

A ROBBINS–MONRO SEQUENCE THAT CAN EXPLOIT PRIOR INFORMATION FOR FASTER CONVERGENCE

BY SIWEI LIU^a , KE MA^b  AND STEPHAN M. GOETZ^c 

University of Cambridge

^as12049@cam.ac.uk; ^bkm834@cam.ac.uk; ^csmg84@cam.ac.uk

We propose a new way to improve the convergence speed of the Robbins–Monro algorithm by introducing prior information about the target point into the Robbins–Monro iteration. We achieve the incorporation of prior information without the need of a—potentially wrong—regression model, which would also entail additional constraints. We show that this prior-information Robbins–Monro sequence is convergent for a wide range of prior distributions, even wrong ones, such as Gaussian, weighted sum of Gaussians, e.g., in a kernel density estimate, as well as bounded arbitrary distribution functions greater than zero. We furthermore analyse the sequence numerically to understand its performance and the influence of parameters. The results demonstrate that the prior-information Robbins–Monro sequence converges faster than the standard one, especially during the first steps, which are particularly important for applications where the number of function measurements is limited, and when the noise of observing the underlying function is large. We finally propose a rule to select the parameters of the sequence.

1. Introduction. The Robbins–Monro (RM) iteration was first proposed by Herbert Robbins and Sutton Monro as a general stochastic sequence, which can be used for almost-surely (a.s.) converging root finding of noisy function observations [38]. Due to its robustness to variability and fast convergence, it has become one of the most widely used algorithms in applied mathematics and related disciplines, such as in dealing with large data sets as well as machine learning [5, 33, 57], weather forecasting [20], engineering [27, 31], medicine, as well as neuroscience [15, 28, 51]. This sequence, also often marketed under the label *stochastic approximation*, can estimate the point x_t where a function $f(x)$ reaches a certain function value $f(x_t) = y_t$, e.g., a root. Function $f(x)$ is unknown but can be observed through and unbiasedly estimated by $y(x)$, which further contains a stochastic component [38]. Thus, all observations are inherently variable or subject to noisy measurements. The sequence can further be modified to find the root of an approximation of the gradient of a function, for example, through a converging finite-difference representation, and result in a stochastic optimisation method for finding the minima of variably or noisy functions.

In the following, let $y(x) = f(x) + e_k$, expectation $E[y(x)] = f(x)$, and e_k be a bounded variability or error term with 0 mean, changing with every function evaluation k . Denote the starting point of this algorithm by x_1 and the target point by x_t , where $f(x_t) = y_t$. The iteration of the RM sequence follows

$$x_{i+1} = x_i - s_i(y_i - y_t).$$

The abbreviation of y_i denotes the function values $y(x_i)$, and the sequence s_i represents the (decreasing) step size or gain. The latter translates a deviation of the observation or measurement of y_i from the target value y_t into a new x , i.e., input value to be tested. As was

MSC2020 subject classifications: Primary 62L20, 62L05; secondary 62L10.

Keywords and phrases: Stochastic approximation, Robbins–Monro algorithm, stochastic root finding, prior information, sequential statistical design, Bayes’ relation.

previously found, the selection of a good step size sequence is closely related to the local steepness of the function f . For the RM algorithm to be a.s. convergent, the step size sequence must further satisfy $\sum_{i=1}^{\infty} s_i = \infty$, $\sum_{i=1}^{\infty} s_i^2 < \infty$. Following the harmonic and Basel series theory of Mengoli [32] and Euler [10], that means that s_i must—in the long run—be decreasing approximately on an order faster than $i^{-0.5}$ and slower than or equal to i^{-1} . Here, $\sum_{i=1}^{\infty} s_i = \infty$ guarantees that the RM iteration will be able to converge to x_t no matter how far x_1 , x_t are from each other, and $\sum_{i=1}^{\infty} s_i^2 < \infty$ guarantees that $\lim_{i \rightarrow \infty} s_i = 0$, and hence the algorithm can be convergent. Robbins and Monro proved that this sequence converges to x_t in quadratic mean ($E[(x_i - x_t)^2] = 0$) when $y(x)$ is bounded [38].

After the RM algorithm was proposed, much work has been done to strengthen the convergence result of the RM algorithm. Blum [2] as well as Robbins and Siegmund [39] demonstrated that the RM algorithm is convergent when $f(x)$ may by itself be unbounded in the strict sense but is bounded by two linear functions. Kallianpur derived that the RM algorithm converges with probability 1 [23]. Kushner and Yin [26], Borkar [4], as well as Ljung et al. [29] strengthened this convergence result from the perspective of dynamical system theory. In their analysis, the RM algorithm is viewed as a discrete-time stochastic process. Each iteration of the algorithm is treated as a step in the evolution of a dynamical system. Having found an appropriate Lyapunov function, they demonstrated that the trajectories of the dynamical system converge to a stable equilibrium point (x_t) by showing that the Lyapunov function decreases over time. Moreover, the asymptotic properties of the RM sequence were also investigated. Chung [7], Hodges [19], and Sacks [41] proved that $\sqrt{s_i}(x_i - x_t)$ is asymptotically normal. Moreover, Chung considered the asymptotic properties of the higher-order moments of $x_i - x_t$ [7]. Fabian [11] later simplified the proof of Chung [7].

Research has studied many modifications and generalisations of the RM sequence. A continuous variant of the RM sequence reads

$$\begin{aligned} \frac{d}{dt}X(t) &= -s(t)Y(t, X(t)), \\ \int_0^{\infty} s(t) dt &= \infty, \\ \int_0^{\infty} s^2(t) dt &< \infty. \end{aligned}$$

Driml and Nedoma demonstrated that the RM algorithm is convergent when $Y(t, x)$ is monotonic in x and $Y(t, x) = M(x) + h(t)$ where $h(t)$ is an ergodic process with zero mean [8]. However, when extending the one-dimensional continuous case to the multidimensional continuous case, Hans and Spacek showed that many results from one dimension do not hold for multidimensional functions [17].

Blum first investigated the convergence of a multidimensional discrete RM algorithm [3]. In his assumptions, the sequence would move in a direction in which $\|x_n - x_t\|$ decreased. Wei [53] and Krishnaiah [25] investigated the asymptotic properties of the multidimensional discrete RM case. Ruppert modified Blum's multidimensional RM iteration and proposed a new sequence with a different stepping rule that aims at decreasing $\|f\|^2$ step by step [40].

Moreover, Joseph [22] and Wu [54] examined the RM iteration for binary data. Xiong and Xu further generalised the results of the binary RM into multiple dimensions [55].

Early in the study of the properties of the RM algorithm, its often slow convergence speed and instability became an important topic. For higher stability of the RM algorithm while keeping its convergence speed, Toulis et al. proposed a proximal version for the RM algorithm to allow better stability by penalising sudden large steps the further the iteration progresses [46]. Moreover, intensive research aimed to increase the RM algorithm's convergence

speed. Dvoretzky found a sequence of step sizes s_i that minimises $E[(x_i - x_t)^2]$ after some fixed number of iterations [9]. Kesten first proposed a widely-used intuitive and accordingly repeatedly re-suggested method for accelerating the convergence of the RM algorithm [24]. His improvement reduces the step size sequence s_i only, if $x_i - x_{i-1}$ and thus the deviation of the function observation from the target value $y_i - y_t$ has the opposite sign compared to the previous one $x_{i-1} - x_{i-2}$ or $y_{i-1} - y_t$; otherwise, the step size s_i will remain the same as s_{i-1} . Venter modified the iteration of the standard RM algorithm and included information about the local derivative based on an estimate of $f'(x_t)$ to improve the rate of convergence and the asymptotic variance of the Robbins–Monro procedure [48]. Still, the convergence speed of the RM iteration shows great potential.

The convergence speed in the first steps of the iteration can determine the overall speed and is furthermore important in many practical applications where only a few function evaluations are possible [16]. Farrell worked on achieving a statistical grasp on the residual deviation of the iteration in the form of $P(|x_{i+1} - x_t| \leq D) \geq 1 - 2\gamma$ after a finite number of steps [12, 13]. Chung derived lower and upper bounds for $E[|x_n - x_t|^r]$ and $E[(x_n - x_t)^r]$ for all n [7]. Kallianpur further demonstrated that $E[(x_n - x_t)^2] = \mathcal{O}(\log n)^{-v}$ for some constant v in his convergence proof of RM algorithm [23]. Polyak used the Lyapunov function to illustrate the rate of convergence of RM [37]. Based on these results, Sielken [42] as well as Stroup and Braun [45] investigated stopping rules for scalar nonlinear $f(x)$. Such rules also promise to shine light on the early convergence speed of RM. Yin investigated stopping rules for the multidimensional case [56]. However, these rules are all in the form of limits, which are not explicit enough for the first few steps. Wada et al. gave an intuitive equation describing the stopping time of the RM algorithm for different target accuracy and initial point x_1 for linear stochastic approximation [50]. Their equation can estimate the RM algorithm’s accuracy also for only limited number of iterations. Wada and Fujisaki generalised the results to nonlinear stochastic approximation [49]. However, by now, there is no further method that can significantly speed up the convergence of the RM algorithm in the first few steps. It is often simply assumed and accepted that the sequence x_i is typically far from the target x_t when i is small [14].

Stochastic approximation methods (and particularly Robbins–Monro) can further solve stochastic optimisation problems (which are equivalent to finding the root of the derivative $f' = 0$ for an increasing derivative f'), which is a major application of the Robbins–Monro sequence. Previous research could demonstrate that if $f(x)$ is twice continuously differentiable and strongly convex, the algorithm can attain the asymptotically optimal convergence rate $\mathcal{O}(1/n)$, that is, $E[f(x_n) - f(x_t)] = \mathcal{O}(1/n)$ [34, 44]. In the general convex case of $f(x)$, the asymptotically optimal convergence rate can only achieve $\mathcal{O}(1/\sqrt{n})$. In the long run, the order of this convergence rate appears already the best possible [35, 36].

In this paper, we intend to accelerate the convergence of the RM iteration, particularly early on, in the first steps. In many scientific and practical problems, information about the convergence point is known on a statistical level, typically the distribution is known from previous usage on similar cases or can be roughly estimated before starting the approximation process. Historical data can, for example, provide such prior information for numerical weather and climate problems. Known population data—likely estimated with the very same RM procedure—may in turn serve for threshold detection in medicine, neuroscience, or psychophysics [21, 30, 47, 52]. As for the problem of slow convergence speed mentioned above, the inclusion of such prior knowledge should improve the convergence speed. This article will introduce a sequence that expands Robbins–Monro to entire distributions and can exploit prior information in the iteration to speed up the convergence.

