Approaching Quartic Convergence Rates for Quasi-Stochastic Approximation with Application to Gradient-Free Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic approximation is a foundation for many algorithms found in machine learning and optimization. It is in general slow to converge: the mean square error vanishes as $O(n^{-1})$. A deterministic counterpart known as quasi-stochastic approximation (QSA) is a viable alternative in many applications, including gradient free optimization and reinforcement learning. It was assumed in recent prior research that the optimal achievable convergence rate is $O(n^{-2})$. It is shown in this paper that through design it is possible to obtain far faster convergence, of order $O(n^{-4+\delta})$, with $\delta > 0$ arbitrary. Two acceleration techniques are introduced for the first time to achieve this rate of convergence. The theory is also specialized within the context of gradient-free optimization, and tested on standard benchmarks. The main results are based on a combination of recent results from number theory and techniques adapted from stochastic approximation theory.

1 Introduction

Stochastic approximation (SA) was introduced in the seminal work of Robbins and Monro [25]. The goal is to solve the root finding problem $\bar{f}(\theta^*) = 0$, in which $\bar{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ is of the form

$$\bar{f}(\theta) := \mathrm{E}[f(\theta, \zeta)] \tag{1}$$

where ζ is a random vector taking values in $\mathbb{R}^m$. The basic algorithm is expressed as the d-dimensional recursion,

$$\theta_{n+1} = \theta_n + \alpha_{n+1} f(\theta_n, \zeta_{n+1}), \qquad n \geq 0, \tag{2}$$

in which $\{\alpha_n\}$ is the non-negative step-size sequence, and $\zeta_{n+1} \xrightarrow{\mathrm{d}} \zeta$ as $n \to \infty$.

The field has attracted a great deal of attention over the past twenty years, motivated in large part by applications to reinforcement learning and optimization [34, 13, 22, 11, 6].

Convergence theory is couched in the *ODE Method* in which trajectories of (2) are compared to solutions of the ODE $\dot{x} = \bar{f}(x)$. The major assumption required to ensure convergence to θ^*, for each initial condition $\theta_0 \in \mathbb{R}^d$, is that the ODE is globally asymptotically stable—see [6] for minimal assumptions on the "noise sequence" $\boldsymbol{\zeta}$. Establishing sharp rates of convergence is a far greater challenge. There is however a rich theory available to achieve the optimal rate of convergence for the mean-square error (MSE), which is in general $\mathrm{E}[\|\theta_n - \theta^*\|^2] = O(n^{-1})$.

There are many applications for which the designer of the algorithm also designs the noise. Notable examples include the introduction of exploration in reinforcement learning or gradient-free optimization. This motivates the use of *quasi-stochastic approximation* (QSA) in which the sequence $\boldsymbol{\zeta}$ is

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

deterministic (e.g. mixtures of sinusoids or pseudo-random numbers). The idea was introduced in [15, 16], but has a much longer history in the context of gradient-free optimization—see [33, 17] for a survey of *extremum seeking control* (ESC).

Theory supporting rates of convergence of QSA has appeared only recently [8, 20]. Analysis and algorithms are posed in a continuous time setting to simplify analysis, and because of the recent success stories justifying algorithm design in continuous time, followed by a careful translation to obtain a discrete time algorithm. See for example recent theory surrounding acceleration methods of Polyak and Nesterov [2, 31, 14]; it is argued in [36] that Runge-Kutta methods are sometimes valuable for an efficient translation.

The notation adopted in [20, Chap. 4] will be used here: the *QSA ODE* is defined as

$$\tfrac{d}{dt}\Theta_t = a_t f(\Theta_t, \xi_t) \tag{3}$$

The deterministic continuous time process $\boldsymbol{\xi}$ will be called the *probing signal*, and plays the role of $\boldsymbol{\zeta}$ in SA; $\boldsymbol{a}$ is called the gain process. The motivation for QSA is two-fold:

(i) It will be seen that the rate of convergence is far faster than SA, subject to careful choice of algorithm architecture.

(ii) In on-line applications the introduction of independent noise may not be realizable, or may impose unnecessary stress on equipment. In QSA design the components of the probing signal might be chosen to be sinusoidal signals of appropriate frequency and magnitude to ensure learning takes place without stress on physical devices.

What is the optimal rate of convergence for QSA? Consider the most basic one-dimensional problem in which $f(\Theta_t, \xi_t) = -\Theta_t + \xi_t$, with $\boldsymbol{\xi}$ a zero-mean signal. The special case $a_t = 1/(1+t)$ results in an approximate average:

$$\Theta_T = \frac{1}{1+T}\Theta_0 + \frac{1}{1+T}\int_0^T \xi_t \, dt$$

If for example $\xi_t = \sin(\omega t)$ then the right hand side converges at rate $O(T^{-1})$, which translates to $O(T^{-2})$ for the "MSE"; this is far faster than the rate $O(T^{-1})$ expected for SA.

The special case of pure averaging is illustrated at the top in Fig. 1. The bold question mark in the figure refers to a question posed in [20, § 4.9]: *can we obtain the same rate of convergence for general non-linear QSA*? The question is answered in the affirmative, achieved through the averaging technique of Polyak and Ruppert, but only under subtle conditions on the QSA ODE that cannot be verified a priori. A brief overview is provided in Thm. 2.2, with the main point presented here. It is first shown that the following approximation holds under mild assumptions:

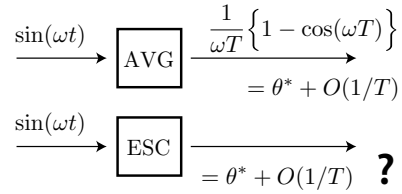

Figure 1: What is the optimal convergence rate for QSA?

$$\Theta_t = \theta^* + a_t[\bar{Y} + \Xi_t^{\mathrm{I}}] + o(a_t) \tag{4}$$

in which $\{\Xi_t^{\mathrm{I}}\}$ is a zero-mean deterministic process, and $\bar{Y} \in \mathbb{R}^d$ is fixed. Based on this bound, it is shown that Polyak-Ruppert averaging results in the bound $\|\Theta_t - \theta^*\|^2 = O(a_t^2)$ whenever $\bar{Y} \neq 0$. This averaging technique always requires $a_t \gg 1/t$, so this rate of convergence is far from $O(t^{-2})$.

