Self-Repellent Random Walks on General Graphs - Achieving Minimal Sampling Variance via Nonlinear Markov Chains (Extended Abstract)*

Vishwaraj Doshi¹, Jie Hu² and Do Young Eun²

¹IQVIA Inc.
²North Carolina State University
vishwaraj.doshi@iqvia.com, {jhu29, dyeun}@ncsu.edu

Abstract

We consider random walks on discrete state spaces, such as general undirected graphs, where the random walkers are designed to approximate a target quantity over the network topology via sampling and neighborhood exploration in the form of Markov chain Monte Carlo (MCMC) procedures. Given any Markov chain corresponding to a target probability distribution, we design a self-repellent random walk (SRRW) which is less likely to transition to nodes that were highly visited in the past, and more likely to transition to seldom visited nodes. For a class of SRRWs parameterized by a positive real $\alpha$, we prove that the empirical distribution of the process converges almost surely to the target (stationary) distribution of the underlying Markov chain kernel. We then provide a central limit theorem and derive the exact form of the arising asymptotic co-variance matrix, which allows us to show that the SRRW with stronger repellence (larger $\alpha$) always achieves a smaller asymptotic covariance, in the sense of Loewner ordering of co-variance matrices. Especially for SRRW-driven MCMC algorithms, we show that the decrease in the asymptotic sampling variance is of the order $O(1/\alpha)$, eventually going down to zero. After generalizing these results for a class of weighted empirical measures, we use them as a stepping stone to show that a similar performance ordering can also be obtained for distributed stochastic optimization tasks using token algorithms. More explicitly, by replacing a Markovian token by a SRRW version with the same target distribution, we show that the asymptotic co-variance of the optimization iterates decreases at rate $O(1/\alpha^2)$ - the performance benefit of using SRRW thereby amplified in the stochastic optimization context. Empirical results support our theoretical findings.

1 Introduction

Random walk-based techniques are a staple in statistics and learning theory. Markov chains such as the Metropolis Hastings random walk, designed to achieve any given target probability distribution as its stationary measure, are widely used as Markov chain Monte Carlo (MCMC) samplers and in distributed optimization via stochastic gradient descent [Sun et al., 2018; Hu et al., 2022]. The local nature of the information required to compute state transition probabilities means that the algorithms scale well and are robustly implementable over state spaces such as large graphs/networks with general topologies. However, classic Markov chains can often be victims of limitations set by the underlying topology of the state space (communication matrix or adjacency matrix of the underlying network structure) leading to correlated samples which can negatively affect the estimator performance. It has also been well established that the time-reversibility requirement for the classical MCMC samplers is one of the causes for their slow convergence (see Section 1 [Andrieu and Livingstone, 2021]). One way in which this problem has been approached in the literature is via construction of non-reversible versions of the base Markov chain [Diaconis et al., 2000; Turitsyn et al., 2011; Chen and Hwang, 2013; Ma et al., 2016; Thin et al., 2020], which is often done by inducing some form of non-backtracking behaviour, that is, avoiding states most recently visited by the random walker [Alon et al., 2007]. This involves the random walker interacting with some of its own past history, and has been shown to possess better efficiency than the original base Markov chain in the sense of the MCMC estimator achieving a smaller asymptotic variance [Neal, 2004; Lee et al., 2012]. Since these non-backtracking based methods only utilize the most recent history of the random walker and are still provably more efficient, it is natural to consider the design of protocols where the random walker interacts with its entire past history to speed up its diffusion and increase its sampling efficiency, especially for sampling over discrete state spaces. This is the approach taken in our paper.

