty$, $\mu(\alpha)$ is the gradient approximate decay rate appearing in (GE_1), and $\delta > 0, \theta \in (0, \mu(\alpha))$ are constants.

(ii) In *backtracking line search ASGM*, since no knowledge of a constant L is assumed, the step size β_k is chosen through a backtracking line search procedure akin to what is commonly done in deterministic line search algorithms [28]. (We list the backtracking line search procedure in Algorithm 1.) As we shall see later in detail, the backtracking line search procedure performed during the kth iteration requires estimation of the objective function and the gradient at the kth iterate $\mathbf{x}_k$, and potentially at several other candidates $\mathbf{x}_k^+(s) = \mathbf{x}_k - sg(n(\mathbf{x}_k); \mathbf{x}_k), s = \gamma^{i-1}\, s_0, i = 1, 2, \ldots$ evaluated before choosing the subsequent iterate $\mathbf{x}_{k+1}$. The respective oracle efforts at $\mathbf{x}_k$ and $\mathbf{x}_k^+(s)$ are given as

$$n_s(\mathbf{x}_k) := \min\{m \geq \eta_k : \max\left(\Gamma_g(\mathbf{x}_k), \sqrt{\Gamma_f(\mathbf{x}_k)}\right) m^{-(\mu_A(\alpha)-\delta)} \leq s^{\frac{1}{2}}\theta\|g(m;\mathbf{x}_k)\|\};$$
$$n_s(\mathbf{x}_k^+(s)) := \min\{m \geq \eta_k : \sqrt{\Gamma_f(\mathbf{x}_k^+(s))}\, m^{-\frac{\alpha}{2}+\frac{\delta}{2}} \leq s^{\frac{1}{2}}\theta\|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|\}. \qquad (SS_\ell)$$

The basis for the oracle effort expressions in (SS_c) and (SS_ℓ) will be clarified by the analysis that is to follow. Loosely speaking, however, the expressions in (SS_c) and (SS_ℓ) are chosen so that, during the kth iteration, the approximated error of the gradient approximate at the iterate $\mathbf{x}_k$ is in "balance" with a measure of proximity of $\mathbf{x}_k$ to a stationary point. This should be immediately evident at least for fixed step ASGM where enough oracle effort is expended to ensure that the term $\Gamma_g(\mathbf{x}_k)n^{-\mu(\alpha)+\delta}$ appearing to the left of the inequality in (SS_c), which is an approximation of the error in the gradient approximate at $\mathbf{x}_k$, stays at a fixed proportion θ with respect to the approximated gradient norm at $\mathbf{x}_k$. A similar logic dictates the choice of the oracle effort expression in (SS_ℓ), as will become clear when we analyze backtracking line search ASGM.

In both (SS_c) and (SS_ℓ), $\{\eta_k\} \to \infty$ is a lower-bound sequence and $\theta \in (0,1), \delta \in (0, \mu(\alpha))$ are user-defined constants. Further mild stipulations on the choice of η_k and θ hold but since such stipulations depend on the nature of the objective function f, we postpone specifying them until we present our results formally. The constant δ can be any positive real number but, as we shall see, smaller δ values result in better work complexities for ASGM.

In what follows, our interest is in characterizing the convergence and work complexity for fixed step ASGM and backtracking line search ASGM for the three contexts $f \in C_L^{1,1}$, $f \in \mathcal{F}_L^{1,1}$, and $f \in \mathcal{S}_{\lambda,L}^{1,1}$. We will analyze the case $f \in C_L^{1,1}$ in its entirety in the sense that the convergence and the work complexity rates will be stated and proved for both the fixed step ASGM and the backtracking line search ASGM flavors. For the context $f \in \mathcal{S}_{\lambda,L}^{1,1}$, we will treat fixed step ASGM in its entirety; for backtracking line search ASGM, however, in an attempt to avoid tedious repetition, we will be content with simply stating the result without proof. We omit results for the context $f \in \mathcal{F}_L^{1,1}$ since we have found that results on work complexity for $f \in \mathcal{F}_L^{1,1}$ do not deviate from the more general context $f \in C_L^{1,1}$.

Before we present the main results, we first state and prove a lemma involving bounds on the oracle effort in general form involving generic constants $p, c_q(\mathbf{x})$ and c_r to facilitate subsequent invocation under different settings.

Lemma 2 *Noting that (GE_1) holds, let n_G be such that for all $n \geq n_G$,*

$$\|g(n;\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \Gamma_g(\mathbf{x})\, n^{-\mu(\alpha)+\delta}. \qquad (19)$$

Let $\mathbf{x} \in \mathbb{R}^d$ *and define the oracle effort*

$$n(\mathbf{x}) = \min\{n \geq \eta : c_q(\mathbf{x})\, n^{-p} \leq c_r \|g(n;\mathbf{x})\|\}, \tag{20}$$

where $\eta \geq n_G + 1$*, and the constants* $p, c_q(\mathbf{x})$ *and* c_r *satisfy* $0 < p \leq \mu(\alpha) - \delta$*,* $0 < \delta < \mu(\alpha)$*,* $c_r > 0$*, and* $c_q(\mathbf{x}) > \Gamma_g(\mathbf{x}) c_r$*. Also, define the following constants.*

$$\eta_0 := (1 - \eta^{-1})^{-\mu(\alpha)+\delta} - 1; \tag{21}$$

$$c_0(\mathbf{x}) := \left(\frac{c_r}{c_q(\mathbf{x})}\right)^{-\frac{1}{p}} \left(1 - \left(\frac{\Gamma_g(\mathbf{x}) c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r}\right)(1 + \eta_0)\right)^{-\frac{1}{p}}. \tag{22}$$

Then, the oracle effort $n(\mathbf{x})$ *satisfies*

$$\left(\frac{c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r}\right)^{-\frac{1}{\mu(\alpha)-\delta}} \|\nabla f(\mathbf{x})\|^{-\frac{1}{\mu(\alpha)-\delta}} \leq n(\mathbf{x}) \leq \max(\eta, 1 + c_0(\mathbf{x}) \|\nabla f(\mathbf{x})\|^{-\frac{1}{p}}). \tag{23}$$

Proof Since $n(\mathbf{x}) \geq \eta > n_G$, we see that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|g(n(\mathbf{x});\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \Gamma_g(\mathbf{x}) n(\mathbf{x})^{-\mu(\alpha)+\delta} \leq \Gamma_g(\mathbf{x}) \left(\frac{c_r}{c_q(\mathbf{x})}\right) \|g(n(\mathbf{x});\mathbf{x})\|. \tag{24}$$

Also, the triangle inequality implies that, for all $\mathbf{x} \in \mathbb{R}^d$

$$\|g(n(\mathbf{x});\mathbf{x}) - \nabla f(\mathbf{x})\| \geq \|g(n(\mathbf{x});\mathbf{x})\| - \|\nabla f(\mathbf{x})\|. \tag{25}$$

The inequalities in (24) and (25) imply that

$$\|g(n(\mathbf{x});\mathbf{x})\| \leq \left(1 - \Gamma_g(\mathbf{x}) \left(\frac{c_r}{c_q(\mathbf{x})}\right)\right)^{-1} \|\nabla f(\mathbf{x})\|, \tag{26}$$

and

$$c_q(\mathbf{x}) n(\mathbf{x})^{-\mu(\alpha)+\delta} \leq c_r \|g(n(\mathbf{x}), \mathbf{x}\| \leq \frac{c_q(\mathbf{x}) c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r} \|\nabla f(\mathbf{x})\|, \tag{27}$$

where the second inequality in (27) follows from (26). By (27), we see that $n(\mathbf{x}) \geq \left(\frac{c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r}\right)^{-\frac{1}{\mu(\alpha)-\delta}} \|\nabla f(\mathbf{x})\|^{-\frac{1}{\mu(\alpha)-\delta}}$ and the left-hand side of (23) holds.

Multiplying both sides of (27) by $((n(\mathbf{x}) - 1)/n(\mathbf{x}))^{-\mu(\alpha)}$, we obtain

$$\begin{aligned} c_q(\mathbf{x})(n(\mathbf{x}) - 1)^{-\mu(\alpha)+\delta} &\leq \left(\frac{c_q(\mathbf{x}) c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r}\right) \|\nabla f(\mathbf{x})\| \left(\frac{n(\mathbf{x}) - 1}{n(\mathbf{x})}\right)^{-\mu(\alpha)+\delta} \\ &\leq \left(\frac{c_q(\mathbf{x}) c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r}\right)(1 + \eta_0) \|\nabla f(\mathbf{x})\|. \end{aligned} \tag{28}$$

Again apply the triangle inequality in (25) but for $n(\mathbf{x}) - 1$, we get

$$\begin{aligned} \|g(n(\mathbf{x}) - 1, \mathbf{x})\| &\geq \|\nabla f(\mathbf{x})\| - \Gamma_g(\mathbf{x})(n(\mathbf{x}) - 1)^{-\mu(\alpha)+\delta} \\ &\geq \|\nabla f(\mathbf{x})\| - \left(\frac{\Gamma_g(\mathbf{x}) c_r}{c_q(\mathbf{x}) - \Gamma_g(\mathbf{x}) c_r}\right)(1 + \eta_0) \|\nabla f(\mathbf{x})\|, \end{aligned} \tag{29}$$

where the last inequality in (29) follows from (28). Since $n(\mathbf{x})$ satisfies (20), we see that if $n(\mathbf{x}) > \eta$ then $n(\mathbf{x}) - 1$ does not satisfy the inequality in (20), that is,

$$\begin{aligned} c_q(\mathbf{x})(n(\mathbf{x})-1)^{-p} &\geq c_r\|g(n(\mathbf{x})-1,\mathbf{x})\| \\ &\geq c_r\left(1-\left(\frac{\Gamma_g(\mathbf{x})c_r}{c_q(\mathbf{x})-\Gamma_g(\mathbf{x})c_r}\right)(1+\eta_0)\right)\|\nabla f(\mathbf{x})\|, \end{aligned} \tag{30}$$

where the last inequality in (30) follows from (29). Employing some algebra, we conclude from (30) that if $n(\mathbf{x}) > \eta$, then $n(\mathbf{x}) \leq 1 + c_0(\mathbf{x})\|\nabla f(\mathbf{x})\|^{-\frac{1}{p}}$, implying that the right-hand side of the inequality in (23) holds.