This text is structured as follows: Section 2 introduces the new sequence that incorporates the prior information about x_t . Sections 3–4 provide two methods to prove the convergence

of our new sequence with linear $f(x)$ and Gaussian prior. The convergence proof of nonlinear $f(x)$ with Gaussian and practically arbitrary prior follows in Sections 5 and 6. The general convergence allows for an alternative proof as coarsely sketched in Section 6.4. We numerically analysed the sequence for demonstrating its performance and better understanding of its parameters and summarised the results in Section 7. Section 8 concludes this paper.

2. Integration of prior information into the Robbins–Monro sequence.

2.1. Formalism.

NOTATION 1. $\mathcal{N}(x | \alpha, \beta^2)$ is the normal distribution with α mean and β^2 variance.

For an integration of prior information about the convergence point, we turn the RM iteration into a sequence of statistical distributions. We intuitively interpret the conventional RM iteration as the solution of a maximum likelihood estimator without any prior information evaluated in each step. For the case of near-Gaussian likelihood, the maximum and the spread would fully determine the shape. With convergence, the distribution would pointwise converge so that the spread in the form of the standard deviation decreases on the order of $1/i$. Thus, the best estimate of x_{i+1} with the information from the previous measurement would obviously be

$$x_{i+1} = \operatorname{argmax}_x P(x_{i+1} | x_i, s_i, \text{Model}),$$

where

$$P(x_{i+1} | x_i, s_i, \text{Model}) = \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), c_i^2\right).$$

Note that $c_i = c_0/i$, c_0 is a positive constant. Moreover, $\mathcal{N}(x | x_i - s_i(y_i - y_t), c_i^2)$ is a normal distribution with $x_i - s_i(y_i - y_t)$ mean and c_i^2 variance.

If we have some prior information of x_t , we can use Bayes' Theorem to derive

$$P(x_t | x_{i+1}) \propto P_{x_t}(x) P(x_{i+1} | x_i, s_i, \text{Model}),$$

where $P_{x_t}(x)$ is the prior information describing the probability distribution of x_t . We want to proceed in the RM sequence containing the prior information with the best estimate of x_t as the next step. That is, we want to choose that x_{i+1} which maximizes $P(x_t | x_{i+1})$. Therefore,

$$(1) \quad x_{i+1} = \operatorname{argmax}_x \left(P_{x_t}(x) \cdot \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), c_i^2\right) \right).$$

For brevity, we will in the following name $P_{x_t}(x)$ the prior distribution and $\mathcal{N}(x | x_i - s_i(y_i - y_t), c_i^2)$ the RM distribution. Further, we will denote the entire sequence prior-information Robbins–Monro sequence here. Note that when the prior distribution in Eq. (1) is a uniform prior (noninformative prior), the prior-information Robbins–Monro sequence will coincide with the standard RM sequence. When the prior information is known and relatively accurate, it should significantly accelerate the convergence speed, especially for the first steps when x_i is far from x_t .

2.2. *Interpretation.* The prior distribution can be practically any positive bounded continuous probability distribution. In the following analysis, we will first assume the prior distribution to be a normal distribution and then generalise the results derived from the normal prior distribution to any practical prior distribution (non-zero to allow for any outcome, bounded, and finite weight under the curve (obviously typically 1) so that the RM distribution

can eventually outgrow its maximum). Ideally, the prior distribution should (relatively) accurately describe the distribution of x_t expected in the specific problem. However, the prior distribution could also be wrong. If it does not suppress the right outcome, it should still allow—albeit typically slower—convergence. The results in our following analysis demonstrate that our algorithm will still be convergent even if the prior information is wrong if it is non-zero and finite in the relevant ranges, specifically a large enough neighbourhood of the actual root of the function.

Note that the mean of the RM distribution in step i is the value x_{i+1} of the standard RM algorithm, and the variance of the RM distribution decreases with the number of iterations. The product of the prior distribution $P_{x_t}(x)$ with the RM distribution $\mathcal{N}(x | x_i - s_i(y_i - y_t), c_i^2)$ allows the former to pull the maximum closer to the maximum of the prior, i.e., where prior information would expect the root of the function $f(x)$. However, whereas the prior distribution stays the same throughout the iteration, the RM distribution becomes narrower and sharper with every step. Furthermore, its maximum grows beyond any value of the bounded prior as well as also any product of the maximum of the prior and a value of the RM distribution outside a neighbourhood of the RM distribution maximum. This neighbourhood likewise decreases on a $1/i$ trend. Thus, the maximum of the RM distribution eventually dominates the argmax operator. From the perspective of the sharper-growing RM distribution, the prior distribution practically degrades into a constant, i.e., a naive prior, in the shrinking neighbourhood of the RM distribution maximum after a sufficient number of iterations. Thus, the prior distribution exerts its influence, particularly during the early steps, until the RM distribution turns into a dominant sharp peak.

3. Convergence for linear function f and Gaussian prior.

NOTATION 2. In all the following proof, $\prod_{u=m}^n Q(u) = 1$ if $m > n$ for any arbitrary function Q .

DEFINITION 3.1. Let x_1^s, x_2^s, \dots be the standard Robbins–Monro sequence. s_1, s_2, \dots are step size gains for the standard Robbins–Monro sequence. Then, the standard Robbins–Monro sequence follows

$$x_{i+1}^s = x_i^s - s_i(y_i^s - y_t).$$

In the Robbins–Monro sequence, the target point we want to find is x_t . Assume $f(x_t) = y_t$, and $\sum_{i=1}^{\infty} s_i = \infty$, $\sum_{i=1}^{\infty} s_i^2 < \infty$. Moreover, $y_i^s = y(x_i^s)$ is the measurement of $f(x)$ at x_i^s with $y_i^s = f(x_i^s) + \varepsilon_i^s$, ε_i^s, \dots are bounded independent random variables with 0 mean.

DEFINITION 3.2. Let x_1, x_2, \dots be the prior-information Robbins–Monro sequence, and s_1, s_2, \dots of Definition 3.1 be the step sizes also for the prior-information Robbins–Monro sequence. Then, the prior-information Robbins–Monro sequence follows

$$x_{i+1} = \operatorname{argmax}_x \left(P_{x_t}(x) \cdot \mathcal{N}(x | x_i - s_i(y_i - y_t), c_i^2) \right),$$

where $c_i = c_0/i$, c_0 is a positive constant. $P_{x_t}(x)$ describes the assumed probability distribution of x_t . Denote $P_{x_t}(x)$ as prior distribution, and designate $\mathcal{N}(x | x_i - s_i(y_i - y_t), c_i^2)$ the RM distribution. $y_i = y(x_i)$ is the measurement of $f(x)$ at x_i . Moreover, $y_i = f(x_i) + \varepsilon_i$ with ε_i, \dots are bounded independent and identically distributed (i.i.d.) random variables with 0 mean and d^2 variance ($d < \infty$). The target point we want to find is likewise x_t , $f(x_t) = y_t$.

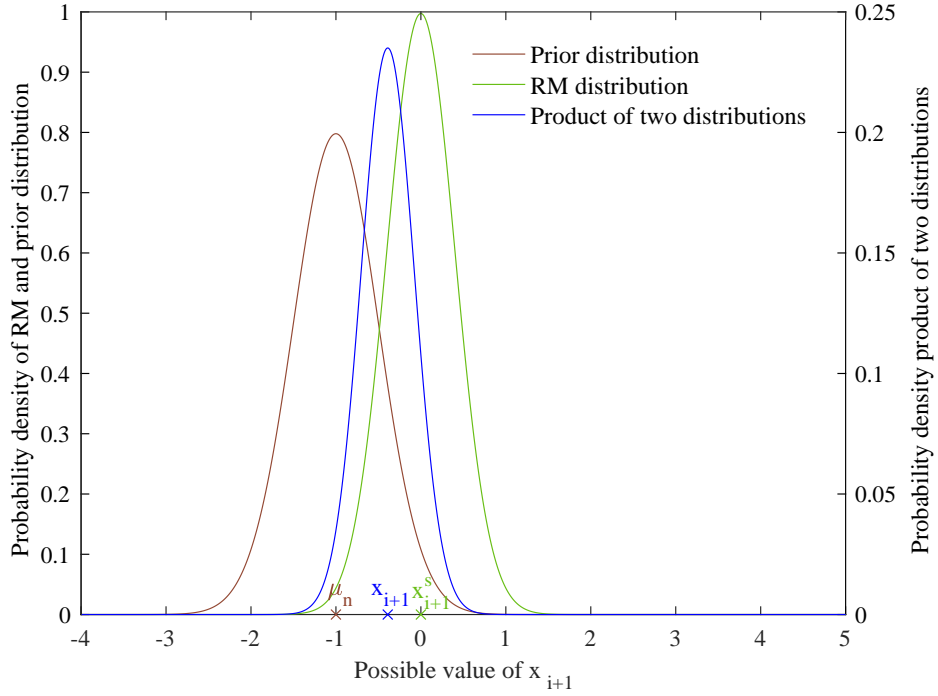


FIG 1. Illustration of the prior-information Robbins–Monro sequence with the two contributions of prior distribution and Robbins–Monro distribution combined to provide the a-posteriori distribution.

ASSUMPTION 1. If $f(x)$ is linear, w.l.o.g., assume $f(x) = ax$, $a > 0$.

ASSUMPTION 2. Assume $x_1 = x_1^s$ in the following proof.

ASSUMPTION 3. If $f(x)$ is a nonlinear function, then assume $f(x)$ is continuously differentiable, positive, and bounded per $M \geq f' \geq m > 0$.

ASSUMPTION 4. Assume w.l.o.g. that the variance of the RM distribution follows $1/i^2$ in the following proof.

ASSUMPTION 5. $\{\varepsilon_i\}$ and $\{\varepsilon_i^s\}$ are symmetric about 0. With this assumption, ε_i will be equal to $-\varepsilon_i$, $\forall i \in \mathbb{Z}^+$.

By default, Assumptions 1–5 hold for the proof below.

Note that we do not require the prior distribution to correctly describe the possible values of x_t .