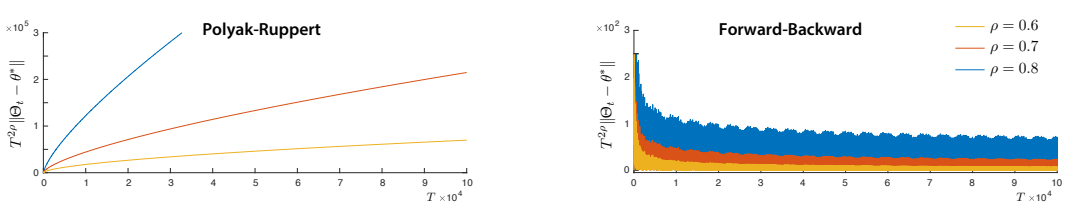

Figure 2: Rates of convergence for PR averaging and Forward-Backward algorithms.

This brings us to the main contributions of this paper:

(i) It is shown that the "ESC" block in Fig. 1 can be designed to obtain a QSA algorithm for which $\bar{Y} = 0$, so that the bound $\|\Theta_t - \theta^*\|^2 = O(t^{-2})$ is obtained.

(ii) The presumption that the rate obtained in (i) is optimal is *fallacy*. It is shown here that for any $\delta > 0$, an algorithm can be designed to achieve $\|\Theta_t - \theta^*\|^2 = O(t^{-4+\delta})$.

Theory and algorithms are developed for general root finding problems. The theory is refined in the context of gradient free-optimization, and the general theory is illustrated through numerical examples in this setting.

While theory is developed in continuous time, simulation studies using Euler approximations are consistent with theory. The plots shown in Fig. 2 are based on a two-dimensional example, whose details can be found in Section 2.2. The change of notation is also explained there: $(2\rho)^2 = 4 - \delta$, where ρ is a design parameter subject to $\rho < 1$. In this example $\bar{Y} \neq 0$ so that Polyak-Ruppert averaging fails (see (16)). The Forward-backward algorithm is a new algorithm introduced in this paper, that achieves the $O(t^{-4+\delta})$ convergence rate. The plots on the right hand side of Fig. 2 show that $n^{4-\delta}\|\theta_n - \theta^*\|^2$ is bounded in the three experiments shown, where θ_n denotes the Euler approximation at iteration n. Standard theory from Quasi-Monte Carlo (QMC) would predict a root mean square convergence rate of $O(\log(n)^d/n)$ for QSA in discrete time, since QMC is a special case (see [1, Section 9.3]). It is shown in the present paper that it is possible to obtain $O(n^{-1})$ convergence rate for QMC in discrete time—potentially $O(n^{-\sqrt{4-\delta}})$—subject to a smoothness condition on the function of $\boldsymbol{\xi}$ and careful design of the probing signal. *The main results of this paper hinge on these new bounds for QMC.*

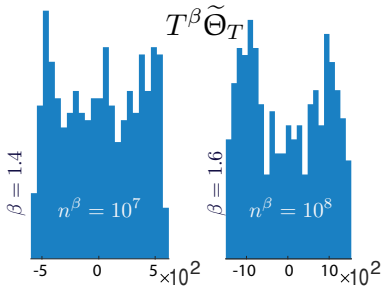

Figure 3: Quasi Monte Carlo using QSA with Polyak-Ruppert averaging. Histograms for $n = 10 \times T = 10^5$.

Fig. 3 illustrates how the techniques introduced in this paper specialize to QMC, based on two instances of the algorithm: one designed using $\beta = \sqrt{4-\delta} = 1.4$, and the other with $\beta = 1.6$; in the first case it is found empirically that $|\widetilde{\Theta}_T| := |\Theta_T - \theta^*| \leq 10^3 T^{-1.4}$ for all T, provided $|\widetilde{\Theta}_0| \leq 25$. Details are found in the supplementary material.

Literature Review: QMC remains an active area of research for applications to estimation and optimization [23]. In standard QMC methodology, a sequence $\{\xi_t : t \geq 0\} \subset H$ (the d-dimensional hypercube $H = [0,1]^d$) is constructed so that sample path averages are convergent: $n^{-1}\sum_1^n f(\xi_t) \to \langle f \rangle := \int f(x)\,dx$ as $n \to \infty$. The new bounds on the rate of convergence obtained here are based on recent refinements of Baker's Theorem [19, 7].

QSA was first applied to finance within the context of QMC [15, 16]. Consistency and bounds on the rate of convergence came much later—see [20, § 4.9] for recent results and history. In particular, in this recent work the convergence rate $\|\Theta_t - \theta^*\|^2 = O(t^{-2})$. was obtained subject to stringent conditions on the model, which was thought to be the best possible convergence rate without imposing additional assumptions.

Gradient-Free Optimization (GFO) concerns minimization of an objective function $\Gamma : \mathbb{R}^d \to \mathbb{R}$ based solely on measurements of $\Gamma(\theta)$ for selected values of $\theta \in \mathbb{R}^d$. This is motivated by problems in which access to gradient information is computationally intractable or not available (such as in model-free online optimization). The algorithms of Keifer and Wolfowitz are probably the first approaches of this kind [12, 1]. Many improvements followed, with much research led by Spall [27, 28, 29], such as the Simultaneous Perturbation Stochastic Approximation (1SPSA and 2SPSA) algorithms of the form (2) with

$$f^{\text{1SPSA}}(\theta_n, \zeta_n) = \frac{1}{\varepsilon}\zeta_{n+1}\Gamma(\theta_n + \varepsilon\zeta_{n+1}) \tag{5a}$$

$$f^{\text{2SPSA}}(\theta_n, \zeta_n) = \frac{1}{2\varepsilon}\big(\zeta_{n+1}[\Gamma(\theta_n + \varepsilon\zeta_{n+1}) - \Gamma(\theta_n - \varepsilon\zeta_{n+1})]\big) \tag{5b}$$

where $\varepsilon > 0$ and $\{\zeta_n\}$ is a zero-mean i.i.d. sequence. See also [5, 3, 10].