4.1 ASGM Convergence and Work Complexity when $f \in \mathcal{C}_L^{1,1}$

We now analyze the performance of ASGM when the objective function f belongs to $\mathcal{C}_L^{1,1}$. Since f is not necessarily convex, ASGM provides no guarantees on the attainment of a local minimum of f. Instead, we will show that ASGM's iterates are such that the corresponding sequence of objective function values are strictly decreasing for large enough k, and that the corresponding sequence of true gradient norms converge to zero. Define $\mathcal{K}(k) := \arg\min_{0\leq j\leq k}\{\|g(n(\mathbf{x}_j;\mathbf{x}_j)\|\}$, and

$$k^*(k) := \max\{\mathcal{K}(k)\}. \tag{31}$$

Then, Theorem 1 analyzes the behavior of the "returned solution sequence"

$$\mathbf{x}_k^* := \mathbf{x}_{k^*(k)}. \tag{32}$$

In words, the returned solution $\mathbf{x}_k^*$ is simply the most recent iterate amongst those having the smallest "observed" gradient-approximate norm. The corresponding iteration and oracle effort are denoted $k^*(k)$ and n_k^* respectively. The use of the "best observed" solution is standard in non-convex settings [26].

Theorem 1 (Fixed Step ASGM for $f \in \mathcal{C}_L^{1,1}$) *Suppose that $f \in \mathcal{C}_L^{1,1}$ and bounded below, with $f^* := \inf\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d\}$. Let ASGM be applied on f with fixed step size $\beta_k = \beta \leq L^{-1}$ and oracle effort rule as in (SS_c) with $\theta \in (0, 1/2)$. Let the lower bound sequence $\{\eta_k\}$ in (SS_c) be such that $\eta_k \to \infty$ and $\eta_k = \mathcal{O}\left(k^{\frac{1}{2\mu(\alpha)-2\delta}}\right)$ as $k \to \infty$.*

1. *The sequence $\{\mathbf{x}_k\}$ is such that*

$$\lim_{k\to\infty}\|\nabla f(\mathbf{x}_k)\| = 0.$$

Suppose the gradient approximate $g(\cdot,\cdot)$ satisfies $g(n,\mathbf{x}) \neq 0$ for all $n, \mathbf{x}$.

2. *The sequence $\{k^*(k)\}$ satisfies $k^*(k) \to \infty$ as $k \to \infty$.*
3. *The sequence $\{\mathbf{x}_k^*\}$ is such that*

$$\lim_{k\to\infty}\|\nabla f(\mathbf{x}_k^*)\| = 0. \tag{33}$$

4. *For large enough k, the oracle effort $n(\mathbf{x}_k)$ satisfies*

$$n(\mathbf{x}_k) \leq (1 + c_0(\mathbf{x}_k))\|\nabla f(\mathbf{x}_k^*)\|^{-\frac{1}{\mu(\alpha)-\delta}}. \tag{34}$$

5. Recalling $w_k = \sum_{j=1}^{k} n(\mathbf{x}_j)$, there exists $\chi_c < \infty$ such that all k,

$$w_k \|\nabla f(\mathbf{x}_k^*)\|^{2+\frac{1}{\mu(\alpha)-\delta}} \leq \chi_c. \tag{35}$$

Proof Since (GE_1) holds, there exists n_G such that for all $\mathbf{x} \in \mathbb{R}^d$ and $n \geq n_G$:

$$\|g(n;\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \Gamma_g(\mathbf{x})\, n^{-\mu(\alpha)+\delta}. \tag{36}$$

Since $\eta_k \to \infty$, we see that $n(\mathbf{x}_k) \geq n_G$ for large enough k, that is, there exists k_G such that if $k \geq k_G$, $(g(n;\mathbf{x}_k) - \nabla f(\mathbf{x}_k))^T g(\mathbf{x}_k) \leq \|g(n;\mathbf{x}_k)\| \|g(n;\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\| \leq \Gamma_g(\mathbf{x}_k)\ (n(\mathbf{x}_k))^{-\mu(\alpha)+\delta} \leq \theta \|g(n;\mathbf{x}_k)\|^2$, implying that

$$\nabla f(\mathbf{x}_k)^T g(n(\mathbf{x}_k);\mathbf{x}_k) \geq (1-\theta)\|g(n(\mathbf{x}_k);\mathbf{x}_k)\|^2. \tag{37}$$

Also, since $f \in C_L^{1,1}$, we see that

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\leq f(\mathbf{x}_k) - \beta \nabla f(\mathbf{x}_k)^T g(n(\mathbf{x}_k);\mathbf{x}_k) + \frac{\beta}{2}\|g(n(\mathbf{x}_k);\mathbf{x}_k)\|^2, \end{aligned} \tag{38}$$

where the last inequality uses our assumption $\beta \leq L^{-1}$. Continuing from (38) and using (37), we get, for $k \geq k_G$, that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \beta(1-\theta)\|g(n(\mathbf{x}_k);\mathbf{x}_k)\|^2 + \frac{\beta}{2}\|g(n(\mathbf{x}_k);\mathbf{x}_k)\|^2, \tag{39}$$

$$\leq f(\mathbf{x}_k) - \frac{\beta(\frac{1}{2}-\theta)}{(1+\theta)^2}\|\nabla f(\mathbf{x}_k)\|^2, \tag{40}$$

where (40) follows since the sampling rule (SS_c) and the inequality in (36) imply that

$$(1+\theta)^{-1}\|\nabla f(\mathbf{x}_k)\| \leq \|g(n(\mathbf{x}_k);\mathbf{x}_k)\| \leq (1-\theta)^{-1}\|\nabla f(\mathbf{x}_k)\| \tag{41}$$

for $k \geq k_G$. Therefore $\beta(\frac{1}{2}-\theta)(1+\theta)^{-2}\|\nabla f(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$, and since $\theta \in (0, 1/2)$, every move during iterations $k \geq k_G$ is a "descent step." Furthermore,

$$\sum_{j=0}^{k} \|\nabla f(\mathbf{x}_j)\|^2 = \sum_{j=0}^{k_G} \|\nabla f(\mathbf{x}_j)\|^2 + \sum_{j=k_G+1}^{k} \|\nabla f(\mathbf{x}_j)\|^2 \tag{42}$$

$$\leq \sum_{j=0}^{k_G} \|\nabla f(\mathbf{x}_j)\|^2 + \frac{(1+\theta)^2}{\beta(\frac{1}{2}-\theta)}[f(\mathbf{x}_{k_G+1}) - f^*], \tag{43}$$

where (43) follows from (40) and the definition of f^*. The inequality in (43) implies $\sum_{j=0}^{\infty} \|\nabla f(\mathbf{x}_j)\|^2 < \infty$, and hence the first assertion of the theorem holds.

To prove the second assertion, we see that $\sum_{j=0}^{\infty} \|\nabla f(\mathbf{x}_j)\|^2 < \infty$ and (41) imply that $g(n(\mathbf{x}_k);\mathbf{x}_k) \to 0$ as $k \to \infty$. Now suppose $k^*(k)$ does not diverge as $k \to \infty$. Then there exists $k_1 < \infty$ such that for all $k > k_1$, $\|g(n(\mathbf{x}_{k_1});\mathbf{x}_{k_1})\| < \|g(n(\mathbf{x}_k);\mathbf{x}_k)\|$. Combining the last two observations, we see that $\|g(n(\mathbf{x}_{k_1});\mathbf{x}_{k_1})\| = 0$, giving rise to a contradiction since we have postulated that $g(n(\mathbf{x}),\mathbf{x}) \neq 0$ for each $n, \mathbf{x}$, and thus $k^*(k) \to \infty$.

We will next prove the third assertion of the theorem. Since we have proved that $k^*(k) \to \infty$, we know that there exists $k_2 < \infty$ such that if $k \geq k_2$, then $k^*(k) \geq k_G$ and $n_k^* \geq n_G$. We thus notice that for $j \geq k_2$,

$$\begin{aligned}\|\nabla f(\mathbf{x}_j)\| &\geq (1-\theta)\|g(n(\mathbf{x}_j);\mathbf{x}_j)\| \geq (1-\theta)\|g(n_j^*;\mathbf{x}_j^*)\| \\ &\geq \frac{1-\theta}{1+\theta}\|\nabla f(\mathbf{x}_j^*)\|, \end{aligned} \tag{44}$$

where the first and third inequalities in (44) follow from (41), and the second inequality in (44) follows from the definition in (32). Using $\sum_{j=0}^{\infty}\|\nabla f(\mathbf{x}_j)\|^2 < \infty$ and (44), we conclude that $\|\nabla f(\mathbf{x}_k^*)\| \to 0$.