LEMMA 3.3. When $f(x) = ax$ is linear ($a > 0$), for any finite $h \in \mathbb{Z}^+$,

$$\prod_{k=h}^{\infty} (1 - as_k) = 0 \quad a.s.$$

PROOF.

$$x_{i+1}^s = x_i^s - s_i(a(x_i^s - x_t) + \varepsilon_i^s)$$

$$\begin{aligned}
x_{i+1}^s - x_t &= (1 - as_i)(x_i^s - x_t) + s_i \varepsilon_i^s \\
(2) \quad &= \prod_{k=1}^i (1 - as_k)(x_1 - x_t) + \sum_{k=1}^i s_k \varepsilon_k^s \prod_{u=k+1}^i (1 - as_u)
\end{aligned}$$

Here, we write

$$\sum_{k=1}^{i-1} s_k \varepsilon_k^s \prod_{u=k+1}^i (1 - as_u) + s_i \varepsilon_i^s$$

into

$$\sum_{k=1}^i s_k \varepsilon_k^s \prod_{u=k+1}^i (1 - as_u)$$

with Notation 2 for simplification. We will frequently use this abbreviation in the following proofs.

Note that $x_1^s = x_i$. Since the standard Robbins–Monro algorithm is convergent a.s. and $E(\varepsilon_k^s) = 0 \forall k$, then

$$\begin{aligned}
&\lim_{i \rightarrow \infty} E[x_{i+1}^s - x_t] \\
&= \lim_{i \rightarrow \infty} E \left[\prod_{k=h}^i (1 - as_k)(x_h^s - x_t) + \sum_{k=h}^i s_k \varepsilon_k^s \prod_{u=k+1}^i (1 - as_u) \right] \\
&= \prod_{k=h}^{\infty} (1 - as_k)(x_h^s - x_t) = 0
\end{aligned}$$

for all finite $h \in \mathbb{Z}^+$.

Moreover, $x_h^s - x_t \neq 0$ a.s. since the value of x_h^s is influenced by $h - 1$ continuous random variables $\varepsilon_1^s, \varepsilon_2^s, \dots, \varepsilon_{h-1}^s$. Therefore,

$$\prod_{k=h}^{\infty} (1 - as_k) = 0 \quad a.s.$$

for all finite $h \in \mathbb{Z}^+$. □

LEMMA 3.4. When $f(x) = ax$ is linear ($a > 0$),

$$\sum_{k=1}^{\infty} s_k^2 d^2 \prod_{u=k+1}^{\infty} (1 - as_u)^2 = 0 \quad a.s.$$

Note that d^2 is the variance of $\{\varepsilon_i\}$.

PROOF. Here, we want to further consider a standard RM sequence that satisfies ε_k^s, \dots are i.i.d with 0 mean and d^2 variance. Note that this assumption will not change the convergence of the standard RM sequence. Calculate the square of Eq. (2) and take the expectation to get

$$(3) \quad E[(x_{i+1}^s - x_t)^2] = \prod_{k=1}^i (1 - as_k)^2 (x_1 - x_t)^2$$

$$(4) \quad + \sum_{k=1}^i s_k^2 d^2 \prod_{u=k+1}^i (1 - as_u)^2$$

since ε_i^s, \dots are i.i.d. random variables with 0 mean and d^2 variance. As the standard RM sequence is a.s. convergent,

$$\lim_{i \rightarrow \infty} E[(x_{i+1}^s - x_t)^2] = 0 \quad a.s.$$

Since Terms (3), (4) ≥ 0 , then Terms (3), (4) $\xrightarrow{i \rightarrow \infty} 0$ a.s.,

$$\sum_{k=1}^{\infty} s_k^2 d^2 \prod_{u=k+1}^{\infty} (1 - as_u)^2 = 0 \quad a.s.$$

□

THEOREM 3.5. *When the prior distribution is normal and $f(x) = ax$ is linear ($a > 0$), the prior-information Robbins–Monro sequence is generated by*

$$x_{i+1} = \operatorname{argmax}_x \left(\mathcal{N}(x \mid \mu_n, \sigma_n^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(a(x_i - x_t) + \varepsilon_i), \frac{1}{i^2}\right) \right),$$

where μ_n, σ_n are constants, $\sigma_n > 0$. Then,

$$\lim_{i \rightarrow \infty} x_{i+1} = x_t \quad a.s.$$

Therefore, the prior-information Robbins–Monro sequence is convergent a.s. for linear $f(x)$ and normal prior distribution.

PROOF.

$$\begin{aligned} x_{i+1} &= \operatorname{argmax}_x \left(\mathcal{N}(x \mid \mu_n, \sigma_n^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(a(x_i - x_t) + \varepsilon_i), \frac{1}{i^2}\right) \right) \\ &= \operatorname{argmax}_x \left(-\frac{(x - \mu_n)^2}{2\sigma_n^2} - \frac{i^2(x - x_i + s_i(a(x_i - x_t) + \varepsilon_i))^2}{2} \right) \\ &= \operatorname{argmax}_x g(x) \end{aligned}$$

Since

$$g''(x) = -\frac{1}{\sigma_n^2} - i^2 < 0, \text{ we know that } g'(x_{i+1}) = 0.$$

We can have

$$\begin{aligned} x_{i+1} &= \frac{\mu_n + i^2 \sigma_n^2 (1 - as_i) x_i + i^2 \sigma_n^2 s_i (ax_t + \varepsilon_i)}{i^2 \sigma_n^2 + 1} \\ x_{i+1} - x_t &= \frac{i^2 \sigma_n^2 (1 - as_i)}{i^2 \sigma_n^2 + 1} (x_i - x_t) + \frac{\mu_n - x_t}{i^2 \sigma_n^2 + 1} + \frac{i^2 \sigma_n^2 s_i \varepsilon_i}{i^2 \sigma_n^2 + 1} \end{aligned}$$

Hence, we can iteratively form

$$(5) \quad \lim_{i \rightarrow \infty} x_{i+1} - x_t = \prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - as_k)}{k^2 \sigma_n^2 + 1} (x_h - x_t)$$

$$(6) \quad + \sum_{k=h}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1}$$

$$(7) \quad + \sum_{k=h}^{\infty} \frac{k^2 \sigma_n^2 s_k \varepsilon_k}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1},$$

where h is a given large number that satisfies $s_k < 1/a$, $\forall k \geq h$. Such h exists since $\sum s_i^2 < \infty$, $s_i \xrightarrow{i \rightarrow \infty} 0$.

By Lemma 3.3,

$$\prod_{k=h}^{\infty} (1 - a s_k) = 0 \quad a.s.$$

Therefore,

$$0 \leq \left| \prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - a s_k)}{k^2 \sigma_n^2 + 1} (x_h - x_t) \right| \leq \left| \prod_{k=h}^{\infty} (1 - a s_k) (x_h - x_t) \right| = 0 \quad a.s.,$$

and Term (5) = 0 a.s.

For Term (6), since $\sum_{k=1}^{\infty} 1/k^2 < \infty$, for any infinitesimal $p > 0$, there exists $Z \in \mathbb{N}$ such that

$$\left| \sum_{k=Z}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2} \right| < p,$$

w.l.o.g., assume $Z > h$. Therefore,

$$0 \leq \left| \sum_{k=Z}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1} \right| \leq \left| \sum_{k=Z}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \right| \leq \left| \sum_{k=Z}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2} \right| < p.$$

Hence,

$$(8) \quad \sum_{k=Z}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1} = 0.$$

Moreover, when $h \leq k \leq Z$, following the same idea in proving in Term (5) = 0, we can get

$$\prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1} = 0 \quad a.s.,$$

and

$$(9) \quad \sum_{k=h}^{Z-1} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1} = 0 \quad a.s.$$

Therefore, adding up Eq. (8) and Eq. (9), we can get Term (6) = 0 a.s.

Calculate the square of Term (7) and take the expectation gets us to

$$(10) \quad E[(7)^2] = \sum_{k=h}^{\infty} \left(\frac{k^2 \sigma_n^2 s_k}{k^2 \sigma_n^2 + 1} \right)^2 d^2 \cdot \prod_{u=k+1}^{\infty} \left(\frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1} \right)^2$$

since $\varepsilon_k, \varepsilon_{k+1}, \dots$ are i.i.d random variables with 0 mean and d^2 variance.

By Lemma 3.3, when $k = 1, 2, 3, \dots, h-1$,

$$\prod_{u=k+1}^{\infty} (1 - as_u) = 0 \quad a.s.$$

Hence,

$$\sum_{k=1}^{h-1} s_k^2 d^2 \prod_{u=k+1}^{\infty} (1 - as_u)^2 = 0 \quad a.s.$$

Moreover, with Lemma 3.4, we can get

$$0 \leq |(10)| \leq \left| \sum_{k=h}^{\infty} s_k^2 d^2 \prod_{u=k+1}^{\infty} (1 - as_u)^2 \right| = 0 \quad a.s.$$

Therefore Eq. (10) = Term (7) = 0 a.s.

Accordingly

$$\lim_{i \rightarrow \infty} x_{i+1} \stackrel{a.s.}{=} x_t,$$

and the prior-information Robbins–Monro sequence is convergent a.s. for linear $f(x)$ and normal prior distribution. \square

REMARK. Note that if the prior distribution $\mathcal{N}(x | \mu_n, \sigma_n^2)$ does not correctly describe the possible distribution of x_t , the prior-information Robbins–Monro sequence is still convergent a.s. This remark holds for all the following convergence proofs.

4. An alternative convergence proof for a linear function f and Gaussian prior. We will also provide an alternative idea to prove the a.s. convergence of the prior-information Robbins–Monro sequence.

ASSUMPTION 6. w.l.o.g, assume $x_t = y_t = 0$.

THEOREM 4.1. *When the prior distribution is normal and $f(x) = ax$ is linear ($a > 0$), the prior-information Robbins–Monro sequence is convergent a.s. when Assumption 6 holds.*

PROOF. In this proof, we will construct a standard RM sequence $\{x_i^s\}$ that has the same limit as our prior-information Robbins–Monro sequence $\{x_i\}$ to show that both $\{x_i\}$ and $\{x_i^s\}$ converges to x_t .

Follow the same idea in Lemma 3.3 and Theorem 3.5, we can get

$$\begin{aligned} \lim_{i \rightarrow \infty} x_{i+1} &= \prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - as_k)}{k^2 \sigma_n^2 + 1} x_h + \sum_{k=h}^{\infty} \frac{\mu_n + k^2 \sigma_n^2 s_k \varepsilon_k}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - as_u)}{u^2 \sigma_n^2 + 1} \\ \lim_{i \rightarrow \infty} x_{i+1}^s &= \prod_{k=h}^{\infty} (1 - as_k) x_h^s + \sum_{k=h}^{\infty} s_k \varepsilon_k^s \prod_{u=k+1}^{\infty} (1 - as_u), \end{aligned}$$

where h is a given large number that satisfies $s_k < 1/a, \forall k \geq h$.