Deterministic versions of SPSA were analyzed in [4] without sharp rates of convergence. The present paper follows the approach of [20, § 4.9], in which the applications to gradient-free optimization result in algorithms that resemble simple versions of the ESC algorithms surveyed in [33, 17].

Notation Throughout the paper we restrict to sinusoidal probing signals,

$$\xi_t = \sum_{i=1}^{K} v^i \cos(2\pi[\,\omega_i t + \phi_i]) \tag{6}$$

for which $\{\phi_i\}$ are arbitrary, and assumptions on the frequencies $\{\omega_i\}$ will be imposed in the main results. The vectors $\{v^i : 1 \le i \le K\}$ are are assumed to span $\mathbb{R}^m$.

Let Φ_t denote the function of time evolving on $\mathbb{C}^K$ with entries $\Phi_t^i = \exp(2\pi j[\omega_i t + \phi_i])$.

For two real-valued functions of time h, g we denote

$$\langle h, g\rangle_T := \int_0^T h_t g_t \, dt\,, \qquad \langle h\rangle \;:= \lim_{T\to\infty} \frac{1}{T}\int_0^T h_t\, dt\,, \quad \tilde{h} = h - \langle h\rangle\,, \qquad h_T^{\mathrm{I}} := \int_0^T \tilde{h}_t\, dt - \bar{h}^{\mathrm{I}} \tag{7a}$$

where $\bar{h}^{\mathrm{I}}$ is a constant chosen so $\langle h^{\mathrm{I}}\rangle = 0$. When h is a function of time through the probing signal, $h_t \equiv h(\xi_t)$, then we adopt the alternative notation $\overline{h} := \langle h(\xi)\rangle$. The bracket notation is also used for vector or matrix valued functions. In particular, we denote $\Sigma_\xi := \langle \xi\xi^{\mathsf{T}}\rangle$, a $K \times K$ matrix. Thm. 2.1 provides conditions ensuring the existence of sample path averages and boundedness of $\{h_T^{\mathrm{I}} : T > 0\}$.

For a scalar-valued function $\delta_t > 0$, we use the notation $O(\delta_t)$ to denote a vector-valued function of t; the notation indicates that there is a constant B such that $\|O(\delta_t)\| \le B\delta_t$ for all t. That is, we are only asserting an *upper bound*. For example, if $\delta_t = 1/(1+t)$ we are not claiming that $\|O(\delta_t)\| \le B\delta_t^2$ is not possible.

Organization: Section 2 surveys recent work and summarizes the main contributions of the paper. Applications to GFO in Section 3 establish new bounds on both bias and convergence rate, along with numerical results illustrating the theory. Conclusions and directions for future research are summarized in Section 4. Technical proofs and additional numerical results can be found in the supplemental material.

2 Towards Quartic Rates

2.1 Probing Design for QMC

Thm. 2.1 below justifies boundedness of the integral (7a) and much more. Through careful design of the frequencies $\{\omega_i\}$ appearing in (6) we can apply refinements of Baker's Theorem, as surveyed in the monograph [7]. While this theory provides many options, we restrict here to a special case:

$$\begin{aligned} &\omega_i = \log(a_i/b_i) > 0\,,\;\; 1 \le i \le K\,,\;\; \textit{where } \{a_i, b_i\} \textit{ are positive integers.} \\ &\{\omega_i : 1 \le i \le K\} \textit{ are linearly independent over the field of rational numbers.} \end{aligned} \tag{8}$$

Under this assumption $\Sigma_\xi := \langle \xi\xi^{\mathsf{T}}\rangle = \frac{1}{2}\sum v^i v^{i\,\mathsf{T}}$.

Theorem 2.1. *Suppose that the function $h\colon \mathbb{R}^d \to \mathbb{R}$ is analytic in a neighborhood of the unit hypercube $[-1,1]^d$, and that ξ_t in* (6) *satisfies* (8)*. Then the limit defining $\langle h\rangle$ in* (7a) *exists with $h_t = h(\xi_t)$, and the following ergodic bounds hold for some constant K_f independent of the phase values $\{\phi_i\}$:*

$$\Big|\frac{1}{T}\int_0^T \tilde{h}(\xi_t)\, dt\Big| \le K_f \frac{1}{T}\,, \qquad T > 0 \qquad \textit{where } \tilde{h} = h - \langle h\rangle. \tag{9}$$

Moreover, there is a function $\hat{h}\colon \mathbb{C}^d \to \mathbb{C}^d$ that is analytic in a neighborhood of the domain $\Omega := \{z \in \mathbb{C}^d : |z_i| = 1\,,\; 1 \le i \le d\}$ satisfying

$$\begin{aligned} &\int_0^T \tilde{h}(\xi_t)\, dt = \hat{h}(\Phi_0) - \hat{h}(\Phi_T)\,, \quad T \ge 0\,, \qquad \bar{h}^{\mathrm{I}} = \hat{h}(\Phi_0) \\ &\textit{and} \quad \langle \hat{h}(\Phi)\rangle = \lim_{T\to\infty}\frac{1}{T}\int_0^T \hat{h}(\xi_t)\, dt = 0 \end{aligned} \tag{10}$$

The theorem provides conditions ensuring the existence of $\bar{h}^{\mathrm{I}}$ and boundedness of $\{h^{\mathrm{I}}_T : T > 0\}$ as defined in (7a). Since $\hat{h}$ is also analytic we are assured of multiple integrals that are also bounded in time. We will require both in the analysis supporting our main results:

$$h^{\mathrm{I}}_T = \int_0^T \tilde{h}(\xi_t)\, dt - \bar{h}^{\mathrm{I}} = -\hat{h}(\Phi_T)\,, \qquad h^{\mathrm{II}}_T = \int_0^T h^{\mathrm{I}}_t dt - \bar{h}^{\mathrm{II}} \tag{11}$$

The proof of Thm. 2.1 is contained in the supplementary material, along with extensions such as the alternative choice of triangle waves—see Thm. C.1.