Next, we prove the fourth assertion that appears in (34). We know that if $k \geq k_2$, then $k^*(k) \geq k_G$ and $n_k^* \geq n_G$. Use (42) and (41) to write for $k \geq k_2$,

$$\begin{aligned}\sum_{j=0}^{k}\|\nabla f(\mathbf{x}_j)\|^2 &\geq \sum_{j=0}^{k_G}\|\nabla f(\mathbf{x}_j)\|^2 + (1-\theta)^2\sum_{j=k_G+1}^{k}\|g(n(\mathbf{x}_j),\mathbf{x}_j)\|^2 \\ &\geq \sum_{j=0}^{k_G}\|\nabla f(\mathbf{x}_j)\|^2 + \left(\frac{1-\theta}{1+\theta}\right)^2(k-k_G)\,\|\nabla f(\mathbf{x}_k^*)\|^2. \end{aligned} \tag{45}$$

Combine (45) and (43) to see that for $k \geq k_2$,

$$\|\nabla f(\mathbf{x}_k^*)\|^2 \leq \left(\frac{(1+\theta)^4}{(1-\theta)^2}\right)\left(\frac{f(\mathbf{x}_0)-f^*}{\beta(\frac{1}{2}-\theta)}\right)(k-k_G)^{-1}. \tag{46}$$

Now, the right-hand side of the assertion of Lemma 2 (applied with $c_q = \Gamma_g(\mathbf{x}_k), c_r = \theta, p = \mu(\alpha)-\delta$) implies that there exists $k_3 < \infty$ such that for $k \geq k_3$,

$$n(\mathbf{x}_k) \leq \max(\eta_k, (1+c_0(\mathbf{x}))\|\nabla f(\mathbf{x}_k^*)\|^{-\frac{1}{\mu(\alpha)-\delta}}). \tag{47}$$

By (46) the term $\|\nabla f(\mathbf{x}_k^*)\|^{-\frac{1}{\mu(\alpha)-\delta}}$ appearing in (47), diverges at a rate faster than the lower bound sequence $\{\eta_k\}$, which is assumed to diverge slower than $k^{\frac{1}{2\mu(\alpha)-2\delta}}$. Furthermore, due to the first assertion of the theorem, (41), and the continuity of $\Gamma_f(\cdot)$, there exist constants k_4, c_0 such that for all $k \geq k_4$, $c_0(\mathbf{x}) \leq c_0$. We thus see that for $k \geq \max(k_2,k_3,k_4)$,

$$n(\mathbf{x}_k) \leq (1+c_0)\|\nabla f(\mathbf{x}_k^*)\|^{-\frac{1}{\mu(\alpha)-\delta}}, \tag{48}$$

proving the fourth assertion of the theorem.

The bound in (48) implies, after ignoring non-integrality issues, that for $k \geq k_5 = \max(k_2,k_3,k_4)$,

$$\begin{aligned} w_k = \sum_{j=0}^{k} n(\mathbf{x}_j) &= \sum_{j=0}^{k_5} n(\mathbf{x}_j) + \sum_{j=k_5}^{k} n(\mathbf{x}_j) \\ &\leq \left(\sum_{j=0}^{k_5} n(\mathbf{x}_j)\right) + (1+c_0)k\|\nabla f(\mathbf{x}_k^*)\|^{-\frac{1}{\mu(\alpha)-\delta}}. \end{aligned} \tag{49}$$

Use (46) and (49) to conclude that the fifth assertion of the theorem holds with $\chi_c = \left(\sum_{j=0}^{k_5} n(\mathbf{x}_j)\right)\left(\frac{(1+\theta)^4}{(1-\theta)^2}\right)\left(\frac{f(\mathbf{x}_0)-f^*}{\beta(\frac{1}{2}-\theta)}\right)^{\frac{1}{2(2+1/\mu(\alpha)-\delta)}} + (1+c_0)\left(\frac{(1+\theta)^4}{(1-\theta)^2}\right)\left(\frac{f(\mathbf{x}_0)-f^*}{\beta(\frac{1}{2}-\theta)}\right)$. □

An implication of Theorem 1 is that the work complexity of fixed step ASGM is $\mathcal{O}\left(\epsilon^{-2-\frac{1}{\mu(\alpha)-\delta}}\right)$, where $\delta \in (0, \mu(\alpha))$ is any (arbitrarily small) positive number. This means that an upper bound to the amount of oracle effort expended by fixed step ASGM's iterates to reach and stay within an ϵ-neighborhood of a zero of ∇f is arbitrarily close to $\mathcal{O}\left(\epsilon^{-2-\frac{1}{\mu(\alpha)}}\right)$. The work complexity expression in Theorem 1 also characterizes the dependence on the quality of the gradient approximator through the decay rate $\mu(\alpha)$. For example, Theorem 1 implies that QMC with central-difference derivatives will yield an upper bound on work complexity that is arbitrarily close to $\mathcal{O}(\epsilon^{-3.5})$ when using fixed step ASGM on smooth non-convex functions. It will be recalled [26] that the corresponding complexity rate when an exact oracle is available is $\mathcal{O}\left(\epsilon^{-2}\right)$, thus providing a sense of the price that is paid due to the lack of an exact oracle.

The stipulation that the lower bound sequence η_k diverge slower than $k^{\frac{1}{2\mu(\alpha)-2\delta}}$ is satisfied typically by logarithmic sequences. e.g., $\eta_k = \log k$. Also, note that while Theorem 1 charaterizes the behavior of the sequence $\{\nabla f(\mathbf{x}_k^*)\}$, this says nothing about convergence to a local minimum. More, e.g., convexity, needs to be assumed about the structure of the function $f(\cdot)$ to guarantee such convergence.

Theorem 1 is an analysis of *fixed step ASGM* where the step size β_k is fixed at $\beta_k = \beta \leq L^{-1}$. So, fixed step ASGM implicitly assumes knowledge of a constant L that satisfies $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$. For contexts where such a constant L is unknown, we propose *backtracking line search ASGM* that mimics a widely used technique in deterministic numerical optimization contexts [28].

Algorithm 1 Adaptive Backtracking Line Search

Inputs:
Current iterate $\mathbf{x}_k \in \mathbb{R}^d$; initial step $s_0 > 0$; line search contraction factor $\gamma \in (0,1)$; line search constant $c_F := \frac{1}{2} - s_0^{1/2}\theta - 2\theta^2$, effective decay rate $\mu_A(\alpha) := \min(\frac{\alpha}{2}, \mu(\alpha))$.
Outputs:
Next step size β_k.
Initialize:
i = 0;
repeat
$i = i + 1$;
$s_i := \gamma^{i-1} s_0$;
$\mathbf{x}_k^+(s_i) := \mathbf{x}_k - s_i g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k)$;

$$n_{s_i}(\mathbf{x}_k) := \min\{m \geq \eta_k : \max\left(\Gamma_g(\mathbf{x}_k), \sqrt{\Gamma_f(\mathbf{x}_k)}\right) m^{-(\mu_A(\alpha)-\delta)} \leq s_i^{\frac{1}{2}}\theta\|g(m;\mathbf{x}_k)\|\};$$

$$n_{s_i}(\mathbf{x}_k^+(s_i)) := \min\{m \geq \eta_k : \sqrt{\Gamma_f(\mathbf{x}_k^+(s))}\, m^{-\frac{\alpha}{2}+\frac{\delta}{2}} \leq s_i^{\frac{1}{2}}\theta\|g(n_{s_i}(\mathbf{x}_k);\mathbf{x}_k)\|\}; \qquad (SS_\ell)$$

until

$$f(n_{s_i}(\mathbf{x}_k^+(s_i)); \mathbf{x}_k^+(s_i)) \leq f(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k) - c_F s_i \|g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k)\|^2. \qquad (50)$$

return $\beta_k = s_i$.

In backtracking line search ASGM, during the kth iteration and starting from the point $\mathbf{x}_k$, a one-dimensional search is undertaken along the approximated negative gradient direction until a point that satisfies a sufficient decrease condition

is identified. As outlined in Algorithm 1, the backtracking line search procedure always starts with the step size s_0 and successively checks if the points $\mathbf{x}_k^+(s_i) := \mathbf{x}_k - \gamma^{i-1} s_0 g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k), i = 1, 2, \ldots$ satisfy the sufficient decrease condition (50). Akin to the Armijo condition [28], the sufficient decrease condition in (50) involves checking if the approximated function value at the candidate point $\mathbf{x}_k^+(s_i)$ lies below the line passing through the point $(\mathbf{x}_k, f(n_{s_i}(\mathbf{x}_k), \mathbf{x}_k))$ and having slope equal to a constant c_F times the approximated negative gradient norm at the point $\mathbf{x}_k$. As we will see, the constant c_F is chosen carefully, to ensure that the backtracking procedure always terminates with a point that satisfies the sufficient decrease condition.