Since Term (6) = Term (5) = 0 a.s.,

$$\sum_{k=h}^{\infty} \frac{\mu_n}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - a s_u)}{u^2 \sigma_n^2 + 1} = 0 \quad a.s.$$

$$\prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - a s_k)}{k^2 \sigma_n^2 + 1} x_h = 0 \quad a.s.$$

Moreover, by Lemma 3.3,

$$\prod_{k=h}^{\infty} (1 - a s_k) x_h^s = 0 \quad a.s.$$

Then, $\lim_{i \rightarrow \infty} x_{i+1}$ and $\lim_{i \rightarrow \infty} x_{i+1}^s$ follow the same distribution if and only if

$$(11) \quad \varepsilon_k^s = \varepsilon_k \cdot \prod_{u=k}^{\infty} \frac{u^2 \sigma_n^2}{u^2 \sigma_n^2 + 1} \quad \forall k \in \mathbb{N}^+.$$

In Eq. (11), since $\{\varepsilon_k\}$ is bounded and i.i.d., $\{\varepsilon_k^s\}$ is bounded and independent. Moreover, $E(\varepsilon_k) = E(\varepsilon_k^s) = 0 \quad \forall k \in \mathbb{N}^+$. Therefore, $\{x_i^s\}$ is a standard RM sequence given that $\sum_{i=1}^{\infty} s_i = \infty, \sum_{i=1}^{\infty} s_i^2 < \infty$. Hence, $\{x_i^s\}$ converges to x_t , and $\{x_i\}$ converges to x_t , which means that the prior-information Robbins–Monro sequence is convergent a.s. for linear $f(x)$ and normal prior distribution. \square

5. Convergence for nonlinear function f and Gaussian prior. Keep the prior distribution to be normal. When the underlying function f , e.g., in a root search, is a nonlinear continuously differentiable function,

$$x_{i+1} = \operatorname{argmax}_x \left(\mathcal{N}(x \mid \mu_n, \sigma_n^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i), \frac{1}{i^2}\right) \right).$$

Through the mean value theorem,

$$f(x_i) - f(x_t) = f'(z_i)(x_i - x_t),$$

where z_i lies between x_i and x_t . Therefore,

$$x_{i+1} = \operatorname{argmax}_x \left(\mathcal{N}(x \mid \mu_n, \sigma_n^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(f'(z_i)(x_i - x_t) + \varepsilon_i), \frac{1}{i^2}\right) \right).$$

THEOREM 5.1. *When the prior distribution is normal and $f(x)$ is nonlinear,*

$$\lim_{i \rightarrow \infty} x_{i+1} = x_t \quad a.s.$$

Therefore, the prior-information Robbins–Monro sequence is convergent a.s. for nonlinear $f(x)$ and normal prior distribution.

PROOF. By exactly the same procedure as for Theorem 3.5,

$$(12) \quad \lim_{i \rightarrow \infty} x_{i+1} - x_t = \prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - s_k f'(z_k))}{k^2 \sigma_n^2 + 1} (x_h - x_t)$$

$$(13) \quad + \sum_{k=h}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - s_u f'(z_u))}{u^2 \sigma_n^2 + 1}$$

$$(14) \quad + \sum_{k=h}^{\infty} \frac{k^2 \sigma_n^2 s_k \varepsilon_k}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - s_u f'(z_u))}{u^2 \sigma_n^2 + 1},$$

where h is a given large number that satisfies $s_k < 1/M$, $\forall k \geq h$.

Note that $M \geq f' \geq m > 0$.

$$|(12)| \leq \left| \prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2}{k^2 \sigma_n^2 + 1} (1 - s_k m) (x_h - x_t) \right|$$

Replacing a by m in Term (5) entails

$$\prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - s_k m)}{k^2 \sigma_n^2 + 1} (x_h - x_t) = 0 \quad a.s.$$

Therefore, Term (12) = 0. Following the same idea, we can get

$$0 \leq |(13)| \leq \left| \sum_{k=h}^{\infty} \frac{\mu_n - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - s_u m)}{u^2 \sigma_n^2 + 1} \right| = 0 \quad a.s.,$$

$$0 \leq E[(14)^2] \leq \sum_{k=h}^{\infty} \left(\frac{k^2 \sigma_n^2 s_k}{k^2 \sigma_n^2 + 1} \right)^2 d^2 \cdot \prod_{u=k+1}^{\infty} \left(\frac{u^2 \sigma_n^2 (1 - m s_u)}{u^2 \sigma_n^2 + 1} \right)^2 = 0 \quad a.s.$$

Therefore, Term (13) = Term (14) = 0 a.s.

$$\lim_{i \rightarrow \infty} x_{i+1} = x_t \quad a.s.$$

Therefore, the prior-information Robbins–Monro sequence is convergent a.s. for nonlinear $f(x)$ and normal prior distribution. \square

6. Convergence for practically arbitrary prior. The underlying function $f(x)$ in this section will be assumed to be a nonlinear function satisfying Assumption 3.

6.1. Weighted sum of Gaussian prior. In this section, before going to the arbitrary prior, we will first assume the prior distribution to be the weighted sum of L normal distributions with different means and the same variance σ^2 . Then,

$$(15) \quad x_{i+1} = \operatorname{argmax}_x \left(\left(\sum_{r=1}^L w_r \mathcal{N}(x \mid \mu_r, \sigma^2) \right) \cdot \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), \frac{1}{i^2}\right) \right),$$

where $0 < w_r \leq 1 \forall r \in \{1, 2, \dots, L\}$ is the weight of the r^{th} normal distribution. If we want $\sum_{r=1}^L w_r \mathcal{N}(x | \mu_r, \sigma^2)$ to be a probability distribution, we need to have $\sum_{r=1}^L w_r = 1$. Note that the solution of Eq. (15) always exists. Then, let the solution of

$$(16) \quad \operatorname{argmax}_x \left(\mathcal{N}(x | \mu_i^A, \sigma^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), \frac{1}{i^2}\right) \right)$$

also be x_{i+1} , which follows if μ_i^A satisfies

$$(17) \quad x_{i+1} = \frac{i^2 \sigma^2 (x_i - s_i(y_i - y_t)) + \mu_i^A}{i^2 \sigma^2 + 1},$$

$$(18) \quad \mu_i^A = (i^2 \sigma^2 + 1)x_{i+1} - i^2 \sigma^2 (x_i - s_i(y_i - y_t)).$$

Note that $\mathcal{N}(x | \mu_i^A, \sigma^2)$ in Eq. (16) will change in each iteration (change with i).

THEOREM 6.1. *The prior-information Robbins–Monro sequence is convergent a.s. when the prior distribution is the weighted sum of a finite number of normal distributions with the same variance σ^2 . The weights w_r satisfy $0 < w_r \leq 1, \forall r \in \{1, 2, \dots, L\}$ and $\sum_{r=1}^L w_r = 1$.*

PROOF.

$$\begin{aligned} x_{i+1} &= \operatorname{argmax}_x \left(\left(\sum_{r=1}^L w_r \mathcal{N}(x | \mu_r, \sigma^2) \right) \cdot \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), \frac{1}{i^2}\right) \right) \\ &= \operatorname{argmax}_x \left(\mathcal{N}(x | \mu_i^A, \sigma^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), \frac{1}{i^2}\right) \right). \end{aligned}$$

Firstly, define

$$\begin{aligned} z_r &= \operatorname{argmax}_x \left(\mathcal{N}(x | \mu_r, \sigma^2) \cdot \mathcal{N}\left(x \mid x_i - s_i(y_i - y_t), \frac{1}{i^2}\right) \right) \\ (19) \quad &= \frac{i^2 \sigma^2 (x_i - s_i(y_i - y_t)) + \mu_r}{i^2 \sigma^2 + 1}. \end{aligned}$$

Obviously, x_{i+1} lies between $\min\{z_r \mid r = 1, 2, \dots, L\}$ and $\max\{z_r \mid r = 1, 2, \dots, L\}$. By substituting Eq. (19) into Eq. (18), we can derive that μ_i^A lies between $\min\{\mu_r \mid r = 1, 2, \dots, L\}$ and $\max\{\mu_r \mid r = 1, 2, \dots, L\}$, and hence $\{\mu_i^A\}$ is a bounded sequence. Then, expand Eq. (17) as in Theorems 3.5 and 5.1, we can get

$$(20) \quad \lim_{i \rightarrow \infty} x_{i+1} - x_t = \prod_{k=h}^{\infty} \frac{k^2 \sigma_n^2 (1 - s_k f'(z_k))}{k^2 \sigma_n^2 + 1} (x_h - x_t)$$

$$(21) \quad + \sum_{k=h}^{\infty} \frac{\mu_k^A - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - s_u f'(z_u))}{u^2 \sigma_n^2 + 1}$$

$$(22) \quad + \sum_{k=h}^{\infty} \frac{k^2 \sigma_n^2 s_k \varepsilon_k}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - s_u f'(z_u))}{u^2 \sigma_n^2 + 1},$$

where h is a given large number that satisfy $s_k < 1/M, \forall k \geq h$.

Following the same proof as for Theorem 5.1, we can see that Term (20) = Term (22) = 0

a.s. since Term (12) = Term (14) = 0 a.s. Let $|\mu_B - x_t| = \max \{|\mu_r - x_t| \mid r = 1, 2, \dots, L\}$, $B \in \{1, 2, \dots, L\}$. Then,

$$0 \leq |(21)| \leq \left| \sum_{k=h}^{\infty} \frac{\mu_B - x_t}{k^2 \sigma_n^2 + 1} \cdot \prod_{u=k+1}^{\infty} \frac{u^2 \sigma_n^2 (1 - s_u m)}{u^2 \sigma_n^2 + 1} \right| = 0 \quad a.s.$$

since Term (13) = Term (6) = 0 a.s.

Therefore, Term (21) = 0 a.s.,

$$\lim_{i \rightarrow \infty} x_{i+1} = x_t \quad a.s.$$

The prior-information Robbins–Monro sequence is convergent a.s. when the prior distribution is the weighted sum of a finite number of normal distributions with the same variance σ^2 . \square

6.2. Discrete arbitrary prior. We want to extend our results to more arbitrary priors. We will use kernel density estimation (KDE) to attain this goal [6]. Firstly, assume the original prior information of x_t ($P_{x_t}(x)$) is discrete. In nature, all the primary measurement results we have are discrete since we cannot perform infinitely many measurements.