A version of (9) in discrete time is contained in Thm. B.9. This result is a big surprise, given that there is so much theory predicting $O(\log(T)^d/T)$ bounds. Much more surprising is that $O(1/T)$ is a very poor rate of convergence for the algorithms developed in this paper, as observed in Fig. 3. This is explained in Section C.1 of the supplementary material, along with examples to show how the theory can be applied to obtain convergence rates of order $O(T^{-2+\delta})$ for specially designed QSA algorithms, and in particular new approaches to QMC.

A few key ideas are presented here: The assumption that h is analytic is imposed so that we can first restrict to complex exponentials, $H(\Phi_t) = \exp(2\pi j t\omega t)$, whose integral is $H^{\mathrm{I}}_t = (2\pi j\omega)^{-1} H(\Phi_t)$ when $\omega \neq 0$. In the proof of Thm. 2.1 we use $\omega = \sum_{i=1}^d n_i\omega_i$ with $\{n_i\}$ integers; they are not necessarily positive, but at least one n_i is assumed non-zero. Extensions of Baker's Theorem, as surveyed in [7], give a strict lower bound of the form $|\omega| \geq \delta n^{-c}$, where $n = 3 + \sum |n_i|$ and δ, c are non-negative constants that are independent of n—summarized in Thm. B.6. This combined with routine Taylor series bounds establishes the desired conclusions. There is unfortunate news when we search for bounds on K_f as a function of dimension. This bound is a function of constants appearing in Thm. B.6, for which present upper bounds grow at least exponentially in d. This *curse of dimensionality* requires application of block descent strategies for applications to optimization in high dimension.

2.2 Quasi-Stochastic Approximation

Before delving into the main contributions, we summarize the main convergence results for QSA from [20, § 4.9]. For all $\theta \in \mathbb{R}^d$, the averaged vector field $\bar{f}$ is defined by the sample path average,

$$\bar{f}(\theta) := \lim_{T\to\infty} \frac{1}{T}\int_0^T f(\theta, \xi_t)\, dt \tag{12}$$

Similar to standard analysis of SA, convergence of the solution of (3) to θ^* is established by comparison with the ODE,

$$\tfrac{d}{dt}\bar{\Theta}_t = a_t \bar{f}(\bar{\Theta}_t), \quad \text{for } t \geq t_0 \text{ and } \bar{\Theta}_{t_0} = \Theta_{t_0}. \tag{13}$$

In SA theory the time t_0 is also a variable, while in QSA theory it is fixed but chosen suitably large for analysis of convergence rates.

The full list of assumptions (QSA1)–(QSA5) may be found in the supplementary material. We settle for a brief summary here: (QSA1) concerns the gain process. In the body of the paper we take $a_t = a_0(1+t)^{-\rho}$ for $\rho \in (1/2, 1]$ and a constant $a_0 > 0$. (QSA2) imposes global Lipschitz bounds on f and $\bar{f}$. (QSA3): the ODE $\frac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t)$ is globally asymptotically stable. (QSA4): $\bar{f}$ is continuously differentiable, $A(\theta) = \partial_\theta \bar{f}(\theta)$ is bounded and Lipschitz continuous, and $A^* = A(\theta^*)$ is Hurwitz. (QSA5) includes existence of functions $\hat{f}$ and $\widehat{A}$ that are Lipschitz continuous in θ solving for each $0 \leq t_0 \leq t_1$,

$$\hat{f}(\theta, \Phi_{t_0}) = \int_{t_0}^{t_1} \tilde{f}(\theta, \xi_t)\, dt + \hat{f}(\theta, \Phi_{t_1}) \qquad \widehat{A}(\theta, \Phi_{t_0}) = \int_{t_0}^{t_1} \tilde{A}(\theta, \xi_t)\, dt + \widehat{A}(\theta, \Phi_{t_1}) \tag{14}$$

where $\tilde{f}(\theta, \xi_t) = f(\theta, \xi_t) - \bar{f}(\theta)$, $\tilde{A}(\theta, \xi_t) = A(\theta, \xi_t) - A(\theta)$ with $A(\theta, \xi) := \partial_\theta f(\theta, \xi)$.

A well known approach to acceleration is Polyak-Ruppert (PR) averaging [26, 24, 9, 22, 21]:

$$\Theta^{\mathsf{PR}}_T = \frac{1}{T - T_0}\int_{T_0}^T \Theta_t\, dt \tag{15}$$

where the interval $[0, T_0]$ is known as the burn-in period. This is viewed as a linear time-varying system, and we sometimes opt for the notation $\Theta_T^{\mathsf{PR}} = \langle \beta_{T,\cdot}^{\mathsf{PR}}, \Theta \rangle_T$, with $\beta_{T,t}^{\mathsf{PR}} = 1/(T - T_0)$ on $[T_0, T]$ and zero elsewhere.

General expressions for Θ_t and Θ_T^{PR} obtained in [20, § 4.9] are summarized by the following:

Theorem 2.2. *Suppose that (QSA1)–(QSA5) hold and the following additional assumptions are imposed: $\{\Theta_t\}$ is obtained using $a_t = 1/(1+t)^\rho$ with $\rho \in (1/2, 1)$, and PR averaging is obtained using $T_0 = (1 - 1/\kappa)T$ with $\kappa > 1$ fixed. Then there is a vector $\bar{Y} \in \mathbb{R}^d$ and a bounded zero-mean signal $\{\Xi_t^{\mathsf{I}}\}$ such that*

$$\Theta_t = \theta^* + a_t[\bar{Y} + \Xi_t^{\mathsf{I}}] + o(a_t)\,, \qquad \Theta_T^{\mathsf{PR}} = \theta^* + a_T c(\kappa, \rho)\bar{Y} + \mathcal{B}_T/T \tag{16}$$

where $\{\mathcal{B}_T\}$ is a bounded process, and

$$c(\kappa, \rho) = \kappa[1 - (1 - 1/\kappa)^{1-\rho}]/(1 - \rho) \qquad \Xi_t^{\mathsf{I}} = \int_0^t f(\theta^*, \xi_r)\, dr$$

It follows from Thm. 2.2 that $\|\Theta_t - \theta^*\| = O(a_t)$, and $\|\Theta_T^{\mathsf{PR}} - \theta^*\| = O(T^{-1})$ if and only if $\bar{Y} = 0$.