Let $G(\mathcal{N}, \mathcal{E})$ be an undirected, connected graph where $\mathcal{N} \triangleq \{1, \cdots, N\}$ denotes the set of nodes and $\mathcal{E}$ denotes the set of edges, where we say $(i,j) \in \mathcal{E}$ if there is an edge between nodes $i, j \in \mathcal{N}$. We use $A = [a_{ij}]_{i,j\in\mathcal{N}}$ to represent the adjacency matrix of the graph, where $a_{ij} > 0$ if...
The **SRRW transition kernel**: Consider the Markov chain kernel (transition matrix) \( K[x] \triangleq \{ K_{ij}[x]\}_{i,j \in \mathcal{N}} \), whose transition probabilities are mappings \( K_{ij} : \Sigma \to [0,1] \), given by
\[
K_{ij}[x] = \frac{P_{ij}r_{\mu_i}(x_j)}{\sum_{k \in \mathcal{N}} P_{ik}r_{\mu_k}(x_k)},
\]
for any probability vector \( x \triangleq \{ x_i \}_{i \in \mathcal{N}} \in \Sigma \). Here, \( \{ r_{\mu_i} \}_{i \in \mathcal{N}} \) is a family of positive functions \( r_{\mu_i} : [0,1] \to \mathbb{R}_+ \), parameterized by \( \mu_i \), with \( r_{\mu_i}(x_i) \) decreasing in \( x_i \in [0,1] \) and \( r_{\mu_i}(\mu_i) = C \), for all \( i \in \mathcal{N} \). Transition probability kernels defined in this fashion, taking probability distributions as argument, are called ‘nonlinear’ Markov kernels [Andrieu et al., 2007; Andrieu et al., 2011], as opposed to classical Markov chains with kernels \( P \) that are often interpreted as linear operators – the transition probabilities at each step being independent of \( x \) (i.e., the case where \( r_{\mu_i}(\cdot) \) is a constant function).

Stochastic processes utilizing nonlinear Markov kernels are called nonlinear Markov chains, and can be simulated/generated using **self-interacting Markov chains** (SIMCs) (see [Del Moral and Miclo, 2004; Del Moral and Miclo, 2006; Moral and Doucet, 2010]). Let \( \{ X_n \}_{n \geq 0} \) be a random walker over \( \mathcal{N} \), and let \( x_n \) be its \( \delta \)-measure or **historical empirical distribution** up to time \( n \geq 0 \), written as
\[
x_n \triangleq \frac{1}{n+1} \sum_{k=0}^{n} \delta_{X_k},
\]
where \( \delta_{X_k} \) is the delta measure whose \( k \)-th entry is one and the rest are zero, thus recording the position of the random walker at time \( k \geq 0 \). The process \( \{ X_n \}_{n \geq 0} \) becomes a SIMC if at each time step \( n \geq 0 \), the random walker makes transitions according to some nonlinear kernel \( K[x_n] \), not necessarily as defined in (1). We say that the process \( \{ X_n \}_{n \geq 0} \) is a SIMC if it is a SIMC with \( K[x_n] \) as in (1). We use the term **self-repellent** since at each time step, the transition probability to a node \( j \in \mathcal{N} \) is proportional to \( r_{\mu_j}(x_j) \), and is thus a decreasing function of the visit count to \( j \in \mathcal{N} \). In other words, the walker is less likely to move to a node that has been visited more often so far (thus self-repellent), as shown in Figure 1.

When \( P_{ij} \propto a_{ij} \) in (1) for each \( i \in \mathcal{N} \), the SRRW is a self-repellent version of the well-known **simple random walk** (SRW) procedure, with the target distribution being proportional to the degree of the nodes, that is, \( \mu_i \propto \deg(i) \) for all \( i \in \mathcal{N} \). Like most general MCMC procedures, the SRRW kernel can also be defined for **any given** sampling distribution \( \mu \in \Int(\Sigma) \), for instance, by setting \( P \) to be the transition matrix of a Metropolis Hastings Random Walk (MHRW) with stationary distribution \( \mu \). For example if \( \mu_i = 1/|N| \), then we can choose \( P_{ij} = \min\left\{ \frac{1}{\deg(i)} \frac{1}{\deg(j)} \right\} \) for all \( (i,j) \in \mathcal{E} \), with \( P_{ii} = 1 - \sum_{j \neq i} P_{ij} \). The matrix \( P \) defined in this manner is the MHRW kernel with the uniform distribution as its stationary measure, and is among the most commonly used kernels for unbiased graph sampling [Lee et al., 2012; Li et al., 2015] and distributed optimization [Sun et al., 2018]. The elegance in the Metropolis Hastings algorithm and the key to its widespread adaptation lies in the fact that at each time step, the entries of \( \mu \) need only to be known for the neighbouring nodes of the random walkers (that is, only local information required), and only up to a constant multiple. This property ensures a robust, scalable implementation of the MHRW, since global constants are often unknown for large networks a priori.