Assume we have L measurements of $f(x)$ (at $x_1^M, x_2^M, \dots, x_L^M$) with the result y_t , and we will use the normal kernel function

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

in our KDE analysis. Note that σ in Theorem 6.1 is exactly the bandwidth of KDE. Then, with these discrete measurement results, we will approximate the original prior distribution $P_{x_t}(x)$ by

$$(23) \quad P_0(x) = \frac{1}{L\sigma} \sum_{r=1}^L K\left(\frac{x_j^M - x}{\sigma}\right),$$

which is the averaged sum of normal distributions with the same variance. Here, $w_r \equiv 1/L$. Therefore, for any discrete prior information of x_t , we can use KDE to pre-process the prior information to get the approximated continuous prior distribution $P_0(x)$ (Eq. (23)) and derive an a.s. convergent prior-information Robbins–Monro sequence.

6.3. Continuous practically arbitrary prior. Now, we will consider the case that the original prior distribution is positive, bounded, and continuous. We will need the following assumptions and notation:

ASSUMPTION 7. $s_i = 1/i$.

ASSUMPTION 8. The function underlying the prior-information Robbins–Monro sequence $f(x)$ satisfies

$$c_2 x - d_2 < f(x) \leq y_t - \delta, \quad \text{for } x < x_t,$$

$$c_1 x + d_1 > f(x) \geq y_t + \delta, \quad \text{for } x > x_t,$$

for some positive constants $c_1, c_2, d_1, d_2, \delta$.

ASSUMPTION 9. The prior distribution $P_{x_t}(x)$ is a finite continuous distribution that satisfies $|P'_{x_t}(x)/P_{x_t}(x)| \leq J$ for some $J > 0$. This condition guarantees that $P'_{x_t}(x)$ is bounded. Furthermore, the prior should be bounded per $0 < P_{x_t}(x) < \infty$.

NOTATION 3. Denote the RM distribution at the i^{th} iteration

$$\mathcal{N}\left(x \mid x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i), \frac{1}{i^2}\right)$$

by $N_i(x)$. The derivative of $N_i(x)$ is $N'_i(x)$.

LEMMA 6.2. Assume Assumptions 7–9 hold. Let

$$\begin{aligned} x_{i+1} &= \operatorname{argmax}_x (P_{x_t}(x) \cdot N_i(x)), \\ \mu_i^c &= x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i). \end{aligned}$$

Then,

$$|x_{i+1} - \mu_i^c| \leq \frac{J}{i^2}.$$

PROOF.

$$\begin{aligned} N_i(x) &= \mathcal{N}\left(x \mid x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i), \frac{1}{i^2}\right) \\ &= \frac{i}{\sqrt{2\pi}} \exp\left(-\frac{i^2\left(x - x_i + s_i(f(x_i) - f(x_t) + \varepsilon_i)\right)^2}{2}\right) \end{aligned}$$

$$N'_i(x) = N_i(x) \cdot (-i^2(x - x_i + s_i(f(x_i) - f(x_t) + \varepsilon_i)))$$

Since $\mu_i^c = x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i)$,

$$\frac{N'_i(x)}{N_i(x)} = -i^2(x - \mu_i^c).$$

When $N_i(x) \cdot P_{x_t}(x)$ attains its maximum at x_{i+1} ,

$$x_{i+1} = \operatorname{argmax}_x (P_{x_t}(x) \cdot N_i(x)),$$

$$N'_i(x_{i+1})P_{x_t}(x_{i+1}) + N_i(x_{i+1})P'_{x_t}(x_{i+1}) = 0$$

$$\frac{N'_i(x_{i+1})}{N_i(x_{i+1})} = -\frac{P'_{x_t}(x_{i+1})}{P_{x_t}(x_{i+1})}.$$

Note that

$$\left| \frac{P'_{x_t}(x_{i+1})}{P_{x_t}(x_{i+1})} \right| \leq J.$$

Therefore,

$$\left| \frac{N'_i(x_{i+1})}{N_i(x_{i+1})} \right| = |i^2(x_{i+1} - \mu_i^c)| \leq J,$$

$$|x_{i+1} - \mu_i^c| \leq \frac{J}{i^2}.$$

□

THEOREM 6.3. *When Assumptions 7–9 hold, the prior-information Robbins–Monro sequence is a.s. convergent.*

PROOF. We will continue to use the notations in Lemma 6.2. We have seen that introducing the prior distribution at the i^{th} iteration will at most pull x_{i+1} away from μ_i^c by J/i^2 . Note that μ_i^c is the next step if the sequence follows the standard RM scheme (without the prior information).

Let $f^s(x_i)$ be the underlying function of a potential standard RM sequence $\{x_{i+1}^s\}$ such that $x_{i+1}^s = x_{i+1} \forall i \in \mathbb{N}$. We will prove that $\{x_{i+1}^s\}$ is a standard RM sequence. Since

$$\begin{aligned} x_{i+1}^s &= x_{i+1} = x_i - s_i(f^s(x_i) - f(x_t) + \varepsilon_i) \\ &\in \left[\mu_i^c - \frac{J}{i^2}, \mu_i^c + \frac{J}{i^2} \right] \\ &= \left[x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i) - \frac{J}{i^2}, x_i - s_i(f(x_i) - f(x_t) + \varepsilon_i) + \frac{J}{i^2} \right] \end{aligned}$$

and $s_i = 1/i$,

$$f^s(x_i) \in \left[f(x_i) - \frac{J}{i}, f(x_i) + \frac{J}{i} \right].$$

Hence, from the above equations, we may easily see that to let $x_{i+1}^s = x_{i+1}$, that is, to let

$$\operatorname{argmax}_x (P_{x_t}(x) \cdot N_i(x)) = x_i - s_i(f^s(x_i) - f(x_t) + \varepsilon_i),$$

$f^s(x_i)$ is at most perturbed by J/i relative to $f(x_i)$.

Then, we will define $f^s(x)$ as follows:

If $x \neq x_i (= x_i^s)$, $\forall i \in \mathbb{N}^+$, or $x = x_i (= x_i^s)$, for some $i \leq 2J/\delta$, let $f^s(x) = f(x)$. Note that we defined δ in Assumption 8.

If $x = x_i (= x_i^s)$, for some $i > 2J/\delta$, the value of $f^s(x_i)$ is chosen such that

$$x_{i+1}^s = x_i - s_i(f^s(x_i) - f(x_t) + \varepsilon_i) = x_{i+1} = \operatorname{argmax}_x (P_{x_t}(x) \cdot N_i(x)).$$

Since the values of x_1, x_2, \dots are influenced by a number of continuous random variables $\varepsilon_1, \varepsilon_2, \dots$, we can see that $\forall i, j \in \mathbb{N}^+$, $x_i \neq x_j$ a.s. Therefore, $f^s(x)$ is well-defined here.

Since $f(x_{i+1}^s) \in [f(x_i) - J/i, f(x_i) + J/i]$ for $i > 2J/\delta$, $f(x_{i+1}^s) \in [f(x_i) - \delta/2, f(x_i) + \delta/2]$. Moreover, $f(x^s) = f(x)$ for the other points x . Therefore,

$$f^s(x) \leq y_t - \delta/2, \quad \text{for } x < x_t,$$

$$f^s(x) \geq y_t - \delta/2, \quad \text{for } x > x_t,$$

given that

$$f(x) \leq y_t - \delta, \quad \text{for } x < x_t,$$

$$f(x) \geq y_t - \delta, \quad \text{for } x > x_t$$

by Assumption 8.

Moreover, since $f(x)$ satisfies

$$f(x) > c_2 x - d_2, \quad \text{for } x < x_t,$$

$$f(x) < c_1 x + d_1, \quad \text{for } x > x_t,$$

for some positive constants c_1, c_2, d_1, d_2 by Assumption 8, we can easily see that $f^s(x)$ satisfies

$$f^s(x) > c_3x - d_3, \quad \text{for } x < x_t,$$

$$f^s(x) < c_4x + d_4, \quad \text{for } x > x_t,$$

for some positive constants c_3, c_4, d_3, d_4 since the difference between $f^s(x)$ and $f(x)$ is very small.

By Blum, the underlying function $f(x)$ of the standard RM sequence needs to satisfy the following conditions [2]:

$$c_2x - d_2 < f(x) < y_t, \quad \text{for } x < x_t,$$

$$c_1x + d_1 > f(x) > y_t, \quad \text{for } x > x_t.$$

Note that $\{\varepsilon_i\}$ have finite variance. Therefore, by Blum's argument, $f^s(x)$ can be the underlying function of a standard RM sequence. $\{x_i\} = \{x_i^s\}$ are a.s. convergent standard RM sequences when $i > 2J/\delta$. These two sequences are generated by $x_{i+1} = x_{i+1}^s = x_i^s - s_i(f^s(x_i) - f(x_t) + \varepsilon_i)$.

When $i \leq 2J/\delta$, we just simply make $x_i^s = x_i = \operatorname{argmax}_x (P_{x_t}(x) \cdot N_i(x))$. Note that the initial few steps will not influence the convergence of the whole sequence. It will not influence whether $\{x_i^s\}$ is a standard RM sequence or not.

Therefore, with Assumptions 7–9, the prior-information Robbins–Monro sequence $\{x_i\}$ is indeed a standard RM sequence. Therefore, it is a.s. convergent. \square

REMARK. Note that we require the original prior distribution $P_{x_t}(x)$ to be always positive and bounded, i.e., $0 < P_{x_t}(x) < \infty$. Otherwise, the prior-information Robbins–Monro sequence may not converge to x_t .

REMARK. We may see that if $P_{x_t}(x)$ is a polynomial in a closed interval $[a_0, a_1]$ and satisfies $\inf_{x \in [a_0, a_1]} P_{x_t}(x) > 0$, the Assumption 9 will be satisfied for $P_{x_t}(x)$ locally in that interval.

Afterwards, we will go back to use $P_0(x)$ (in Eq. (23)) to estimate the continuous positive bounded prior distribution $P_{x_t}(x)$. The KDE $P_0(x)$ has several important properties. These properties are largely a result of the characteristics of the normal kernel and the non-parametric nature of KDE [18]:

(1) Smoothness: $P_0(x)$ produces a smooth density estimate. The Gaussian kernel is infinitely differentiable, which means that the resulting density estimate is also smooth and has no discontinuities.

(2) Non-Negative: The estimated density function is always non-negative.

(3) Normalisation: The integral over the entire space of $P_0(x)$ equals 1, satisfying the property of a probability density function.

(4) Bandwidth Dependence: The choice of bandwidth (σ) has a significant effect on $P_0(x)$. A small bandwidth leads to a density estimate that is more sensitive to fluctuations in the data (high variance but low bias), potentially resulting in an overfit. A large bandwidth smooths out these fluctuations (low variance but high bias), which might oversimplify the structure of the data.

(5) Pointwise Convergence: As the sample size increases and thus the number of Gaussians in the series, $P_0(x)$ converges to $P_{x_t}(x)$ at each point.