The oracle effort sizes $n_{s_i}(\mathbf{x}_k)$ and $n_{s_i}(\mathbf{x}_k^+(s_i))$ used to construct the function and gradient approximates appearing in (50) are specified through the rule (SS_ℓ) appearing in Algorithm 1. The oracle effort rule (SS_ℓ) is designed to keep the square-root of the error approximates of $f(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k)$ and $f(n_{s_i}(\mathbf{x}_k^+(s_i)), \mathbf{x}_k^+(s_i))$, and the error approximate of $\|g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k)\|$, within a fixed proportion of the gradient norm $\|g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k)\|$. It will be seen in the ensuing analysis that the need for maintaining such a balance between the various error approximates and the gradient norm is rooted in a certain fundamental inequality that governs $C_L^{1,1}$ functions.

We first demonstrate through Lemma 3 that the backtracking linesearch procedure Algorithm 1 always terminates successfully, that is, it takes as input the current iterate $\mathbf{x}_k$ and successfully returns a step size β_k for obtaining the next iterate $\mathbf{x}_{k+1}$. Lemma 3 also provides an upper bound for the total amount of oracle effort expended by Algorithm 1 before termination.

Lemma 3 *Suppose that $f \in C_L^{1,1}$ and bounded below. Let ASGM with backtracking line search procedure Algorithm 1 be applied on the function f with constants $\theta \in (0,1)$, $s_0 > 0$, and $c_F \in (0,1)$. Let these constants be chosen so that $s_0^{1/2}\theta < 1$, $\theta \leq \frac{1}{4}(\sqrt{s_0+4} - \sqrt{s_0})$, and $c_F = \frac{1}{2} - s_0^{1/2}\theta - 2\theta^2$. If the iterate $\mathbf{x}_k$ satisfies $\|\nabla f(\mathbf{x}_k)\| \neq 0$, then the following assertions hold.*

1. *The backtracking line search procedure in Algorithm 1 always terminates, that is, for each k, the termination criterion (50) for the recursive loop in Algorithm 1 is satisfied after a finite number of steps.*
2. *Define $\mu_A(\delta) := \min(\alpha/2, \mu(\alpha))$ and $k_{G,F} := \min\{k : \eta_k \geq n_{G,F}\}$, where the constant $n_{G,F}$ is such that for all $n \geq n_{G,F}$,*

$$\|g(n;\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \Gamma_g(\mathbf{x})\, n^{-\mu(\alpha)+\delta}; \quad \|f(n;\mathbf{x}) - f(\mathbf{x})\| \leq \Gamma_f(\mathbf{x})\, n^{-\alpha+2\delta}. \tag{51}$$

Then, during the kth iteration, for $k \geq k_{G,F}$, the backtracking line search procedure in Algorithm 1 is guaranteed to terminate in $I^* := 2 - (\log\gamma)^{-1}(\log s_0 + \log L)$ or fewer steps. Furthermore, the oracle effort $n(\mathbf{x}_k)$ expended by Algorithm 1 during the kth iteration satisfies

$$n(\mathbf{x}_k) \leq 2I^* \max\left(\eta_k, 1 + \max\left(c_A(\mathbf{x}_k), c_A^*(\mathbf{x}_k)\right) \max\left(1, \|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}\right)\right), \tag{52}$$

where $\eta_A := (1-\eta_0^{-1})^{-(\mu_A(\alpha)-\delta)} - 1$,

$$c_A(\mathbf{x}_k) := \left(\frac{\beta_k^{\frac{1}{2}}\theta}{\max(\sqrt{\Gamma_f(\mathbf{x}_k)}, \Gamma_g(\mathbf{x}_k))}\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \left(1-\frac{\theta}{1-\theta}(1+\eta_A)\right)^{-\frac{1}{\mu_A(\alpha)-\delta}},$$

$$c_A^*(\mathbf{x}_k) := \left(\frac{\beta_k^{\frac{1}{2}}\theta}{\sqrt{\Gamma_f^*(\mathbf{x}_k)}}\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \left(1-\frac{\theta}{1-\theta}(1+\eta_A)\right)^{-\frac{1}{\mu_A(\alpha)-\delta}},$$

$$\Gamma_f^*(\mathbf{x}_k) := \sup\{\Gamma_f(\mathbf{y}) : \mathbf{y} \in B\left(\mathbf{x}_k, \frac{s_0}{1+\beta_k^{\frac{1}{2}}\theta}\|\nabla f(\mathbf{x}_k)\|\right)\},$$

and β_k *is the step size returned by Algorithm 1 upon termination.*

Proof Let $n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+)$ be the oracle efforts in Algorithm 1 expressed as a function of a step s, that is,

$$n_s(\mathbf{x}_k) := \min\{m \geq \eta_k : \max\left(\Gamma_g(\mathbf{x}_k), \sqrt{\Gamma_f(\mathbf{x}_k)}\right) m^{-(\mu_A(\alpha)-\delta)} \leq s^{\frac{1}{2}}\theta\|g(m;\mathbf{x}_k)\|\};$$

$$n_s(\mathbf{x}_k^+(s)) := \min\{m \geq \eta_k : \sqrt{\Gamma_f(\mathbf{x}_k^+(s))}\, m^{-\frac{\alpha}{2}+\frac{\delta}{2}} \leq s^{\frac{1}{2}}\theta\|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|\}. \quad (53)$$

Suppose for now that the following two conditions hold: (i) $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+(s))) \geq n_{G,F}$, and (ii) $s \in (0, \min(L^{-1}, s_0))$. Then the guarantees in (51) hold, and using Cauchy-Schwarz [5] inequality we write

$$\begin{aligned}
\nabla f(\mathbf{x}_k)^T g(n_s(\mathbf{x}_k);\mathbf{x}_k) &\geq \|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|^2 - \|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|\|\nabla f(\mathbf{x}_k) - g(n_s(\mathbf{x}_k);\mathbf{x}_k)\| \\
&\geq \|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|^2 - \|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|\, \Gamma_g(\mathbf{x}_k)\, n_s(\mathbf{x}_k)^{-\mu(\alpha)+\delta} \\
&\geq \|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|^2 - s^{\frac{1}{2}}\theta\|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|^2 \quad (54) \\
&= (1-s^{\frac{1}{2}}\theta)\|g(n_s(\mathbf{x}_k);\mathbf{x}_k)\|^2, \quad (55)
\end{aligned}$$

where (54) follows from the expression for $n_s(\mathbf{x}_k)$ in (53). (Since $s \leq s_0 < \theta^{-2}$, (55) implies that $-g(n_s(\mathbf{x}_k);\mathbf{x}_k)$ forms a descent direction for f at the point $\mathbf{x}_k$.) Recalling the notation $\mathbf{x}_k^+(s) := \mathbf{x}_k - sg(n_s(\mathbf{x}_k);\mathbf{x}_k)$, and since $f \in C_L^{1,1}$, we can write

$$\begin{aligned}
f(\mathbf{x}_k^+(s)) &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}_k^+(s) - \mathbf{x}_k) + \frac{1}{2}L\|\mathbf{x}_k^+(s) - \mathbf{x}_k\|^2 \\
&\leq f(\mathbf{x}_k) - s\nabla f(\mathbf{x}_k)^T g(n_s(\mathbf{x}_k);\mathbf{x}_k)) + \frac{1}{2}s^2L\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 \\
&\leq f(\mathbf{x}_k) - \left(1 - s^{1/2}\theta - \frac{1}{2}sL\right) s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2, \quad (56)
\end{aligned}$$

where the last inequality follows from (55). Also, since $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+)) \geq n_{G,F}$, the function quality bound in (51) holds and we have

$$|f(n_s(\mathbf{x}_k);\mathbf{x}_k) - f(\mathbf{x}_k)| \leq \Gamma_f(\mathbf{x}_k) n_s(\mathbf{x}_k)^{-\alpha+2\delta}; \quad (57)$$

$$|f(n_s(\mathbf{x}_k^+(s));\mathbf{x}_k^+(s)) - f(\mathbf{x}_k^+(s))| \leq \Gamma_f(\mathbf{x}_k^+(s))\, n_s(\mathbf{x}_k^+(s))^{-\alpha+2\delta}. \quad (58)$$

Combining (56) with (57) and (58), we see that

$$
\begin{aligned}
&f(n_s(\mathbf{x}_k^+(s));\mathbf{x}_k^+(s)) - f(n_s(\mathbf{x}_k);\mathbf{x}_k) \\
&\quad \le \Gamma_f(\mathbf{x}_k^+(s)) n_s(\mathbf{x}_k^+(s))^{-\alpha+2\delta} + \Gamma_f(\mathbf{x}_k) n_s(\mathbf{x}_k)^{-\alpha+2\delta} \\
&\qquad\qquad - \left(1 - s^{1/2}\theta - \frac{1}{2}sL\right) s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 \\
&\quad \le 2s\theta^2\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 - \left(1 - s^{1/2}\theta - \frac{1}{2}sL\right) s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 \qquad (59) \\
&\quad \le -\left(1 - s^{1/2}\theta - \frac{1}{2}sL - 2\theta^2\right) s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 \\
&\quad \le -\left(1 - s^{1/2}\theta - \frac{1}{2} - 2\theta^2\right) s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 \qquad (60) \\
&\quad \le -\left(\frac{1}{2} - s_0^{1/2}\theta - 2\theta^2\right) s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2 \qquad (61) \\
&\quad = -c_F s\|g(n_s(\mathbf{x}_k);\mathbf{x}_k))\|^2, \qquad (62)
\end{aligned}
$$

