

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

Vishwaraj Doshi<sup>1</sup>, Jie Hu<sup>2</sup> and Do Young Eun<sup>2</sup>

<sup>1</sup>IQVIA Inc.

<sup>2</sup>North 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 repulsion (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.

\*This is an abridged version of [Doshi *et al.*, 2023], recipient of the Outstanding Paper Award at ICML 2023, along with its follow-up research [Hu *et al.*, 2024] that was presented orally at ICLR 2024.

## 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  $\mathcal{G}(\mathcal{N}, \mathcal{E})$  be an undirected, connected graph where  $\mathcal{N} \triangleq \{1, \dots, N\}$  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  $\mathbf{A} = [a_{ij}]_{i,j \in \mathcal{N}}$  to represent the adjacency matrix of the graph, where  $a_{ij} > 0$  if

$(i, j) \in \mathcal{E}$ , and zero otherwise;  $\mathcal{N}(i) \triangleq \{j \in \mathcal{N} \mid (i, j) \in \mathcal{E}\}$  refers to the set of neighbors of node  $i$ ;  $\deg(i) \triangleq \sum_{j \in \mathcal{N}} a_{ij}$  will refer to the degree of each node  $i \in \mathcal{N}$ . Denote by  $\Sigma$  the  $N$ -dimensional probability simplex over  $\mathcal{N}$ , with  $\text{Int}(\Sigma)$  denoting its interior, and let  $\mathbf{P} \triangleq [P_{ij}]_{i,j \in \mathcal{N}}$  be the transition probability matrix of an ergodic, time-reversible Markov chain over  $\mathcal{N}$ , with its stationary distribution  $\boldsymbol{\mu} \triangleq [\mu_i]_{i \in \mathcal{N}}$ . Without loss of generality, we assume  $P_{ij} > 0$  if and only if  $a_{ij} > 0$ . In this setup, we design *Self-Repellent Random Walks* (SRRWs) on general graphs<sup>1</sup> indexed by a tunable parameter  $\alpha \geq 0$ , all of which can sample from  $\boldsymbol{\mu} \in \Sigma$ , and then study their sampling ‘efficiency’ as a function of  $\alpha$  (with  $\alpha = 0$  being equivalent to the baseline Markov chain with transition kernel  $\mathbf{P}$ ).

**The SRRW transition kernel:** Consider the Markov chain kernel (transition matrix)  $\mathbf{K}[\mathbf{x}] \triangleq [K_{ij}[\mathbf{x}]]_{i,j \in \mathcal{N}}$ , whose transition probabilities are mappings  $K_{ij} : \Sigma \rightarrow [0, 1]$ , given by

$$K_{ij}[\mathbf{x}] \triangleq \frac{P_{ij} r_{\mu_j}(x_j)}{\sum_{k \in \mathcal{N}} P_{ik} r_{\mu_k}(x_k)}, \quad (1)$$

for any probability vector  $\mathbf{x} \triangleq [x_i]_{i \in \mathcal{N}} \in \Sigma$ . Here,  $\{r_{\mu_i}\}_{i \in \mathcal{N}}$  is a family of positive functions  $r_{\mu_i} : [0, 1] \rightarrow \mathbb{R}_+$  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  $i \in \mathcal{N}$ .<sup>2</sup> 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  $\mathbf{P}$  that are often interpreted as linear operators – the transition probabilities at each step being independent of  $\mathbf{x}$  (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  $\mathbf{x}_n$  be its *occupational measure* or *historical empirical distribution* up to time  $n \geq 0$ , written as

$$\mathbf{x}_n \triangleq \frac{1}{n+1} \sum_{k=0}^n \delta_{X_k}, \quad (2)$$

where  $\delta_{X_k}$  is the delta measure whose  $X_k$ ’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  $\mathbf{K}[\mathbf{x}_n]$ , not necessarily as defined in (1). We say that the process  $\{X_n\}_{n \geq 0}$  is a SRRW if it is a SIMC with  $\mathbf{K}[\mathbf{x}]$  as in (1). We

<sup>1</sup>We consider graphs because they represent a generalization of (discrete) finite state spaces by imposing a communication (adjacency) matrix. The existence of an edge between two nodes (states) represents a non-zero probability of state transitions between the two nodes.