(6) Asymptotic Properties: $P_0(x)$ has well-defined asymptotic properties. As the number of data points increases, $P_0(x)$ approaches $P_{x_t}(x)$, provided the bandwidth is chosen appropriately.

(7) Boundedness: From Eq. (23), we can see that $0 < P_0(x) \leq 1/(\sqrt{2\pi}\sigma)$, where the bandwidth σ determines its maximum. Moreover, given the bandwidth σ , $P_0(x)$ is always finite. Note that the spread of $x_1^M, x_2^M, \dots, x_L^M$ will also influence the maximum of $P_0(x)$. If they are centralised, the maximum of $P_0(x)$ will be larger. If they are decentralised, the maximum of $P_0(x)$ will be lower.

(8) Handling Multimodal Distributions: $P_0(x)$ is capable of representing multimodal distributions (distributions with multiple peaks).

(9) Local Property: The estimation at each point is influenced mainly by nearby data points due to the local nature of the Gaussian kernel. This property allows $P_0(x)$ to adapt to the local structure of the data.

(10) Global Property: The Gaussian kernel, with its infinite support, means that every data point has some influence on every estimation point, which is beneficial for capturing global structures.

(11) Robustness and Flexibility: $P_0(x)$ does not make strong assumptions about $P_{x_t}(x)$, rendering it particularly useful in situations where $P_{x_t}(x)$ is unknown or complicated.

(12) Kernel Symmetry: The Gaussian kernel is symmetric, which simplifies the mathematical treatment of $P_0(x)$ and ensures that the density estimate is not skewed in any particular direction around a data point.

If we draw L random discrete samples from $P_{x_t}(x)$ to estimate $P_{x_t}(x)$ by $P_0(x)$ (in Eq. (23)), the approximated prior distribution $P_0(x)$ is always positive and finite. Hence, the prior-information Robbins–Monro sequence with $P_0(x)$ (as the prior distribution) is always convergent. If the original continuous prior distribution $P_{x_t}(x) = 0$ or ∞ somewhere, without pre-processing it, the corresponding prior-information Robbins–Monro sequence may not converge to x_t .

By the Law of Large Numbers, when the sample size L is sufficiently large, the distribution of the discrete samples will be very close to the original continuous prior distribution $P_{x_t}(x)$. Higgins showed that the order of mean integrated square error (MISE) between $P_0(x)$ and $P_{x_t}(x)$ is $\mathcal{O}(\sigma^4) + \mathcal{O}(1/L\sigma)$ [18]. Therefore, we can see that when σ is close to 0 ($\sigma > 0$) and $L \gg 1/\sigma$, the MISE between $P_0(x)$ and $P_{x_t}(x)$ will be very small. Hence, the KDE $P_0(x)$ with a small bandwidth (σ) and large sample size (L) will be a good estimate of $P_{x_t}(x)$. Since the prior-information Robbins–Monro sequence with $P_0(x)$ as the prior distribution is a.s. convergent according to Theorem 6.1, this sequence can be an estimate of the prior-information Robbins–Monro sequence with any prior distribution $P_{x_t}(x)$ of practical concern if $0 < P_{x_t}(x) < \infty$.

Moreover, Alspach and Sorenson previously supported that the Gaussian sum in the form of

$$(24) \quad P_A(x) = \sum_{r=1}^L w_r N(x | a_r, \sigma^2)$$

(where $\sum_{r=1}^L w_r = 1$, $w_r \geq 0$ for all r) converges uniformly to any probability density function of practical value as the number of terms L increase and the variance σ^2 approaches 0 [1]. If σ is small enough but nonzero, any positive finite continuous prior distribution $P_{x_t}(x)$ can be approximated by a Gaussian sum in the form of Eq. (24) to any accuracy. Moreover, Sorenson and Alspach further showed that if $P_{x_t}(x)$ is continuous at all but a finite number of locations, we can still use $P_A(x)$ to approximate $P_{x_t}(x)$ [43]. Moreover, we proved that the prior-information Robbins–Monro sequence, whose prior distribution is a convex combination of same-variance Gaussians, is a.s. convergent per Theorem 6.1. In combination,

the prior-information Robbins–Monro sequence is a.s. convergent for all practical positive finite prior distributions $P_{x_A}(x)$, which well approximate any prior distribution $P_{x_t}(x)$ that is continuous at all but a finite number of points.

6.4. More general convergence. The above proof covers most practical prior distributions as it includes all priors that can be represented by a Gaussian kernel density estimate. We could alternatively assume any bounded and (beyond a finite number of discontinuous steps) continuous prior $P(x)$ greater than 0 so that the target x_t and also any point between every x_i and the target x_t can be reached by the sequence. The boundedness excludes concepts such as Dirac distributions, which would derail the current sequence as it was defined and would need a modified formalism. Convergence is guaranteed if a standard RM sequence $\{x_i^s\}$ exists that reproduces the exact same sequence of $\{x_i\}$ with a control sequence $\{s_i^s\}$ according to

$$(25) \quad \begin{aligned} x_{i+1}^s &= x_i^s - s_i^s(y_i^s - y_t), \\ x_i^s &= x_i \quad \forall i, \\ y_i^s &= y_i \quad \forall i, \end{aligned}$$

so that

$$(26) \quad \sum s_i^s = \infty,$$

$$(27) \quad \sum (s_i^s)^2 < \infty.$$

Please note that with a prior distribution unequal to a constant, $\{s_i^s\}$ does not coincide with s_i from the conventional RM sequence. Instead, $\{s_i^s\}$ incorporates the effect of the prior initially pulling the sequence closer to its own centre of gravity. The existence of s_i^s is trivial as it can be systematically constructed. For proving convergence, this sequence has to fulfill Conditions (26) and (27) in the long run.

7. Numerical analysis.

NOTATION 4. $X = \text{Lognormal}(\alpha, \beta^2)$ is a random variable such that $\ln X$ follows the normal distribution $\mathcal{N}(x | \alpha, \beta^2)$ with α mean and β^2 variance.

NOTATION 5. $X = 10^{\mathcal{U}(\alpha, \beta)}$ is a distribution such that $\lg X = \log_{10} X$ follows the uniform distribution $\mathcal{U}(\alpha, \beta)$.

We performed all numerical analysis in MATLAB R2016b (9.1.0.441655) 64-bit (win64). A key objective of this section is to find the optimal $c_{i=0}$ in

$$x_{i+1} = \operatorname{argmax}_x \left(\mathcal{N}(x | \mu_n, \sigma_n^2) \cdot \mathcal{N}(x | x_i - s_i(a(x_i - x_t) + \varepsilon_i), c_i^2) \right),$$

where $c_i = c_0/i$, c_0 is a positive constant, so that the prior-information Robbins–Monro sequence converge fastest. The expectation of ε_i is 0 and the standard deviation of ε_i is d , that is, $\varepsilon_i = \mathcal{N}(0, d^2)$.

We will first find the parameters related to the optimal value of c_0 and demonstrate that the c_i sequence should reflect the uncertainty of the observation of the function f . Then, we will give an equation describing the relationship between the optimal c_0 with these parameters by linear regression. Afterwards, we will find out how much the prior-information Robbins–Monro sequence with the optimal c_0 performs better than the standard RM algorithm.

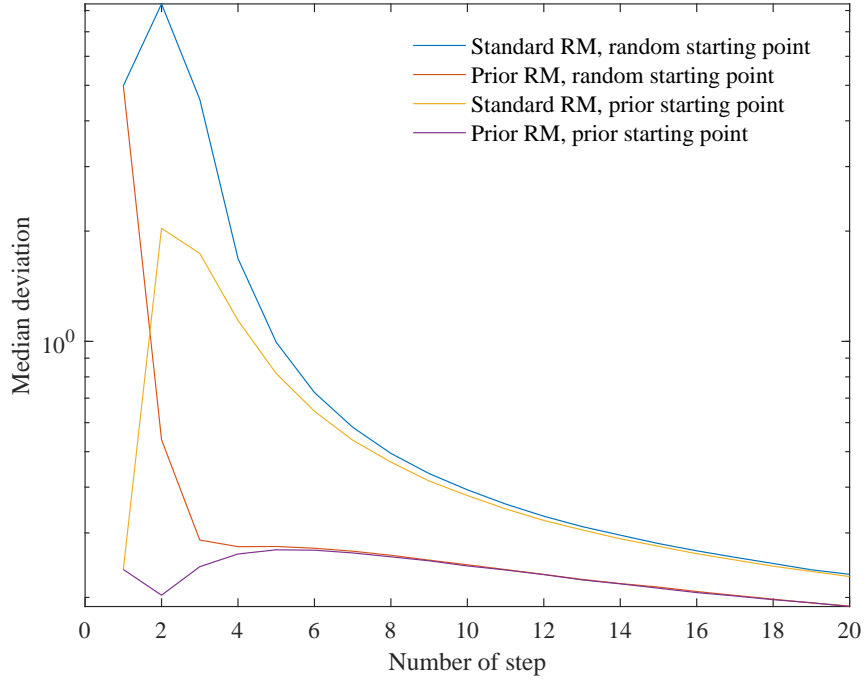


FIG 2. The median deviation ($|x_i - x_t|$ or $|x_i^s - x_t|$) of 400,000 runs for four algorithms in the first 20 steps ($c_0 = 0.3$, $d = 1$).

In this section, the prior distribution is constantly $\mathcal{N}(0.5, (0.25)^2)$, and the underlining function of the RM algorithm is $f(x) = ax$. For each run (not each step/iteration) of the (standard or prior) RM algorithm, a will be randomly chosen again according to $\text{Lognormal}(0, (0.5)^2)$ and x_t will be randomly chosen under $\mathcal{N}(0.5, (0.25)^2)$. The step size gain sequence $s_i = s_1/i$, where s_1 is randomly chosen from $10^{\mathcal{U}(0,1)}$. For each time, we want to measure $f(x_i)$, the result we get is $y(x_i) = y_i = f(x_i) + \varepsilon_i$. Since we will only run the algorithms for a finite number of iterations and repetitions, ε_i can be seen as a bounded random variable in our simulation. For all the simulations below, we always run the whole algorithm for many times. All the results we show in this section will be the median (or the average of medians) of a number of runs (for example, the median deviation (of 400,000 runs) of the prior-information Robbins–Monro sequence at the 10^{th} iteration).

Four algorithms are investigated in Figure 2. The algorithms with prior starting points will choose their starting points according to the prior distribution $\mathcal{N}(0.5, (0.25)^2)$. The algorithms with naive random starting points will choose their starting points according to a uniform distribution of $\mathcal{U}(-10, 10)$. For each algorithm, we will investigate the deviation ($|x_i - x_t|$ or $|x_i^s - x_t|$) when $i = 1, 2, \dots, 20$.