The mysterious vector $\bar{Y}$ is defined in terms of another sample path average:

$$\bar{Y} = [A^*]^{-1}\bar{\Upsilon}, \text{ where } \quad \bar{\Upsilon} = -\lim_{T\to\infty} \frac{1}{T}\int_0^T \left\{ \int_0^t [A(\theta^*, \xi_r) - A^*]\, dr \right\} f(\theta^*, \xi_t)\, dt \tag{17}$$

Existence comes from (14) and (QSA5). Nothing like $\bar{Y}$ appears in standard SA theory.

If $A(\theta, \xi) = A(\theta)$ for all θ and ξ, then $\bar{Y} = 0$. This is a very restrictive special case, though it does hold for QMC for which $A(\theta, \xi) = -I$.

Numerical example: We demonstrate in the next subsection that the assumptions on the probing signal imposed in Thm. 2.1 are also useful for QSA algorithm design. This would appear far stronger than needed: for one, the full rank condition for Σ_ξ can be achieved under far weaker assumptions. In two dimensions, $\Sigma_\xi = \frac{1}{2}I$ for each of the following special cases: $\xi_t = [\cos(t), \cos(2t)]^\intercal$ and $\xi_t = [\cos(t), \sin(t)]^\intercal$. However, in each case we can construct a second order polynomial $h\colon \mathbb{R}^2 \to \mathbb{R}$ such that $h(\xi_t) = 0$ for all t, so that the *excitation* implied by the full rank condition is lost through a simple nonlinearity. This is only the first sign of trouble.

We consider next a numerical example designed with four positive frequencies that are linearly independent over the rationals, denoted $[\omega_{1,1}\,,\, \omega_{2,1}\,,\, \omega_{1,2}\,,\, \omega_{2,2}]$. Consider the linear QSA ODE,

$$\tfrac{d}{dt}\Theta_t = a_t f(\Theta_t, \xi_t) = (1+t)^{-\rho}[(A^* + A_t^\circ)\Theta_t + 10 b_t^\circ]$$

$$\textit{with} \quad A_t^\circ = \begin{bmatrix} 4\sin(\omega_{1,1}t) & \sin(\omega_{2,1}t) \\ \sin(\omega_{1,2}t) & 4\sin(\omega_{2,2}t) \end{bmatrix} \qquad b_t^\circ = \begin{bmatrix} 2\cos(\omega_{1,1}t) \\ \cos(\omega_{2,2}t) \end{bmatrix} \qquad \textit{and } A^* \textit{ Hurwitz.}$$

We have $\bar{f}(\theta) = A^*\theta$ and hence $\theta^* = 0$. To fit the notation above we must take $\xi_t \in \mathbb{R}^6$, which can be chosen so that $\Sigma_\xi = \frac{1}{2}I$. The expression $\bar{Y} = -20[2/\omega_{1,1}, 1/\omega_{2,2}]^\intercal$ follows from (17).

The plots shown on the left in Fig. 2 were obtained using PR-averaging applied to this model using $A^* = -0.8I$ and frequencies $[\pi\,,\, \sqrt{3}\,,\, 4\,,\, \sqrt{5}]/5$. The convergence rate is $O(a_t)$ because $\bar{Y} \neq \mathsf{0}$.

This bad news is easily avoided by following the guidelines in the next subsection.

2.3 Acceleration

Tighter bounds for PR averaging Recall from Section 1 that the main contribution of this paper is the introduction of algorithms to achieve MSE bounds that are arbitrarily close to $O(T^{-4})$. The next result asserts that this can be obtained using standard averaging, but only under restrictive conditions.

Proposition 2.3. *Under the assumptions of Thm. 2.2,*

$$\Theta_T^{\mathsf{PR}} = \theta^* + a_T[c(\kappa, \rho) + o(1)]\bar{Y} + O(T^{-2\rho}) \tag{18}$$

where $c(\kappa, \rho) > 0$ is defined below (16)*. Consequently, Θ_T^{PR} converges to θ^* with rate $O(T^{-2\rho})$ if and only if $\bar{Y} = 0$.*

Overview of proof: The authors of [8, 20] believed that the best possible convergence rate of $\|\Theta_T^{\mathsf{PR}} - \theta^*\|$ was $1/T$, and for this simple goal it suffices to establish boundedness of the process $\{\mathcal{B}_T\}$ appearing in Thm. 2.2. The proof of (18) begins with the decomposition $\mathcal{B}_T = [A^*]^{-1}\{\varepsilon_T^Y - \varepsilon_T^Z + \varepsilon_T^\Upsilon + \varepsilon_T^a\}$, and then lengthy computations establish that each term is bounded by $O(T^{1-2\rho} + \|\bar{Y}\|)$. Thm. 2.2 then completes the proof. □

Acceleration by probing design It is possible to design the probing signal $\boldsymbol{\xi}$ so that $\bar{Y} = 0$ for general non-linear QSA and hence fast rates can always be achieved:

Theorem 2.4. *If the assumptions of Thm. 2.2 hold and* $\boldsymbol{\xi}$ *defined in* (6) *satisfies* (8)*, then* $\bar{Y} = 0$.

Overview of proof: From (17) it suffices to show that $\bar{\Upsilon} = 0$. For this we combine (14) and (17) to obtain $\bar{\Upsilon}_i = \bar{z}_i$ for each i, where $z(\xi) = \widehat{A}(\theta^*, \xi) f(\theta^*, \xi) \in \mathbb{R}^d$ for each $\xi \in \mathbb{R}^m$. It is shown in Corollary B.5 that steady-state means of this form are always zero under the given assumptions. □

Forward-backward filtering We present a technique to achieve a near quartic MSE convergence rate without imposing additional structure on the probing signal.