<sup>2</sup>As we shall see later,  $x_i$  will be directly proportional to the visit count to any node  $i \in \mathcal{N}$ , since  $\mathbf{x} \in \Sigma$  will be the empirical distribution of the self-repellent random walk.

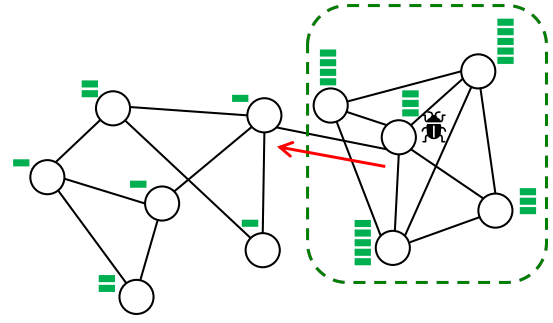


Figure 1: Diagram of SRRW. The green bars indicate the number of previous visits to each node, and the red arrow indicates tendency of moving towards neighbors visited less often in the past.

use the term *self-repellent* since at each time step, the transition probability to a node  $j \in \mathcal{N}$  is proportional to  $r_{\mu_j}([\mathbf{x}_n]_j)$  where  $[\mathbf{x}_n]_j = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_k=j\}}$ , 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  $\boldsymbol{\mu} \in \text{Int}(\Sigma)$ , for instance, by setting  $\mathbf{P}$  to be the transition matrix of a Metropolis Hastings Random Walk (MHRW) with stationary distribution  $\boldsymbol{\mu}$ . For example if  $\mu_i = 1/N$ , that is  $\boldsymbol{\mu} = \frac{1}{N} \mathbf{1}$  – the uniform distribution over the set of nodes  $\mathcal{N}$ , 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  $\mathbf{P}$  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  $\boldsymbol{\mu}$  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.

Our SRRW construction begins with  $r_{\mu_i}(\cdot)$  taking a polynomial form for all  $i \in \mathcal{N}$ , given by

$$r_{\mu_i}(x_i) \triangleq \left( \frac{x_i}{\mu_i} \right)^{-\alpha}, \quad \forall \alpha \geq 0, \quad (3)$$

where the parameter  $\alpha \geq 0$  can be perceived as the *strength of the self-repellence mechanism* designed into the SRRW transition kernel. Similar to the MHRW transition kernel, only the local information regarding entries of  $\boldsymbol{\mu}$  needs to be known at any given time step, and up to only a constant multiple. For convenience, we formalize this property as *scale-invariance (S.I.)*: an SRRW kernel possesses S.I. if for all