Figure 2 illustrates that the prior-information Robbins–Monro sequence converges faster than the standard RM, especially in the first few iterations and when x_1 (or x_1^s) is far from x_t . Subsequently, we are naturally concerned about the parameters related to the optimal value of c_0 that let the prior-information Robbins–Monro sequence converge fastest.

In Figure 3 A, the reader may readily recognise that the prior-information Robbins–Monro sequence with $d = 0.15$ performs best around the 10^{th} iteration when c_0 lies around 0.12. From $c_0 = 0.12$ on and with larger c_0 , the residual deviation of the prior-information Robbins–Monro sequence at the 10^{th} iteration increases, even exceeding the deviation of the standard RM when $c_0 > 0.3$.

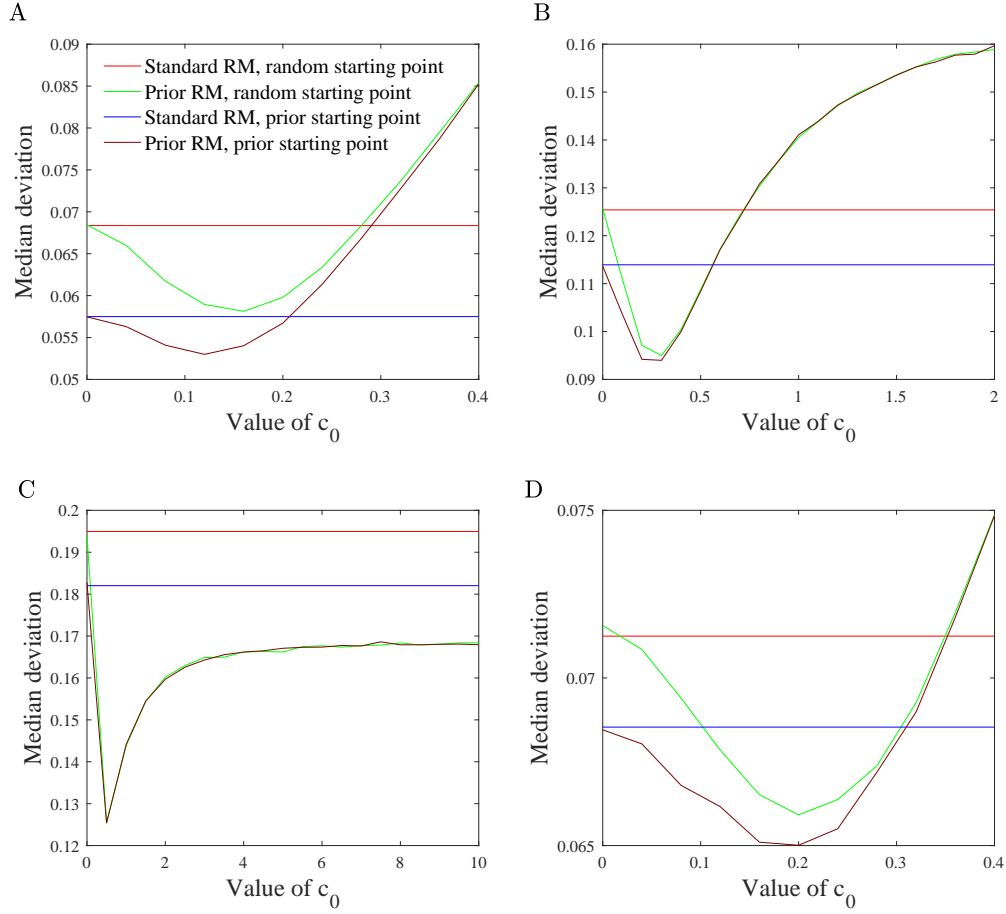


FIG 3. The median deviation ($|x_i - x_t|$ or $|x_i^s - x_t|$) of 400,000 runs at the 10^{th} step for the four algorithms when A $d = 0.15$, B $d = 0.3$, C $d = 0.48$, and D at 20^{th} iteration when $d = 0.3$.

An intuitive explanation for this phenomenon is a trade-off between the prior distribution and the noise contribution or uncertainty of the function observation. If d is small, i.e., when the measurement of $f(x)$ is more accurate and less variable, a large value c_0 will spread out the RM distribution and flatten its maximum so that the prior distribution's influence dominates and pulls x_{i+1} too close to μ_n . Thus, the prior distribution would dominate. However, typically $\mu_n \neq x_t$. Then x_{i+1}^s may be closer to x_t than x_{i+1} when c_0 is too large.

Comparing Figure 3 A and B, the error of measuring $f(x)$ is doubled. Moreover, the optimal value of c_0 also roughly doubles, increasing from around 0.12 to 0.3. When the variability d of the observation increases, x_{i+1} and x_{i+1}^s naturally converge slower as this variability has to be averaged out over more iterations. An adjustment of the variance of the RM distribution automatically increases the influence of the prior distribution (increase c_0) and places x_{i+1} closer to μ_n already in the first few iterations. The prior-information RM consequently accelerates the convergence since μ_n is an estimate of x_t . Again, however, the influence of the prior distribution may not be too large in this balance. Otherwise, it will take time until the accumulated information in the RM distribution (and the noise averaged out with every iteration) requires too many iterations before it outbalances the prior.

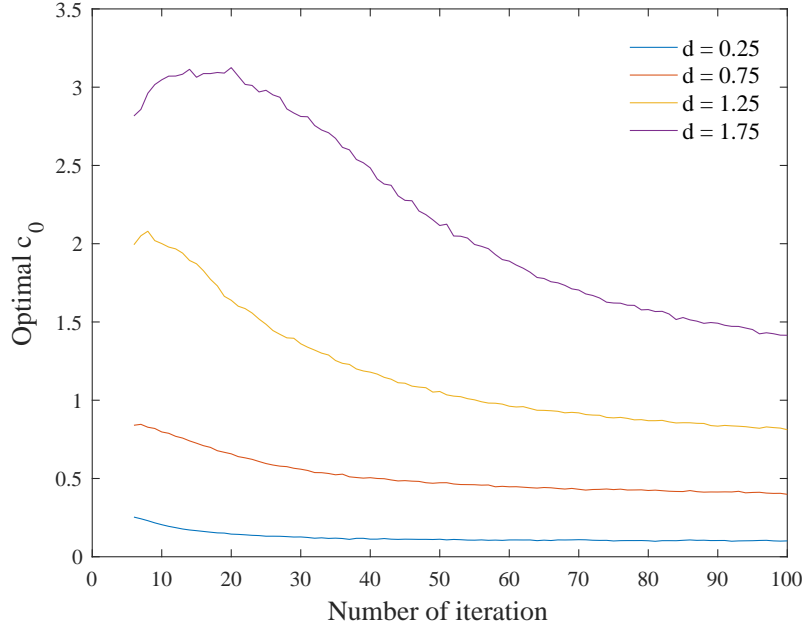


FIG 4. The graph of the optimal c_0 of the prior distribution for $d \in \{0.25, 0.75, 1.25, 1.75\}$ and iterations 6–100. Optimal c_0 for each d and iteration is the average of 700 medians derived from 6,000 runs. (run the whole algorithm for $700 \times 6,000$ times)

If $d \geq 0.48$ (see Figure 3 C), the residual deviation of the prior-information Robbins–Monro sequence is always smaller than that of the standard RM at the 10th iteration for all studied c_0 . Hence, for large observation variability d , the prior-information Robbins–Monro sequence consistently outperforms standard RM.

Figure 3 D displays the residual deviation at the 20th iteration when $d = 0.3$. The optimal c_0 value appears to be around 0.2. Comparing Figures 3 B and D, we can see that the best-performing c_0 for the prior-information Robbins–Monro sequence also depends on the number of steps into the iteration. When the number of steps increases, the optimal c_0 tends to decrease.

In conclusion, the optimal initial variance c_0 of the RM distribution is related to the observation variability d and the number of steps into the iteration. We identified the optimal c_0 , i.e., the one with the smallest residual deviation from the convergence point, for each $d \in [0, 2]$ over 6–100 iterations. In this numerical simulation, we chose the starting point of the iteration according to the prior distribution $\mathcal{N}(0.5, (0.25)^2)$. The first five iterations are excluded as these steps are highly unstable and rather random, so that they do not yet contain much information. A linear regression leads to

$$(28) \quad c_0 = 1.32 \times d - 0.0089 \times \langle \text{Iteration} \rangle + 0.13.$$

The root-mean-squared error of the linear regression is 0.317, and $R^2 = 0.865$. The F-statistic vs. a constant model is 61400, and $p \ll 10^{-10}$. The linear regression result reflects the above-identified increase of the optimal c_0 with increasing observation variability d and a minor decrease with the number of iterations. We presented the c_0 –iteration relationship curve for $d \in \{0.25, 0.75, 1.25, 1.75\}$ in Figure 4.

We further investigate the performance of the prior-information Robbins–Monro sequence with the optimal c_0 . Then, we will run the RM algorithm (with and without prior) and stop at

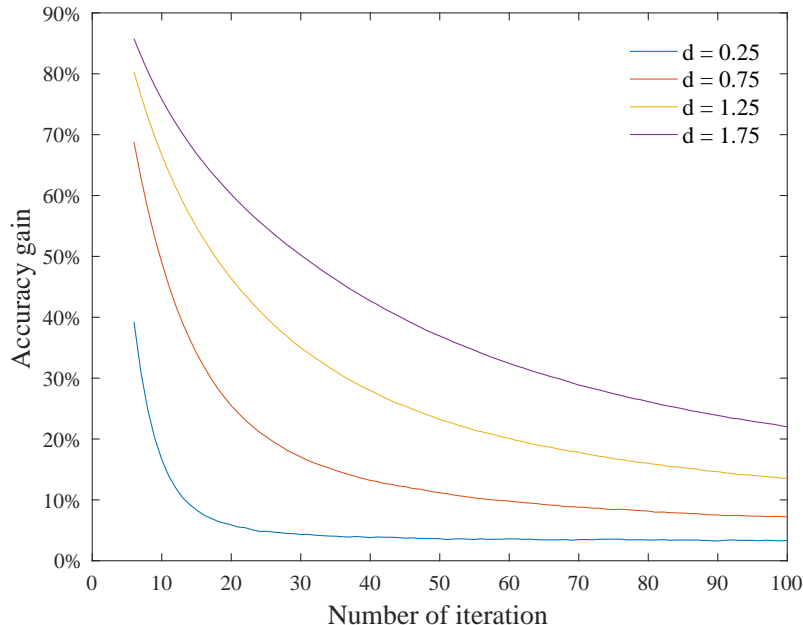


FIG 5. The graph of the performance of the prior-information Robbins–Monro sequence with the optimal c_0 for $d \in \{0.25, 0.75, 1.25, 1.75\}$ and iterations 6–100. Accuracy gain for each d and iteration is the average of 700 medians derived from 6,000 runs. (run the whole algorithm for $700 \times 6,000$ times)

the iteration we designated to see the performance of the prior-information Robbins–Monro sequence at that iteration with the given optimal c_0 . The results are shown in Figure 5. Given the function observation variability d , we denote the standard and prior-information Robbins–Monro sequences’ deviation at the i^{th} iteration ($|x_i^s - x_t|$ and $|x_i - x_t|$) respectively by d_i^s and d_i . The value in Figure 5 is the ratio of $(d_i^s - d_i)/d_i^s$ and thus reflects the gain of the prior-information RM over the standard RM. The starting points of both algorithms are chosen randomly according to the prior distribution $\mathcal{N}(0.5, (0.25)^2)$, i.e., the prior distribution.