Forward-backward (FB) filtering is defined by a pair of QSA ODEs:

$$\tfrac{d}{dt}\Theta_t^- = a_t f(\Theta_t^-, \xi_t^-)\,, \quad \Theta_T^{\mathsf{PR}^-} = \langle \beta_{T,\cdot}^{\mathsf{PR}}, \Theta^- \rangle_T \tag{19a}$$

$$\Theta_T^{\mathsf{FB}} = \tfrac{1}{2}[\Theta_T^{\mathsf{PR}} + \Theta_T^{\mathsf{PR}^-}] \tag{19b}$$

where $\xi_t^- := \xi_{-t}$, and Θ_T^{PR} is defined in (15).

Theorem 2.5. *If the assumptions of Thm. 2.2 hold, then,* $\Theta_T^{\mathsf{FB}} = \theta^* + O(T^{-2\rho})$.

Overview of proof: We refer to ξ_t as the forward probing signal, and ξ_t^- as backward probing. The analogous vector $\bar{Y}$ for the QSA ODE (19a) is denoted $\bar{Y}^-$. The identity $\bar{Y}^- = -\bar{Y}$ is established through consideration of (17), and the desired bound is then obtained by combining (18) with the definition (19b). □

The conclusion of Thm. 2.5 is illustrated in Fig. 2. The plots were obtained using $\kappa = 4$ and the three values $\rho \in \{0.6, 0.7, 0.8\}$. As predicted by Thm. 2.2, PR averaging fails to achieve the convergence rate of $O(T^{-2\rho})$, while FB filtering succeeds for each value of ρ tested.

3 Gradient-Free Optimization

3.1 Gradient-Free Optimization and QSA

We now turn to the application of the methods introduced by Section 2 to GFO. It is assumed that the function Γ is C^2 and that it has a unique minimizer denoted θ^{opt}. Two approaches are considered in the following: For $\varepsilon > 0$, a $d \times d$ matrix $M > 0$ and each $\Theta_0 \in \mathbb{R}^d$,

$$\textbf{1qSGD:} \qquad f^{1\mathrm{Q}}(\Theta_t, \xi_t) = -\frac{1}{\varepsilon} M \xi_t \Gamma(\Theta_t + \varepsilon \xi_t) \tag{20}$$

$$\textbf{2qSGD:} \qquad f^{2\mathrm{Q}}(\Theta_t, \xi_t) = -\frac{1}{2\varepsilon} M \xi_t [\Gamma(\Theta_t + \varepsilon \xi_t) - \Gamma(\Theta_t - \varepsilon \xi_t)] \tag{21}$$

where $K = m = d$ in (6). These are the natural analogs of Spall's 1SPSA and 2SPSA algorithms.

The mean vector fields $\bar{f}^{1\mathrm{Q}}, \bar{f}^{2\mathrm{Q}}$ are identical, provided the probing signal is symmetric:

Proposition 3.1. *Suppose (QSA1)-(QSA5) hold and that* Γ *is* C^2 *with unique minimizer* θ^{opt}*. Assume, moreover that* $M = \Sigma_\xi = I$*, and that the probing signal is of the form* (6) *with* $\phi_i = 1/4$ *for each* i*. Then, the average vector fields for* 1*qSGD and* 2*qSGD are equal:*

$$\bar{f}(\theta) = \langle f^{1\mathrm{Q}}(\theta, \xi) \rangle = \langle f^{2\mathrm{Q}}(\theta, \xi) \rangle = -M \Sigma_\xi \nabla \Gamma(\theta) + O(\varepsilon^2) \tag{22}$$

For notational convenience we set $M = I$ and assume $\Sigma_\xi = I$ in the subsequent analysis.

For the special case of a strongly convex objective Γ, current literature provides bounds on the bias of 2qSGD [30]. An application of Prop. 3.1 implies that the bias for 1qSGD is identical:

Corollary 3.2. *Suppose the assumptions of Prop. 3.1 hold, and* Γ *is strongly convex. Then,* $\|\theta^* - \theta^{\mathsf{opt}}\| \le O(\varepsilon^2)$ *for either* 1*qSGD or* 2*qSGD.*

3.2 Numerical Examples

The numerical results surveyed here and in the supplementary material used a crude approximation of the QSA ODE: an Euler scheme with sampling interval of 1 sec. It is found that the results remain consistent with theory for moderate dimension. A metric for testing this theory was via the empirical covariance, obtained using $M \geq 50$ independent runs to obtain $\Theta_T^i : \{1 \leq i \leq M\}$, and then

$$\overline{\Sigma}_T = \frac{1}{M}\sum_{i=1}^{M} \Theta_T^i \Theta_T^{i\,\intercal} - \bar{\Theta}_T \bar{\Theta}_T^{\intercal}\,, \qquad \bar{\Theta}_T = \frac{1}{M}\sum_{i=1}^{M} \Theta_T^i \tag{23}$$

Theory predicts that root mean square error $T^{2\rho}\sqrt{\mathrm{tr}(\overline{\Sigma}_T)}$ will be bounded in T subject to conditions on the algorithm and probing signal.