where (59) follows from the expressions for $n_s(\mathbf{x}_k)$ and $n_s(\mathbf{x}_k^+(s))$ in (53) and recalling that $\mu_A(\alpha) := \min(\alpha/2, \mu(\alpha)) \le \alpha/2$; (60) follows since we have assumed $s \le L^{-1}$; (61) follows since $s \le s_0$; and (62) follows from the definition of c_F. Comparing the termination criterion in (50) of Algorithm 1 with (62) we thus see that the Algorithm 1 will terminate if the two conditions (i) $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+)) \ge n_{G,F}$ and (ii) $s \in (0, \min(L^{-1}, s_0))$ are satisfied during some step in the recursive loop of Algorithm 1. The condition in (ii) will be satisfied if the backtracking line search iteration number i exceeds $I^* := \min\{i : s_0\gamma^{i-1} \le L^{-1}\}$ simply because the step size is reduced sequentially by a factor $\gamma \in (0,1)$ during each iteration of the backtracking line search. To see that the condition in (i), that is, $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+)) \ge n_{G,F}$ will also be "ultimately" satisfied, let us suppose $k < k_{G,F}$. (The constant $k_{G,F}$ is defined in the statment of the lemma.) Then the expressions for $n_s(\mathbf{x}_k)$ and $n_s(\mathbf{x}_k^+)$ show that there exists $s^*(\mathbf{x}_k)$ such that $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+)) \ge n_{G,F}$ for all $s \le s^*(\mathbf{x}_k)$. On the other hand, if $k \ge k_{G,F}$, we see that $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+)) \ge n_{G,F}$ by the definition of $k_{G,F}$. The first assertion of the lemma is thus proved.

As part of proving the first assertion, we argued that if $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+(s))) \ge n_{G,F}$ and $s \in (0, \min(L^{-1}, s_0))$, the termination criterion in (50) will be satisfied. Now, if $k \ge k_{G,F}$, we know by the definition of $k_{G,F}$ that $\min(n_s(\mathbf{x}_k), n_s(\mathbf{x}_k^+(s))) \ge n_{G,F}$. Furthermore, we see that due to the backtracking line search, Algorithm 1 terminates in I steps or less where

$$
I^* := \min\{i : s_0\gamma^{i-1} \le L^{-1}\} \le 2 - \frac{\log s_0 + \log L}{\log \gamma}. \qquad (63)
$$

Next, recalling that β_k is the output (i.e., step length) returned by Algorithm 1, we invoke Lemma 2 with $c_q(\mathbf{x}_k) := \max(\Gamma_g(\mathbf{x}_k), \sqrt{\Gamma_f(\mathbf{x}_k)})$, $c_r := \beta_k^{\frac{1}{2}}\theta$, and $p := \mu_A(\alpha) - \delta = \min(\alpha/2, \mu(\alpha)) - \delta$ to see that

$$
n_{s_i}(\mathbf{x}_k) \le n_{\beta_k}(\mathbf{x}_k) \le \max(\eta_k, 1 + c_A(\mathbf{x}_k)\,\|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}). \qquad (64)
$$

Similarly, noting the expression for $n_{s_i}(\mathbf{x}_k^+(s_i))$ in (SS_ℓ), invoke Lemma 2 with $c_q(\mathbf{x}_k^+(s_i)) := \sqrt{\Gamma_f(\mathbf{x}_k^+(s_i))}$, $c_r := s_i^{\frac{1}{2}}\theta$, and $p = \alpha/2 - \delta/2$ so that

$$n_{s_i}(\mathbf{x}_k^+(s_i)) \leq \max(\eta_k, 1 + \tilde{c}_A(\mathbf{x}_k^+(s_i))\, \|\nabla f(\mathbf{x}_k)\|^{-2/(\alpha-\delta)}), \tag{65}$$

where

$$\tilde{c}_A(\mathbf{x}_k^+(s_i)) := \left(\frac{s_i^{\frac{1}{2}}\theta}{\sqrt{\Gamma_f(\mathbf{x}_k^+(s_i))}}\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \left(1 - \frac{\theta}{1-\theta}(1+\eta_A)\right)^{-\frac{1}{\mu_A(\alpha)-\delta}}.$$

We notice that since $\beta_k \leq s_i$,

$$\begin{aligned}
\tilde{c}_A(\mathbf{x}_k^+(s_i)) &\leq \left(\frac{\beta_k{}^{\frac{1}{2}}\theta}{\sqrt{\Gamma_f(\mathbf{x}_k^+(s_i))}}\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \left(1 - \frac{\theta}{1-\theta}(1+\eta_A)\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \\
&\leq \left(\frac{\beta_k{}^{\frac{1}{2}}\theta}{\sqrt{\Gamma_f^*(\mathbf{x}_k)}}\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \left(1 - \frac{\theta}{1-\theta}(1+\eta_A)\right)^{-\frac{1}{\mu_A(\alpha)-\delta}} \qquad (66) \\
&=: c_A^*(\mathbf{x}_k), \qquad (67)
\end{aligned}$$

where (66) holds since $k \geq k_{G,F}$ implying that $\|g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k)\| \leq (1+\beta_k^{\frac{1}{2}}\theta)^{-1}\|\nabla f(\mathbf{x}_k)\|$ and hence $\mathbf{x}_k^+(s_i) := \mathbf{x}_k - s_i g(n_{s_i}(\mathbf{x}_k); \mathbf{x}_k) \in B\left(\mathbf{x}_k, \frac{s_0}{1+\beta_k^{\frac{1}{2}}\theta}\|\nabla f(\mathbf{x}_k)\|\right)$. Plugging (67) in (65), we get

$$n_{s_i}(\mathbf{x}_k^+(s_i)) \leq \max(\eta_k, 1 + c_A^*(\mathbf{x}_k)\, \|\nabla f(\mathbf{x}_k)\|^{-2/(\alpha-\delta)}). \tag{68}$$

Combining (64) and (68), and noticing that $\mu_A(\alpha) - \delta \leq \alpha/2 - \delta/2$, we see that the oracle effort $n_{s_i}(\mathbf{x}_k) + n_{s_i}(\mathbf{x}_k^+(s_i))$ during the ith step of Algorithm 1 satisfies, for $k \geq k_{G,F}$,

$$\begin{aligned}
&n_{s_i}(\mathbf{x}_k) + n_{s_i}(\mathbf{x}_k^+(s_i)) \leq \\
&\qquad 2\max\left(\eta_k, 1 + \max(c_A(\mathbf{x}_k)\,, c_A^*(\mathbf{x}_k)\,)\max\left(1, \|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}\right)\right).
\end{aligned} \tag{69}$$

Using (69) and since we have argued that Algorithm 1 terminates in $I^* := \min\{i : s_0\gamma^{i-1} \leq L^{-1}\} \leq 2 - \frac{\log s_0 + \log L}{\log \gamma}$ or fewer steps when $k \geq k_{G,F}$, the total oracle effort $n(\mathbf{x}_k)$ expended by Algorithm 1 when $k \geq k_{G,F}$ satisfies

$$n(\mathbf{x}_k) \leq 2I^* \max\left(\eta_k, 1 + \max(q\, c_A(\mathbf{x}_k)\,, c_A^*(\mathbf{x}_k)\,)\|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}\right).$$

This proves the second assertion of the lemma. □

We next present Theorem 2 that uses Lemma 3 to establish that the sequence of true gradient norms at the iterates generated by backtracking line search ASGM converges to zero. Theorem 2 also establishes an upperbound on the resulting work complexity.

Theorem 2 (Backtracking Line Search ASGM for $f \in \mathcal{C}_L^{1,1}$) *Let the postulates of Lemma 3 hold and let the lower bound sequence $\{\eta_k\}$ in (SS_ℓ) be such that $\eta_k \to \infty$ and $\eta_k = \mathcal{O}\left(k^{\frac{1}{2\mu_A(\alpha)-2\delta}}\right)$ as $k \to \infty$.*

1. *The sequence $\{\mathbf{x}_k\}$ is such that $\lim_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| \to 0$.*

Suppose the gradient approximate $g(\cdot,\cdot)$ satisfies $g(n(\mathbf{x}),\mathbf{x}) \neq 0$ for all $n, \mathbf{x}$. Then,

2. *The sequence $\{k^*(k)\}$ (defined in (31)) satisfies $k^*(k) \to \infty$ as $k \to \infty$.*
3. *The sequence $\{\mathbf{x}_k^*\}$ (defined in (32)) satisfies $\lim_{k\to\infty} \|\nabla f(\mathbf{x}_k^*)\| \to 0$.*
4. *For $k \geq k_{G,F}$, the oracle effort $n(\mathbf{x}_k)$ satisfies*

$$n(\mathbf{x}_k) \leq 2I^* \left(1 + \max\left(c_A(\mathbf{x}_k), c_A^*(\mathbf{x}_k)\right) \max\left(1, \|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}\right)\right),$$

   where $k_{G,F}, c_A(\mathbf{x}_k)$, $c_A^*(\mathbf{x}_k)$, and I^* are as defined in Lemma 3.
5. *There exists a constant $\chi_u < \infty$ such that the total oracle effort $w_k = \sum_{j=1}^k n(\mathbf{x}_j)$ expended until the kth iteration of backtracking line search ASGM satisfies*

$$w_k \|\nabla f(\mathbf{x}_k^*)\|^{2+\frac{1}{\mu_A(\alpha)-\delta}} < \chi_u. \tag{70}$$