Results in Figure 5 illustrate that for any observation variability d , the prior-information Robbins–Monro sequence will perform better than the standard RM sequence with the corresponding optimal c_0 . After many iterations, the advantage washes out though. The prior-information Robbins–Monro sequence outperforms the standard RM primarily early into the iteration and/or for large function observation variability d .

8. Conclusion. This paper proposed a novel way to improve the convergence speed of the Robbins–Monro algorithm by turning it into a sequence of distributions and introducing prior information about the target point x_t . We proved the convergence of the prior-information Robbins–Monro sequence when the prior distribution is Gaussian or an averaged sum of same-variance Gaussians. The latter can approximate most practical arbitrary prior distributions using kernel density estimation with a normal kernel to derive an a.s. convergent prior-information Robbins–Monro sequence. We also gave the convergence proof of the prior-information Robbins–Monro sequence with an arbitrary continuous prior under some restrictions.

We further analysed the sequence for linear $f(x)$ with a random slope, which should represent the most practical functions locally. In the numerical results, the prior-information Robbins–Monro sequence consistently outperforms the standard Robbins–Monro iteration,

especially in the first iterations when x_1 is far from x_t and/or when the function's observation variability d is large. The optimal initial spread of the RM distribution (c_0) is tightly linked to the random measurement error d of $f(x)$. A smaller correction can refine this value depending on the intended number of iterations. A regression curve derived from our analysis can inform the best choice of c_0 .

A multidimensional prior-information Robbins–Monro sequence and the continuous-time prior-information Robbins–Monro sequence still require further research.

Author contributions. SMG conceived, supervised, secured funding for the study, and provided the sequence for introducing prior information into the RM iteration. SMG also provided part of the code. SMG and SL conceived and sketched concepts for the proofs. SL completed all the proofs in this paper. SL further studied and characterised the sequence and its performance. SMG and KM checked the proofs. SL and KM also wrote the rest of the code for numerical analysis and visualisation. SL and SMG wrote the manuscript, KM edited the text and figures, and all authors reviewed, commented on, and approved the final version of the manuscript.

REFERENCES

- [1] ALSPACH, D. and SORENSON, H. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE transactions on automatic control* **17** 439–448.
- [2] BLUM, J. R. (1954). Approximation methods which converge with probability one. *The Annals of Mathematical Statistics* 382–386.
- [3] BLUM, J. R. (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics* 737–744.
- [4] BORKAR, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint* **48**. Springer.
- [5] BOTTOU, L., CURTIS, F. E. and NOCEDAL, J. (2018). Optimization methods for large-scale machine learning. *SIAM review* **60** 223–311.
- [6] CHEN, Y. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology* **1** 161–187.
- [7] CHUNG, K. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics* 463–483.
- [8] DRIML, M. and NEDOMA, J. (1960). Stochastic approximations for continuous random processes. In *Trans. of the second Prague conference on information theory* 145–148.
- [9] DVORETSKY, A. (1955). *On stochastic approximation*. Mathematics Division, Office of Scientific Research, US Air Force.
- [10] EULER, L. (1744). Variæ observationes circa series infinitas. *Commentarii academiae scientiarum Petropolitanae* **9** 160–188.
- [11] FABIAN, V. (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics* 1327–1332.
- [12] FARRELL, R. (1962). Bounded length confidence intervals for the zero of a regression function. *The Annals of Mathematical Statistics* 237–247.
- [13] FARRELL, R. H. (1959). *Sequentially determined bounded length confidence intervals*. University of Illinois at Urbana-Champaign.
- [14] GLYNN, P. W. and WHITT, W. (1992). The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability* **2** 180–198.
- [15] GOETZ, S. M., ALAVI, S. M., DENG, Z.-D. and PETERCHEV, A. V. (2019). Statistical model of motor-evoked potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27** 1539–1545.
- [16] GÖTZ, S., WHITING, P. and PETERCHEV, A. (2011). Threshold estimation with transcranial magnetic stimulation: algorithm comparison. *Clinical Neurophysiology* **122** S197.
- [17] HANS, O. and SPACEK, A. (1960). Random fixed point approximation by differentiable trajectories. In *Trans. 2nd Prague Conf. Information Theory* 203–213. Publ. House Czechoslovak Acad. Sci Prague.
- [18] HIGGINS, J. J. (2004). *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA.
- [19] HODGES, J. L. and LEHMANN, E. L. (1956). Two approximations to the Robbins–Monro process. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* **3** 95–105. University of California Press.

- [20] IOOSS, B. and LONCHAMPT, J. (2021). Robust tuning of Robbins-Monro algorithm for quantile estimation – Application to wind-farm asset management. In *ESREL 2021*.
- [21] JONES, L. A. and TAN, H. Z. (2013). Application of Psychophysical Techniques to Haptic Research. *IEEE Transactions on Haptics* **6** 268–284. <https://doi.org/10.1109/TOH.2012.74>
- [22] JOSEPH, V. R. (2004). Efficient Robbins–Monro procedure for binary data. *Biometrika* **91** 461–470.
- [23] KALLIANPUR, G. (1954). A note on the Robbins-Monro stochastic approximation method. *The Annals of Mathematical Statistics* **25** 386–388.
- [24] KESTEN, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics* 41–59.
- [25] KRISHNAIAH, P. (1969). Simultaneous test procedures under general MANOVA models. *Multivariate analysis-II*.
- [26] KUSHNER, H. J. and YIN, G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer.
- [27] LAI, T. L. (2003). Stochastic approximation. *The Annals of Statistics* **31** 391–406.
- [28] LI, Z., PETERCHEV, A. V., ROTHWELL, J. C. and GOETZ, S. M. (2022). Detection of motor-evoked potentials below the noise floor: rethinking the motor stimulation threshold. *Journal of Neural Engineering* **19** 056040.
- [29] LJUNG, L., PFLUG, G. and WALK, H. (2012). *Stochastic approximation and optimization of random systems* **17**. Birkhäuser.
- [30] MA, K., HAMADA, M., DI LAZZARO, V., HAND, B., GUERRA, A., OPIE, G. M. and GOETZ, S. M. Correlating active and resting motor thresholds for transcranial magnetic stimulation through a matching model. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*.
- [31] MARTI, K. (2003). Stochastic optimization methods in optimal engineering design under stochastic uncertainty. *Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM)* **83** 795–811.
- [32] MENGOLI, P. (1650). *Novae quadraturae arithmeticae, seu de additione fractionum*. Iacobus Montius, Bolognia.
- [33] MOULINES, E. and BACH, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems* **24**.
- [34] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* **19** 1574–1609.
- [35] NEMIROVSKI, A. and YUDIN, D. (1978). On Cezari’s convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Mathematics. Doklady* **19** 258–269.
- [36] NEMIROVSKI, A. S. and YUDIN, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience.
- [37] POLYAK, B. (1976). Convergence and convergence rate of iterative stochastic algorithms. 1. general case. *Automation and Remote Control* **37** 1858–1868.
- [38] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics* 400–407.
- [39] ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. *Optimizing Methods in Statistics* 233–257.
- [40] RUPPERT, D. (1985). A Newton-Raphson version of the multivariate Robbins-Monro procedure. *The Annals of Statistics* **13** 236–245.
- [41] SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics* **29** 373–405.
- [42] SIELKEN JR, R. L. and STATISTICS, F. S. U. T. D. O. (1973). Some stopping times for stochastic approximation procedures. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **27** 79–86.
- [43] SORENSON, H. W. and ALSPACH, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica* **7** 465–479.
- [44] SPALL, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons.
- [45] STROUP, D. F. and BRAUN, H. I. (1982). On a new stopping rule for stochastic approximation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **60** 535–554.
- [46] TOULIS, P., HOREL, T. and AIROLDI, E. M. (2021). The proximal Robbins–Monro method. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83** 188–212.
- [47] TREUTWEIN, B. (1995). Adaptive psychophysical procedures. *Vision Research* **35** 2503–2522.
- [48] VENTER, J. (1967). An extension of the Robbins-Monro procedure. *The Annals of Mathematical Statistics* **38** 181–190.
- [49] WADA, T. and FUJISAKI, Y. (2015). A stopping rule for stochastic approximation. *Automatica* **60** 1–6.

- [50] WADA, T., ITANI, T. and FUJISAKI, Y. (2010). A stopping rule for linear stochastic approximation. *IEEE Conference on Decision and Control (CDC)* **49** 4171–4176.
- [51] WANG, B., PETERCHEV, A. V. and GOETZ, S. M. (2022). Analysis and Comparison of Methods for Determining Motor Threshold with Transcranial Magnetic Stimulation. *bioRxiv* 495134. <https://doi.org/10.1101/2022.06.26.495134>
- [52] WASSERMANN, E. M. (2002). Variation in the response to transcranial magnetic brain stimulation in the general population. *Clinical Neurophysiology* **113** 1165–1171.
- [53] WEI, C. (1987). Multivariate adaptive stochastic approximation. *The annals of statistics* 1115–1130.
- [54] WU, C. J. (1985). Efficient sequential designs with binary data. *Journal of the American statistical Association* **80** 974–984.
- [55] XIONG, C. and XU, J. (2018). Efficient Robbins–Monro procedure for multivariate binary data. *Statistical Theory and Related Fields* **2** 172–180.
- [56] YIN, G. (1988). A stopped stochastic approximation algorithm. *Systems & Control Letters* **11** 107–115.
- [57] ZHANG, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on machine learning* 116.