![Figure 1: Diagram of SRRW. The green bars indicate the number of previous visits to each node, and the red arrow indicates tendency of moving towards neighbors visited less often in the past.](image-url)
i, j ∈ N

(i) Computing $K_{ij}[x]$ only requires knowing $\mu_k$ for $k \in \mathcal{N}(i)$, and only up to a constant multiple for any $i \in \mathcal{N}$;

(ii) $K_{ij}[C'[x] = K_{ij}[x]$ for any constant $C' > 0$.

Indeed, we show in [Doshi et al., 2023, Appendix D] that out of all possible forms for the functions $r_{\mu}(\cdot)$, only the polynomial form as in (3) possesses the S.I. property. Henceforth, we restrict ourselves to the polynomial form of $r_{\mu}(\cdot)$.

2 Our Contributions

2.1 Almost Sure Convergence and CLT

We show that given any MCMC kernel $P$ which samples from a target distribution $\mu$, the corresponding SRRW is asymptotically more efficient as a random walk-based sampler. We do this by first showing that

$$x_n \xrightarrow{\text{a.s.}} \mu, \quad \forall \alpha \geq 0. \quad (4)$$

We then provide second-order convergence results in the form of a central limit theorem (CLT); that is, we show that there exists an asymptotic co-variance matrix $V(\alpha) \in \mathbb{R}^{N \times N}$ parameterized by $\alpha \geq 0$, such that

$$\sqrt{n}(x_n - \mu) \xrightarrow{\text{dist.}} N(0, V(\alpha)). \quad (5)$$

We obtain these results by first viewing the SRRW as a stochastic approximation (SA) algorithm with state-dependent noise [Harold J. Kushner, 1997; Fort, 2015], allowing us to form a connection between the stochastic process and a deterministic system of ordinary differential equations (ODEs). We establish global convergence results for this ODE system, which lay the foundations for proving the almost sure convergence of the stochastic SRRW process.

2.2 Co-Variance Ordering in Parameter $\alpha$

For any $\alpha \geq 0$ we derive the exact form of $V(\alpha)$ in terms of $\alpha$ and the spectrum (eigenvalues and eigenvectors) of $P$. This allows us to show that kernels parameterized by larger $\alpha$ are asymptotically more efficient samplers; that is, they achieve smaller sampling variance. This is done by showing that the asymptotic covariance matrices follow a Loewner ordering:

$$V(\alpha_1) \prec_L V(\alpha_2), \quad \forall \alpha_1 > \alpha_2 \geq 0. \quad (6)$$

In other words, as long as the numerical/computational stability of the random walk implementation can be ensured, larger values of $\alpha$ are always more favourable in terms of achieving a smaller (asymptotic) sampling variance. In [Doshi et al., 2023, Corollary 4.7], we also derive an upper bound on the ratio of its sampling variance to that of the baseline Markov chain, and show that this upper bound goes down to zero as $\alpha \to \infty$ with speed $O(1/\alpha)$. This is surprising because asymptotically for large enough $\alpha$, the SRRW, which is a stochastic process whose trajectories are constrained by ‘walking’ on the underlying communication matrix of the network, achieves smaller sampling variance than an i.i.d. sampler\footnote{The corresponding to a sampler that can visit any node $i$ with probability $\mu_i$, independent of its previous position at any given time. Clearly, in the graph setting, this requires the sampler to ‘jump’ to any other node by ignoring the underlying network structure altogether – something which random walkers on general graphs are not allowed to do.} whose variance is always a constant positive value.

2.3 Application to Distributed Stochastic Optimization

We study a family of distributed stochastic optimization algorithms, known as token algorithms, where gradients are sampled by a token traversing a network of agents in random-walk fashion [Sun et al., 2018; Hu et al., 2022; Even, 2023]. Typically, these random-walks are chosen to be Markov chains that asymptotically sample from the desired target distribution $\mu$, and play a critical role in the convergence of the optimization iterates. Our paper [Hu et al., 2024], as illustrated in Figure 2, takes a novel approach by replacing the standard linear Markovian token by SRRW - a nonlinear Markov chain. The SRRW-driven token algorithm is described as follows.