Proof Since the postulates of Lemma 3 hold, we see that the backtracking line search procedure in Algorithm 1 terminates successfully. Specifically, the $(k+1)$th iterate $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k g(n(\mathbf{x}_k); \mathbf{x}_k)$ satisfies

$$f(n(\mathbf{x}_k); \mathbf{x}_{k+1}) \leq f(n(\mathbf{x}_k); \mathbf{x}_k) - c_F \beta_k \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|^2. \tag{71}$$

Also, the function approximator quality condition (FE_1) guarantees that for $k \geq k_{G,F}$, we have $|f(n(\mathbf{x}_{k+1}); \mathbf{x}_{k+1}) - f(\mathbf{x}_{k+1})| \leq \Gamma_f(\mathbf{x}_{k+1}) n(\mathbf{x}_{k+1})^{-\alpha+2\delta}$ and that $|f(n(\mathbf{x}_k); \mathbf{x}_k) - f(\mathbf{x}_k)| \leq \Gamma_f(\mathbf{x}_k) n(\mathbf{x}_k)^{-\alpha+2\delta}$. (See statement of Lemma 3 for definition of $k_{G,F}$.) Combining these with (71), we get for $k \geq k_{G,F}$ that

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \Gamma_f(\mathbf{x}_{k+1}) n(\mathbf{x}_{k+1})^{-\alpha+2\delta} + \Gamma_f(\mathbf{x}_k) n(\mathbf{x}_k)^{-\alpha+2\delta} - c_F \beta_k \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|^2 \\ &\leq f(\mathbf{x}_k) + 2\theta^2 \beta_k \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|^2 - c_F \beta_k \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|^2 \qquad (72) \\ &= f(\mathbf{x}_k) - \beta_k \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|^2 (\frac{1}{2} - s_0^{\frac{1}{2}} - 4\theta^2), \qquad (73) \end{aligned}$$

where (72) follows from the oracle effort rule (SS_ℓ) used in Algorithm 1, and (73) from the definition of c_F. Now notice that the stipulation on θ and s_0 guarantees that the multiplier $(\frac{1}{2} - s_0^{\frac{1}{2}} - 4\theta^2)$ in (73) satisfies $\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2 > 0$. Also, from Lemma 3, we know that the procedure in Algorithm 1 terminates in $I^* \leq (2 - \frac{\log s_0 + \log L}{\log \gamma})$ or fewer steps if $k \geq k_{G,F}$. Since the step size is reduced by a factor γ during each step, this implies that for $k \geq k_{G,F}$, $\beta_k \geq \gamma L^{-1}$. Furthermore, the function approximator quality stipulation (FE_1) and the oracle effort rule in Algorithm 1 guarantee that for $k \geq k_{G,F}$,

$$\|g(n(\mathbf{x}_k); \mathbf{x}_k) - \nabla f(\mathbf{x}_k)\| \leq \Gamma_g(\mathbf{x}_k) n(\mathbf{x}_k)^{-\mu(\alpha)+\delta} \leq \beta_k^{1/2} \theta \|g(n(\mathbf{x}_k); \mathbf{x}_k)\| \tag{74}$$

and hence

$$\frac{\|\nabla f(\mathbf{x}_k)\|}{(1 - s_0^{\frac{1}{2}}\theta)} \geq \frac{\|\nabla f(\mathbf{x}_k)\|}{(1 - \beta_k^{\frac{1}{2}}\theta)} \geq \|g(n(\mathbf{x}_k); \mathbf{x}_k)\| \geq \frac{\|\nabla f(\mathbf{x}_k)\|}{(1 + \beta_k^{\frac{1}{2}}\theta)} \geq \frac{\|\nabla f(\mathbf{x}_k)\|}{(1 + s_0^{\frac{1}{2}}\theta)}. \tag{75}$$

Since $\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2 > 0$ and we have argued that $\beta_k \geq \gamma L^{-1}$, and using the right-hand side of (75) in the inequality (73), we get for $k \geq k_{G,F}$ that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \gamma L^{-1} \left(\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2 \right) \frac{\|\nabla f(\mathbf{x}_k)\|^2}{1 + s_0^{1/2}\theta}. \tag{76}$$

Summing (76) across iterations, we then get

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_{k_{G,F}}) - \gamma L^{-1} \left(\frac{\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2}{1 + s_0^{1/2}\theta} \right) \left(\sum_{i=k_{G,F}}^{k} \|\nabla f(\mathbf{x}_i)\|^2 \right). \tag{77}$$

Recalling that $f(\mathbf{x}_i) \geq f^* := \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, we have

$$\sum_{i=k_{G,F}}^{k} \|\nabla f(\mathbf{x}_i)\|^2 \leq \left(\frac{1 + s_0^{1/2}\theta}{\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2} \right) \frac{L}{\gamma} \left(f(\mathbf{x}_{k_{G,F}}) - f^* \right). \tag{78}$$

The inequality in (78) implies that as $k \to \infty$, $\sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|^2 < \infty$, and the first assertion of the theorem is proved.

To prove the second assertion, we see that $\sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|^2 < \infty$ and (75) imply that $g(n(\mathbf{x}_k); \mathbf{x}_k) \to 0$ as $k \to \infty$. Now suppose $k^*(k)$ does not diverge as $k \to \infty$. Then there exists $\tilde{k}_1 < \infty$ such that for all $k > \tilde{k}_1$, $\|g(n(\mathbf{x}_{\tilde{k}_1}); \mathbf{x}_{\tilde{k}_1})\| < \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|$. Combining the last two observations, we see that $\|g(n(\mathbf{x}_{\tilde{k}_1}); \mathbf{x}_{\tilde{k}_1})\| = 0$ leading to a contradiction since we have postulated that $g(n(\mathbf{x}), \mathbf{x}) \neq 0$ for each $n, \mathbf{x}$, and thus $k^*(k) \to \infty$.

We will next prove the third assertion of the theorem. Since we have proved that $k^*(k) \to \infty$, we know that there exists $\tilde{k}_2 < \infty$ such that if $k \geq \tilde{k}_2$, then $k^*(k) \geq k_{G,F}$ and $n_k^* \geq n_{G,F}$. We thus notice that for $j \geq \tilde{k}_2$,

$$\begin{aligned} \|\nabla f(\mathbf{x}_j)\| &\geq (1 - s_0^{\frac{1}{2}}\theta) \|g(n(\mathbf{x}_j); \mathbf{x}_j)\| \geq (1 - s_0^{\frac{1}{2}}\theta) \|g(n_j^*; \mathbf{x}_j^*)\| \\ &\geq \frac{1 - s_0^{\frac{1}{2}}\theta}{1 + s_0^{\frac{1}{2}}\theta} \|\nabla f(\mathbf{x}_j^*)\|, \end{aligned} \tag{79}$$

where the first and third inequalities in (79) follow from (75), and the second inequality in (79) follows from the definition of $\mathbf{x}_j^*$. Using $\sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|^2 < \infty$ and (79), we conclude that $\|\nabla f(\mathbf{x}_k^*)\| \to 0$.

We will next prove the fourth assertion. Again we write for $k \geq \tilde{k}_2$ that

$$\begin{aligned} \sum_{j=0}^{k} \|\nabla f(\mathbf{x}_j)\|^2 &\geq \sum_{j=0}^{k_{G,F}} \|\nabla f(\mathbf{x}_j)\|^2 + (1 - s_0^{\frac{1}{2}}\theta)^2 \sum_{j=k_{G,F}+1}^{k} \|g(n(\mathbf{x}_j), \mathbf{x}_j)\|^2 \\ &\geq \sum_{j=0}^{k_{G,F}} \|\nabla f(\mathbf{x}_j)\|^2 + \left(\frac{1 - s_0^{\frac{1}{2}}\theta}{1 + s_0^{\frac{1}{2}}\theta} \right)^2 (k - k_{G,F}) \|\nabla f(\mathbf{x}_k^*)\|^2. \end{aligned} \tag{80}$$

Combine (80) and (78) to see that for $k \geq \tilde{k}_2$,

$$\|\nabla f(\mathbf{x}_k^*)\|^2 \leq \left(\frac{1 - s_0^{\frac{1}{2}}\theta}{1 + s_0^{\frac{1}{2}}\theta} \right)^2 \left(\frac{1 + s_0^{1/2}\theta}{\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2} \right) \frac{L}{\gamma} \left(f(\mathbf{x}_{k_{G,F}}) - f^* \right) (k - k_{G,F})^{-1}. \tag{81}$$

Now, the second assertion of Lemma 3 implies that for $k \geq k_{G,F}$

$$n(\mathbf{x}_k) \leq 2I^* \max\left(\eta_k, 1 + \max\left(c_A(\mathbf{x}_k)\,, c_A^*(\mathbf{x}_k)\right) \max\left(1, \|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}\right)\right). \tag{82}$$

The inequality in (81), however, implies that the term $\|\nabla f(\mathbf{x}_k^*)\|^{-\frac{1}{\mu_A(\alpha)-\delta}}$ appearing in (82) diverges at a rate faster than the lower bound sequence $\{\eta_k\}$ (which has been assumed to diverge slower than $k^{\frac{1}{2(\mu_A(\alpha)-\delta)}}$). We thus see that there exists $\tilde{k}_3$ such that for $k \geq \tilde{k}_3$,

$$n(\mathbf{x}_k) \leq 2I^* \left(1 + \max\left(c_A(\mathbf{x}_k)\,, c_A^*(\mathbf{x}_k)\right) \max\left(1, \|\nabla f(\mathbf{x}_k)\|^{-\frac{1}{\mu_A(\alpha)-\delta}}\right)\right). \tag{83}$$

This proves the fourth assertion of the theorem.