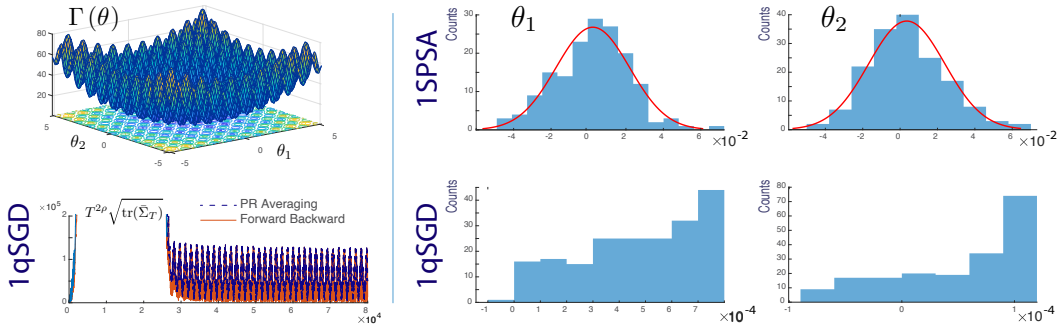

Figure 4: Rastrigin objective (top left), scaled trace of empirical covariance (bottom left), histograms of estimation error for 1SPSA with PR averaging (top middle and top right), histograms of estimation error for 1qSGD with PR averaging (bottom middle and bottom right).

Rastrigin: The two qSGD algorithms were compared in this standard benchmark [32], for which a plot of the objective can be found on the upper left in Fig. 4. In each case the values $\varepsilon = 0.25$, $\kappa = 5$ and $\rho = 0.85$ were used. The stochastic counterparts were included in these experiments in which the exploration sequence $\{\zeta_n\}$ was chosen Gaussian and i.i.d., $\mathcal{N}(0, \Sigma)$. Further details on the experiment are postponed to the supplementary material.

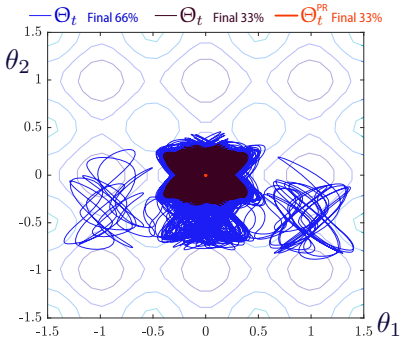

Figure 5: $\{\Theta_T\}$ for the final 66% of the run, and $\{\Theta_T^{\mathrm{PR}}\}$ for the final 33%.

As predicted by theory, the root-MSE $T^{2\rho}\mathrm{tr}\,(\bar{\Sigma}_T)^{\frac{1}{2}}$ is bounded in T when using FB filtering or PR averaging: see the plot on the lower left in Fig. 4. Histograms are shown only for 1qSGD and 1SPSA using PR-averaging. The results obtained for 1qSGD with FB filtering were similar. Results obtained using 2qSGD and 2SPSA were similar qualitatively. Outliers were identified with Matlab's *isoutlier* function and excluded in these plots. Outliers were found in 20% of the independent runs for 1SPSA, and none for 1qSGD. Outliers for deterministic algorithms were observed in other experiments, but fewer than in their stochastic counterparts.

Fig. 5 shows part of the trajectory of $\{\Theta_T\}$ for a short run. We see that $\{\Theta_T\}$ oscillates between saddle points and local extrema before settling around the minimizer $\theta^{\mathsf{opt}} = 0$ near the end of the run. The corresponding PR-estimates very closely approximate θ^{opt} for the final 33% of the run.

Ackley: This is another standard benchmark [32], whose objective function is shown on the right in Fig. 6 in the special case $d = 2$.

Fig. 6 shows the scaled root-MSE for dimensions $d = 2$ and $d = 30$ with $\rho = 0.85$. Both PR averaging and FB filtering are successful in achieving $O(T^{-2\rho})$ convergence rates. However, we see that $\mathrm{tr}\,(\bar{\Sigma}_T)^{\frac{1}{2}}$ appears to approach a steady-state only for $d = 2$, suggesting that the run length is not sufficiently large for $d = 30$. The much larger covariance is expected due to the curse of dimensionality for QSA discussed after Thm. 2.1.

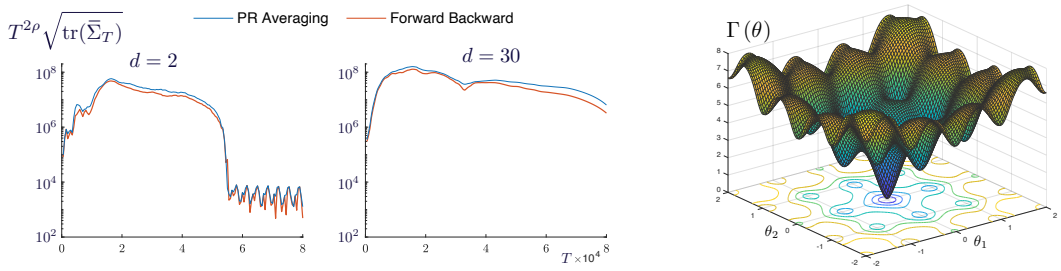

Figure 6: Scaled root-MSE for $d = 2$ (left) and $d = 30$ (middle). At right is the Ackley objective for $d = 2$.

Three-Hump Camel: This is another standard benchmark [32], whose objective function is plotted on the right in Fig. 7. Results from experiments using PR averaging, surveyed in Section C.2, show that the root-MSE is similar to what is obtained for the Ackley objective with $d = 2$. In this example the value $\Gamma(\Theta_t^{\mathsf{PR}})$ converges rapidly towards its optimal value of zero.

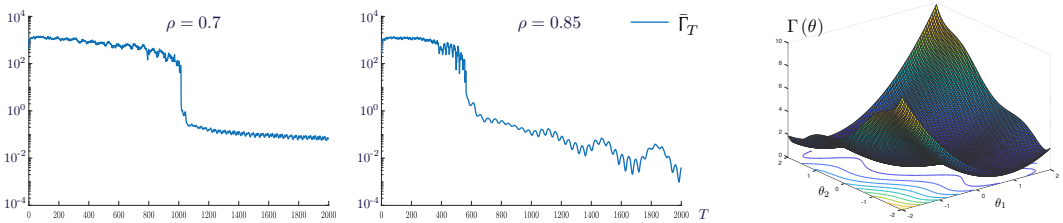

Figure 7: Average of $\Gamma(\Theta_T^i)$ over $1 \le i \le M$ independent runs for the three-hump camel.