\[\begin{align*}
\text{Draw:} \quad X_{n+1} &= K X_n, [x_n] \\
\text{Update:} \quad x_{n+1} &= x_n + \gamma_{n+1} (X_{n+1} - x_n), \quad (7a) \\
\theta_{n+1} &= \theta_n + \beta_{n+1} H(\theta_n, X_{n+1}), \quad (7b)
\end{align*}\]
where \( \{ \beta_n \} \) and \( \{ \gamma_n \} \) are step size sequences decreasing to zero, and \( K[x] \) is the SRRW kernel (1) with \( r_\mu \), as in (3). Iterations (7a) and (7b) (i.e. without considering (7c)) correspond to a generalized SRRW process, with now \( x_n \) being a weighted empirical measure. Setting \( \gamma_n = 1/(n+1) \) coincides with the unweighted case discussed earlier. The \( H(\theta_n, X_n) \) terms in the update rule (7c) driving the optimization iterates \( \{ \theta_n \} \) embeds gradient information such that solving \( E_{X_n-\mu}[H(\theta^*, X)] = 0 \) is equivalent to obtaining a local minimizer for the distributed optimization problem. This is because (7c) corresponds to the more general stochastic approximation (SA) iteration, and numerous stochastic optimization algorithms which can be expressed as token algorithms, such as stochastic gradient descent (SGD) [Hu et al., 2022] and stochastic heavy ball (SHB) [Gadat et al., 2018; Li et al., 2022], can be expressed as a special case of SA.

Thus, since the update rule (7) is essentially a SA algorithm with SRRW driven noise sequence, we call it SA-SRRW. The detailed algorithmic setup and model assumptions on SA-SRRW can be found in [Hu et al., 2024, Sections 1 and 2].

The update rule (7) is in fact an example of a two-timescale stochastic approximation with state dependent (controlled) Markov noise. For any \( \alpha \geq 0 \), we prove almost sure convergence of the optimization iterates \( \theta_n \), as well as a CLT. That is, we show that

\[
\theta_n \xrightarrow{n \to \infty} \theta^* \quad \text{and} \quad (\theta_n - \theta^*)/\sqrt{\beta_n} \xrightarrow{n \to \infty} V_\theta(\alpha),
\]

where \( V_\theta(\alpha) \) is the asymptotic co-variance matrix for the iterates \( \{ \theta_n \} \).

Our key result involves the case when \( \beta_n = o(\gamma_n) \), for which we say that \( \theta_n \) is on a slower timescale compared to \( x_n \). In this case, in addition to proving a co-variance ordering as in Section 2.2, i.e.,

\[
V_\theta(\alpha_1) \preceq L V_\theta(\alpha_2), \quad \forall \alpha_1 > \alpha_2 \geq 0,
\]

we also prove that \( V_\theta(\alpha) \) decreases to zero at a rate of \( O(1/\alpha^2) \) [Hu et al., 2024, Proposition 3.4]. This is especially surprising, since we achieve an amplification in performance gain when using the SRRW for distributed stochastic optimization, as compared to the Monte Carlo sampling application where the performance gain in \( \alpha \) is in \( O(1/\alpha) \).

3 Conclusion

Our work [Doshi et al., 2023] introduces the SRRW as a drop-in replacement for any Markov chain with stationary distribution \( \mu \), and shows that it achieves better asymptotic performance for Monte Carlo sampling tasks as the degree of repellency \( \alpha \) increases. Our follow-up work [Hu et al., 2024] extends the scope from sampling to distributed stochastic optimization (and stochastic approximations in general), showing that certain combinations of step-sizes also achieve accelerated performance in \( \alpha \). Numerically tests show that the asymptotic performance guarantee is also realized for finite time steps, as can be seen in Figure 3. Our work is an instance where the asymptotic analysis approach allows the design of better algorithms despite the usage of unconventional noise sequences such as nonlinear Markov chains like the SRRW, for which traditional finite-time analytical approaches fall short, thus advocating their wider adoption.

Acknowledgments

This work [Doshi et al., 2023] was done primarily while Vishwaraj Doshi was with the Operations Research Graduate Program, North Carolina State University. This work was supported in part by National Science Foundation under Grant Nos. CNS-2007423 and IIS-1910749.

References