For proving the fifth assertion, we use (78) to write

$$\|\nabla f(\mathbf{x}_k^*)\|^2 \leq \left(\frac{1 + s_0^{1/2}\theta}{\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2}\right) \frac{L}{\gamma} \left(f(\mathbf{x}_{k_{G,F}}) - f^*\right)(k+1)^{-1}. \tag{84}$$

Also, due to $\sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|^2 < \infty$, (75), and the assumed continuity of functions $\Gamma_f(\cdot), \Gamma_g(\cdot)$, there exists constants $\tilde{k}_4, c_A$, and c_A^* such that for all $k \geq \tilde{k}_4$, $c_A(\mathbf{x}_k) \leq c_A$ and $c_A^*(\mathbf{x}_k) \leq c_A^*$. This, and the inequality in (82) imply that for $k \geq \tilde{k}_5 := \max(\tilde{k}_2, \tilde{k}_3, \tilde{k}_4)$, the total work $w_k = \sum_{i=0}^{k} n(\mathbf{x}_i)$ satisfies

$$\begin{aligned} w_k &\leq w_{\tilde{k}_5} + (2 - \frac{\log s_0 L}{\log \gamma}) \sum_{i=\tilde{k}_5}^{k} (1 + \max(c_A\,, c_A^*)\|\nabla f(\mathbf{x}_i)\|^{-1/\mu_A(\alpha)-\delta}), \\ &\leq w_{\tilde{k}_5} + (2 - \frac{\log s_0 L}{\log \gamma}) \left(1 + \max(c_A\,, c_A^*)\|\nabla f(\mathbf{x}_k)\|^{-1/\mu_A(\alpha)-\delta}\right)(k+1). \end{aligned} \tag{85}$$

Combining (84) and (85), we conclude that the fifth assertion of the theorem holds with $\chi_u = (w_{\tilde{k}_5} + 2 - \frac{\log s_0 L}{\log \gamma})((\frac{1+s_0^{1/2}\theta}{\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2})\frac{L}{\gamma}(f(\mathbf{x}_{k_{G,F}}) - f^*))^{\frac{1}{2}(2+1/\mu_A(\alpha)-\delta)} + (2 - \frac{\log s_0 L}{\log \gamma}) \max(c_A\,, c_A^*)((\frac{1+s_0^{1/2}\theta}{\frac{1}{2} - s_0^{\frac{1}{2}}\theta - 4\theta^2})\frac{L}{\gamma}(f(\mathbf{x}_{k_{G,F}}) - f^*))$. □

An implication of Theorem 2 is that the work complexity of backtracking line search ASGM is $\mathcal{O}\left(\epsilon^{-2-\frac{1}{\mu_A(\alpha)-\delta}}\right)$, where $\mu_A(\alpha) = \min(\alpha/2, \mu(\alpha))$ and δ is any (arbitrarily small) positive number. So, we see that this complexity result for backtracking line search ASGM matches the corresponding complexity for ASGM with fixed step size except that the constant $\mu(\alpha)$ in the fixed step size context is replaced by the constant $\mu_A(\alpha)$.

Backtracking line search ASGM is a "practical" algorithm in that it assumes no knowledge about any structural constants of the function f. Furthermore, Lemma 3 and Theorem 2 together provide a characterization that is based on a complete "book-keeping" of the operations within backtracking line search ASGM.

4.2 ASGM Convergence and Work Complexity when $f \in \mathcal{S}^{1,1}_{\lambda,L}$

In this section we assume $f \in \mathcal{S}^{1,1}_{\lambda,L}$, that is, f is smooth and strongly convex. Such problems frequently arise in classification problems [9, 8] where f is the sum of a convex function and a regularization term. As noted in [8], a consideration of $\mathcal{S}^{1,1}_{\lambda,L}$ is important also because it sheds light on the *local* behaviour of ASGM, that is, the behaviour of ASGM in the vicinity of first-order critical point when implemented on a smooth (and potentially non-convex) function f.

Since $f(\cdot)$ has been assumed to be bounded below, $f \in \mathcal{S}^{1,1}_{\lambda,L}$ is guaranteed to attain its unique minimum

$$\mathbf{x}^* = \arg\inf\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d\}. \tag{86}$$

We now present an analysis of fixed step ASGM through Theorem 3 which asserts ASGM's iterates converge to the minimum $\mathbf{x}^*$ as $k \to \infty$. Theorem 3 also characterizes the iteration complexity and the work complexity associated with generating such a sequence.

Theorem 3 (Fixed Step ASGM for $f \in \mathcal{S}^{1,1}_{\lambda,L}$) *Suppose that $f \in \mathcal{S}^{1,1}_{\lambda,L}$ and bounded below. Let ASGM be applied on f with fixed step size $\beta_k = \beta \leq L^{-1}$ and oracle effort rule as in (SS_c) with $\theta \in (0, 1/2)$. Let the lower bound sequence $\{\eta_k\}$ in (SS_c) be such that $\eta_k \to \infty$ and $\eta_k = o^{-1}\left(k^{\frac{1}{2\mu(\alpha)-2\delta}}\right)$ as $k \to \infty$. Define n_G such that for all $n \geq n_G$:*

$$\|g(n;\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \Gamma_g(\mathbf{x})\, n^{-\mu(\alpha)+\delta}. \tag{87}$$

and $k_G := \min\{k : \eta_k \geq n_G\}$. Then the following assertions hold.

1. *Defining $r := \frac{1}{2}\lambda\beta(1-2\theta)(1+\theta)^{-2}$, for all $k \geq k_G$,*

$$\left(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right) \leq (1-r)\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right). \tag{88}$$

2. *For all $k \geq k_G$,*

$$\frac{\|\nabla f(\mathbf{x}_k)\|^2}{(1-r)^k} \leq \frac{2L^2}{\lambda}\,\frac{f(\mathbf{x}_{k_G}) - f(\mathbf{x}^*)}{(1-r)^{k_G-1}} =: s_u. \tag{89}$$

3. *For $k \geq k_G$, $n(\mathbf{x}_k) \geq \left(\frac{1-\theta}{\theta}\right)^{\frac{1}{\mu(\alpha)-\delta}} \left(s_u\,(1-r)^k\right)^{-\frac{1}{2(\mu(\alpha)-\delta)}}$.*
4. *There exist constants $\chi_{\ell,S}, \chi_{u,S} \in (0,\infty)$ such that the total oracle effort $w_k = \sum_{j=1}^{k} n(\mathbf{x}_j)$ expended until the kth iteration of fixed step ASGM satisfies*

$$\chi_{\ell,S} < w_k\,\|\nabla f(\mathbf{x}_k)\|^{\frac{1}{\mu(\alpha)-\delta}} < \chi_{u,S}. \tag{90}$$

Proof Employing standard arguments on functions belonging to $\mathcal{S}^{1,1}_{\lambda,L}$, we get

$$\lambda\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right) \leq \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{2L^2}{\lambda}\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right). \tag{91}$$

Also, we have argued in the proof of Theorem 1 that if $k \geq k_G$, then

$$(1+\theta)^{-1}\|\nabla f(\mathbf{x}_k)\| \leq \|g(n(\mathbf{x}_k);\mathbf{x}_k)\| \leq (1-\theta)^{-1}\|\nabla f(\mathbf{x}_k)\|. \tag{92}$$

By (40) and (91), we see that for $k \geq k_G$,

$$\left(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right) \leq (1-r)\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right), \tag{93}$$

and the first assertion of the theorem holds.

Use the right-hand side of (91) and the first assertion of the theorem to conclude that the second assertion of the theorem holds as well.

For proving the third assertion, observe from the stopping rule (SS_ℓ) employed in fixed step ASGM that

$$\Gamma_g(\mathbf{x}_k) n(\mathbf{x}_k)^{-\mu(\alpha)+\delta} \leq \theta \|g(n(\mathbf{x}_k); \mathbf{x}_k)\|. \tag{94}$$

Noting that $\Gamma_g(\mathbf{x}_k) > 1$, and combining (92), (93), and (94), we see that for $k \geq k_G$,

$$n(\mathbf{x}_k) \geq \left(\frac{1-\theta}{\theta}\right)^{\frac{1}{\mu(\alpha)-\delta}} \|\nabla f(\mathbf{x}_k)\|^{-\frac{1}{\mu(\alpha)-\delta}} \geq \left(\frac{1-\theta}{\theta}\right)^{\frac{1}{\mu(\alpha)-\delta}} \left(s_u (1-r)^k\right)^{-\frac{1}{2(\mu(\alpha)-\delta)}}. \tag{95}$$