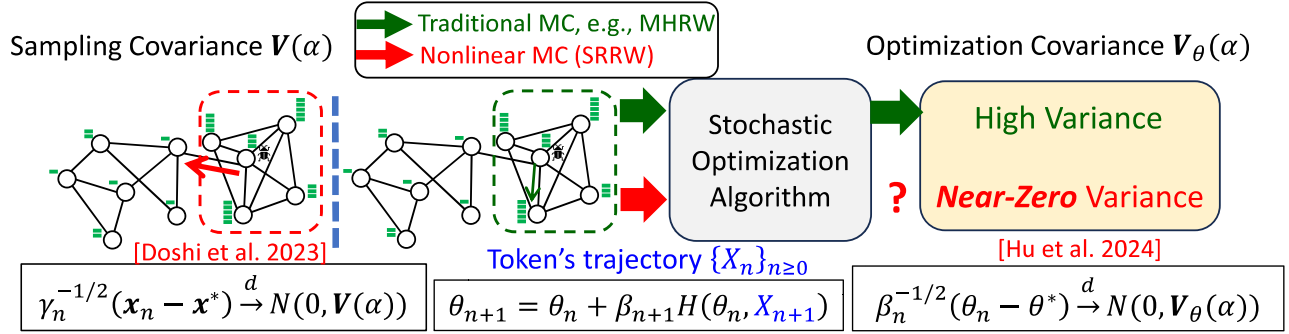


Figure 2: Visualization of token algorithms using SRRW versus traditional MC in distributed learning. The SRRW has a smaller tendency to get trapped within sub-regions of the graph, leading to smaller asymptotic covariance.

$i, j \in \mathcal{N}$

- (i) Computing  $K_{ij}[\mathbf{x}]$  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)  $K_{ij}[C'\mathbf{x}] = K_{ij}[\mathbf{x}]$  for any constant  $C' > 0$ .

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

## 2 Our Contributions

### 2.1 Almost Sure Convergence and CLT

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

$$\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \boldsymbol{\mu}, \quad \forall \alpha \geq 0. \quad (4)$$

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*  $\mathbf{V}(\alpha) \in \mathbb{R}^{N \times N}$  parameterized by  $\alpha \geq 0$ , such that

$$\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}) \xrightarrow[n \rightarrow \infty]{\text{dist.}} N(\mathbf{0}, \mathbf{V}(\alpha)). \quad (5)$$

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  $\mathbf{V}(\alpha)$  in terms of  $\alpha$  and the spectrum (eigenvalues and eigenvectors) of  $\mathbf{P}$ . 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*:<sup>3</sup>

$$\mathbf{V}(\alpha_1) <_L \mathbf{V}(\alpha_2), \quad \forall \alpha_1 > \alpha_2 \geq 0. \quad (6)$$

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  $\alpha \rightarrow \infty$  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<sup>4</sup> 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  $\boldsymbol{\mu}$ , 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.

$$\text{Draw: } X_{n+1} \sim \mathbf{K}_{X_n, \cdot}[\mathbf{x}_n] \quad (7a)$$

$$\text{Update: } \mathbf{x}_{n+1} = \mathbf{x}_n + \gamma_{n+1}(\boldsymbol{\delta}_{X_{n+1}} - \mathbf{x}_n), \quad (7b)$$

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \beta_{n+1}H(\boldsymbol{\theta}_n, X_{n+1}), \quad (7c)$$

<sup>3</sup>Matrices  $\mathbf{A}, \mathbf{B}$  follow the Loewner ordering  $\mathbf{A} <_L \mathbf{B}$  if  $\mathbf{A} \neq \mathbf{B}$  and  $\mathbf{B} - \mathbf{A}$  is positive semi-definite.

<sup>4</sup>This corresponds 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 permitted to do.