4 Conclusions

While it is convenient to design exploration around i.i.d. signals, and this approach opens the doors to many powerful tools from probability theory, we have shown that deterministic "noise" has significant benefits. Convergence rates can be accelerated dramatically provided the algorithm and deterministic probing signals are chosen with care.

There are many avenues for future research:

△ QSA algorithms appear to suffer from the curse of dimensionality that is well known in QMC [1]. We have found in experiments that the run length $T > 0$ must be very large to see the benefit of the $O(T^{-2\rho})$ bound when d is large, or $\rho > 0.9$. Future refinements of these algorithms will be motivated by these shortcomings.

△ The optimal convergence rate for QSA is still unknown. The rates obtained in this paper are only upper bounds. It is possible that the optimal convergence rate for QSA algorithms is actually faster than $O(T^{-2\rho})$; in fact, the rate $O(T^{-1-\rho})$ is conjectured based on recent analysis.

△ It is not clear that the constraint $\rho > \frac{1}{2}$ is required in this deterministic setting, and may be removed if we can improve the bounds in our analysis. We have found in some experiments that the use of PR averaging results in very fast convergence even when this constraint is violated.

△ In stochastic optimization it is common to choose $\rho = 0$ (a constant gain or step-size). Stability theory and convergence rate bounds with averaging is another topic of research. We conjecture that bounds may be simple when noise is purely additive, following [21, 22].

△ Under what conditions is $\sqrt{\mathrm{tr}(\overline{\Sigma}_T)}$ strictly smaller using the FB algorithm as compared to PR averaging? We find in experiments that the FB algorithm usually outperforms the others, but we have yet to find tools to provide bounds that are rich enough to compare algorithms.

References

[1] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, New York, 2007.

[2] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, I. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

[3] S. Bhatnagar and V. S. Borkar. Multiscale chaotic SPSA and smoothed functional algorithms for simulation optimization. *Simulation*, 79(10):568–580, 2003.

[4] S. Bhatnagar, M. C. Fu, S. I. Marcus, and I.-J. Wang. Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(2):180–209, 2003.

[5] S. Bhatnagar, H. Prasad, and L. Prashanth. *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*. Lecture Notes in Control and Information Sciences. Springer, London, 2013.

[6] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint (2nd ed., to appear)*. Hindustan Book Agency and online `https://tinyurl.com/yebenpn4`, Delhi, India and Cambridge, UK, 2020.

[7] Y. Bugeaud. *Linear forms in logarithms and applications*, volume 28 of *IRMA Lectures in Mathematics and Theoretical Physics*. EMS Press, 2018.

[8] S. Chen, A. Devraj, A. Berstein, and S. Meyn. Revisiting the ODE method for recursive algorithms: Fast convergence using quasi stochastic approximation. *Journal of Systems Science and Complexity*, 34(5):1681–1702, 2021.

[9] A. Fradkov and B. T. Polyak. Adaptive and robust control in the USSR. *IFAC–PapersOnLine*, 53(2):1373–1378, 2020. 21th IFAC World Congress.

[10] A. Gosavi. *Simulation-based optimization*. Springer, Berlin, 2015.

[11] P. Karmakar and S. Bhatnagar. Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Math. Oper. Res.*, 43(1):130–151, 2018.

[12] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466, September 1952.

[13] V. R. Konda and J. N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *Ann. Appl. Probab.*, 14(2):796–819, 2004.

[14] N. B. Kovachki and A. M. Stuart. Continuous time analysis of momentum methods. *J. of Machine Learning Research*, 22(17):1–40, 2021.

[15] B. Lapeybe, G. Pages, and K. Sab. Sequences with low discrepancy generalisation and application to Robbins-Monro algorithm. *Statistics*, 21(2):251–272, 1990.

[16] S. Laruelle and G. Pagès. Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Methods and Applications*, 18(1):1–51, 2012.

[17] S. Liu and M. Krstic. Introduction to extremum seeking. In *Stochastic Averaging and Stochastic Extremum Seeking*, Communications and Control Engineering. Springer, London, 2012.

[18] E. M. Matveev. An explicit lower bound for a homogeneous rational linear form in logarithms of algebraic numbers. *Izvestiya: Mathematics*, 62(4):723–772, aug 1998.

[19] E. M. Matveev. An explicit lower bound for a homogeneous rational linear form in the logarithms of algebraic numbers. ii. *Izvestiya: Mathematics*, 64(6):1217, 2000.

[20] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press (to appear)., Cambridge, 2021.

[21] W. Mou, C. Junchi Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *arXiv e-prints*, page arXiv:2004.04719, Apr. 2020.

[22] E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459, 2011.

[23] A. B. Owen and P. W. Glynn. *Monte Carlo and Quasi-Monte Carlo Methods*. Springer, 2016.

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

[25] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[26] D. Ruppert. Efficient estimators from a slowly convergent Robbins-Monro processes. Technical Report Tech. Rept. No. 781, Cornell University, School of Operations Research and Industrial Engineering, Ithaca, NY, 1988.

[27] J. C. Spall. A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. of the American Control Conf.*, pages 1161–1167. IEEE, 1987.

[28] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992.

[29] J. C. Spall. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112, 1997.

[30] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2003.

[31] W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Proc. Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[32] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved May 16, 2022, from `http://www.sfu.ca/~ssurjano`.

[33] Y. Tan, W. H. Moase, C. Manzie, D. Nešić, and I. Mareels. Extremum seeking from 1922 to 2010. In *Proc. of the 29th Chinese control conference*, pages 14–26. IEEE, 2010.

[34] J. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

[35] G. Wüstholz, editor. *A Panorama of Number Theory*–Or–*The View from Baker's Garden*. Cambridge University Press, 2002.

[36] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *Proc. of the Intl. Conference on Neural Information Processing Systems*, pages 3904–3913, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.2 for an overview of the assumptions. A complete list is found in the supplementary material
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are in the supplementary material, with overviews of all significant results in the main body.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code and data are not provided, but all details required for reproduceability are provided
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details on experiments are specified in the supplementary material
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] It is mentioned in the supplementary material
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]