Let's now prove the fourth assertion of the theorem. We see that the left-hand side of (90) holds with $\chi_{\ell,S} = (\theta^{-1}-1)^{\frac{1}{\mu(\alpha)-\delta}}$ from the first inequality in (95) and since $n(\mathbf{x}_k) \leq w_k$. For proving the right-hand side of (90), we first notice from the third assertion that $n(\mathbf{x}_k)$ grows faster than geometrically, and hence, for some $\tilde{c} \in (1,\infty)$, $\sum_{j=0}^{k} n(\mathbf{x}_j) \leq \tilde{c}\, n(\mathbf{x}_k)$. In addition, similar to (48), since (89) holds and $\eta_k = o^{-1}\left(k^{\frac{1}{2\mu(\alpha)-2\delta}}\right)$, there exists $\tilde{k} > 0$ such that for all $k > \tilde{k}$, $n(\mathbf{x}_k) \leq (1+c_0)\|\nabla f(\mathbf{x}_k)\|^{-\frac{1}{\mu(\alpha)-\delta}}$. Taking $\tilde{\tilde{k}} = \max\{\tilde{k}, k_G\}$, we then have $w_k \leq \sum_{j=1}^{\tilde{\tilde{k}}} n(\mathbf{x}_j) + \tilde{c}\,(1+c_0)\|\nabla f(\mathbf{x}_k)\|^{-\frac{1}{\mu(\alpha)-\delta}}$, thus proving the right-hand side of (90) with $\chi_{u,S} = s_u^{\frac{1}{2(\mu(\alpha)-\delta)}}\left(\sum_{j=1}^{\tilde{\tilde{k}}} n(\mathbf{x}_j)\right) + \tilde{c}\,(1+c_0)$. □

Theorem 3 notes that that the work complexity of fixed step ASGM when $f \in \mathcal{S}_{\lambda,L}^{1,1}$ is $\mathcal{O}\left(\epsilon^{\frac{1}{\mu(\alpha)-\delta}}\right)$, where $\delta > 0$ is any (arbitrarily small) positive number. This assurance implies a faster convergence rate than what was guaranteed when $f \in C_L^{1,1}$; such acceleration is understandable given the stronger structure that comes with f being smooth and strongly convex. The rate specified in Theorem 3 is still substantially slower than the linear rate $\mathcal{O}(-\log\epsilon)$ achieved in presence of an exact oracle [26]. The inability of ASGM's iterates to achieve a linear rate is rooted in the polynomial decay of error implicit in (FE_1).

We end this section by stating a result analogous to Theorem 3 but for backtracking line search ASGM. Since the method of proof provides no insight over and above the analysis that was detailed for the corresponding context when $f \in C_L^{1,1}$, we omit a proof.

Theorem 4 (Backtracking Line Search ASGM for $f \in \mathcal{S}_{\lambda,L}^{1,1}$) *Let ASGM with backtracking line search procedure Algorithm 1 be applied on the function f with constants $\theta \in (0,1)$, $s_0 > 0$, $c_F \in (0,1)$ satisfying $s_0^{1/2}\theta < 1$, $\theta \leq \frac{1}{4}(\sqrt{s_0+4} - \sqrt{s_0})$, and $c_F = \frac{1}{2} - s_0^{1/2}\theta - 2\theta^2$. Furthermore, let $\{\eta_k\} \to \infty$, $\eta_k = o^{-1}\left(k^{\frac{1}{2\mu(\alpha)-2\delta}}\right)$ as $k \to \infty$. Then, given $k_{G,F}$ as defined in Lemma 3, the following assertions hold.*

1. *Let* $r := \frac{\gamma\lambda}{L}\left(\frac{\frac{1}{2}-s_0^{\frac{1}{2}}\theta-4\theta^2}{1+s_0^{1/2}\theta}\right) < 1$ *and* $\mathbf{x}^*$ *as in (86), we have for all* $k \geq k_{G,F}$,

$$\left(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right) \leq (1-r)\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right).$$

2. *For all* $k \geq k_{G,F}$,

$$\frac{\|\nabla f(\mathbf{x}_k)\|^2}{(1-r)^k} \leq \frac{2L^2}{\lambda}\frac{f(\mathbf{x}_{k_{G,F}}) - f(\mathbf{x}^*)}{(1-r)^{k_{G,F}-1}} =: s_u.$$

3. *There exist* $\chi_{\ell,S}, \chi_{u,S} \in (0,\infty)$ *such that the total oracle effort* $w_k = \sum_{j=1}^{k} n(\mathbf{x}_j)$ *satisfies* $\chi_{\ell,S} < w_k \|\nabla f(\mathbf{x}_k)\|^{\frac{1}{\mu(\alpha)-\delta}} < \chi_{u,S}$.

5 Concluding Remarks

ASGM is a procedure that mimics gradient search for unconstrained optimization in settings where the objective function can only be approximated using an inexact oracle such as quasi-Monte Carlo or numerical quadrature. ASGM accommodates both derivative-based and derivative-free contexts, the latter through the use of approximated derivatives, e.g., by finite-differencing function values obtained using the inexact oracle. ASGM's strength is its incorporation of adaptive sampling — it exploits knowledge of error bounds on the approximates to decide how much oracle effort to expend during each iteration. Specifically, ASGM expends just enough oracle effort to ensure that the approximated gradient norm at the incumbent iterate is within a fixed constant of the norm of the error in the approximated gradient at the incumbent iterate. Three other remarks pertaining to ASGM are in order.

1. As Table 1 notes, ASGM's work complexity is superior (when using QMC) to that of corresponding methods such as SGD due to the faster inherent rate of QMC compared to Monte Carlo. Such dominance should warrant ASGM with QMC as a serious contender to methods such as SGD.
2. ASGM's complexity rates are poorer than when an exact oracle for the function f is available and used within corresponding gradient search algorithms. The difference in complexities, as displayed in Table 1 clearly reflect the "price" of not having an exact oracle.
3. The methods we have outlined in this paper provide a framework of analysis for obvious extensions that incorporate second order information, functional constraints, and contexts where the function f is non-smooth.

References

1. S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York, NY., 2007.
2. S. Bhatnagar, N. Hemachandra, and V. Mishra. Stochastic approximation algorithms for constrained optimization via simulation. *ACM Transactions on Modeling and Computer Simulation*, 21(3):15:1–15:22, 2011.

3. S. Bhatnagar, N. Hemachandra, and V. Mishra. Stochastic approximation algorithms for discrete parameter simulation optimization. *IEEE Transactions on Automation Science and Engineering*, 8(4):780–793, 2011.
4. M. Bianchetti, S. Kucherenko, and S. Scoleri. Pricing and risk management with high-dimensional quasi monte carlo and global sensitivity analysis. 2015. Preprint.
5. P. Billingsley. *Probability and Measure*. Wiley, New York, NY., 1995.
6. J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, NY., 1997.
7. V. S. Borkar. Stochastic approximation with two time scale. *Systems and Control Letters*, 29:291–294, 1997.
8. L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. 2016.
9. S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–358, 2015.
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. D. Cruz-Uribe and C. J. Neugebauer. Sharp error bounds for the Trapezoidal Rule and Simpson's Rule. *Journal of Inequalities in Pure and Applied Mathematics*, 3(4), 2002.
12. P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Dover Publications, Inc., Mineola, New York, 1984.
13. P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, NY., 2004.
14. S. G. Henderson and B. L. Nelson, editors. volume 13 of *Handbooks in Operations Research and Management Science: Simulation*. Elsevier, 2006.
15. A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
16. S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, NY., 1975.
17. J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.
18. H. Kushner and D. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, NY., 1978.
19. H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York, NY., 2003.
20. H. M. Markowitz, G. P. Todd, and W. F. Sharpe. *Mean-variance analysis in portfolio choice and capital markets*. Wiley, New York, NY, 2000.
21. A. Mokkadem and M. Pelletier. A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49:1523, 2011.
22. W. J. Morokoff and R. E. Caflisch. Quasi-monte carlo integration. *Journal of Computational Physics*, 122(2):218–230, 1995.
23. W. Nasr. *Analysis and Approximations for Time Dependant Queueing Models*. PhD thesis, School of Industrial Engineering, Purdue University, West Lafayette, IN., 2008.

24. 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.
25. A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, New York, NY, 1983.
26. Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science + Business Media, LLC, New York, NY, 2004.
27. H. Niederreiter. *Random Number Generation and Quasi-Monte-Carlo methods*. SIAM, Philadelphia, PA., 1992.
28. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, Berlin, 2006.
29. S. H. Paskov and J. F. Traub. Faster evaluation of financial derivatives. *Journal of Portfolio Management*, 22(1):113–120, 1995.
30. R. Pasupathy. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58:889–901, 2010.
31. R. Pasupathy and S. Ghosh. Simulation optimization: A concise overview and implementation guide. INFORMS TutORials. INFORMS, 2013.
32. R. Pasupathy and S. Kim. The stochastic root-finding problem: overview, solutions, and open questions. *ACM TOMACS*, 21(3), 2011.
33. R. Pasupathy and B. W. Schmeiser. Retrospective-approximation algorithms for multidimensional stochastic root-finding problems. *ACM TOMACS*, 19(2):5:1–5:36, 2009.
34. 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.
35. B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
36. A. Ralston and P. Rabinowitz. *First Course in Numerical Analysis*. McGraw-Hill Book Company, New York, 1978.
37. H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
38. S. Ross. *Stochastic Processes*. Wiley, New York, NY., 1995.
39. A. Ruszczynski and A. Shapiro, editors. *Stochastic Programming. Handbook in Operations Research and Management Science*. Elsevier, New York, NY., 2003.
40. P. Sabino. Efficient quasi-monte simulations for pricing high-dimensional path-dependent options. *Decisions in Economics and Finance*, 32(1):49–65, 2009.
41. A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, PA, 2009.
42. J. C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., Hoboken, NJ., 2003.
43. H. Woźniakowski. Average case complexity of multivariate integration. *Bulletin of the American Mathematical Society*, 24(1):185–194, 1991.