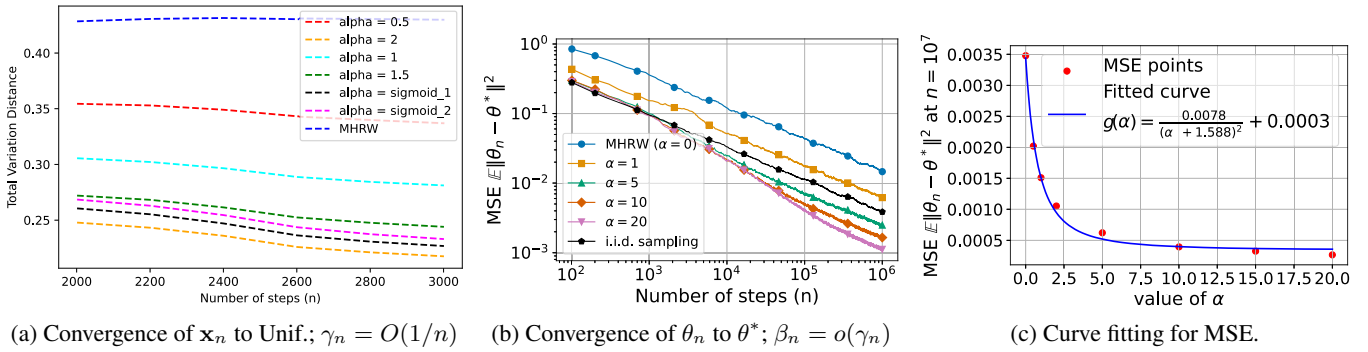


Figure 3: Simulations of SRRW: (a) shows convergence of the empirical measure  $\mathbf{x}_n$  to a Uniform distribution target (Unif.), (b) is SA-SRRW, where (7c) follows an SGD iteration, applied to  $L_2$ -regularized binary classification problem. Both (a) and (b) show the expected performance ordering for various  $\alpha$  values, while (c) shows MSE in (b) decreasing at  $O(1/\alpha^2)$  speed.

where  $\{\beta_n\}$  and  $\{\gamma_n\}$  are step size sequences decreasing to zero, and  $\mathbf{K}[\mathbf{x}]$  is the SRRW kernel (1) with  $r_{\mu_i}$  as in (3). Iterations (7a) and (7b) (i.e. without considering (7c)) correspond to a generalized SRRW process, with now  $\mathbf{x}_n$  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  $\mathbb{E}_{X \sim \mu}[H(\theta^*, X)] = 0$  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

$$\theta_n \xrightarrow[n \rightarrow \infty]{a.s.} \theta^*, \text{ and } (\theta_n - \theta^*)/\sqrt{\beta_n} \xrightarrow[n \rightarrow \infty]{dist.} \mathbf{V}_\theta(\alpha), \quad (8)$$

where  $\mathbf{V}_\theta(\alpha)$  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  $\mathbf{x}_n$ . In this case, in addition to proving a co-variance ordering as in Section 2.2, i.e.,

$$\mathbf{V}_\theta(\alpha_1) <_L \mathbf{V}_\theta(\alpha_2), \quad \forall \alpha_1 > \alpha_2 \geq 0, \quad (9)$$

we also prove that  $\mathbf{V}_\theta(\alpha)$  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 repency  $\alpha$  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

[Alon *et al.*, 2007] Noga Alon, Itai Benjamini, Eyal Lubetzky, and Sasha Sodin. Non-backtracking random walks mix faster. *Communications in Contemporary Mathematics*, 9(04):585–603, 2007.

[Andrieu and Livingstone, 2021] Christophe Andrieu and Samuel Livingstone. Peskun–tierney ordering for markovian monte carlo: Beyond the reversible scenario. *The Annals of Statistics*, 49(4):1958–1981, 2021.

[Andrieu *et al.*, 2007] Christophe Andrieu, Ajay Jasra, Arnaud Doucet, and Pierre Del Moral. Non-linear markov chain monte carlo. *Esaim: Proceedings*, 19:79–84, 01 2007.

- [Andrieu *et al.*, 2011] Christophe Andrieu, Ajay Jasra, Arnaud Doucet, and Pierre Del Moral. On nonlinear Markov chain Monte Carlo. *Bernoulli*, 17(3):987 – 1014, 2011.
- [Chen and Hwang, 2013] Ting-Li Chen and Chii-Ruey Hwang. Accelerating reversible markov chains. *Statistics & Probability Letters*, 83(9):1956–1962, 2013.
- [Del Moral and Miclo, 2004] Pierre Del Moral and Laurent Miclo. On convergence of chains with occupational self-interactions. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2041):325–346, 2004.
- [Del Moral and Miclo, 2006] Pierre Del Moral and Laurent Miclo. Self-interacting markov chains. *Stochastic Analysis and Applications*, 24:615–660, 07 2006.
- [Diaconis *et al.*, 2000] Persi Diaconis, Susan Holmes, and Radford M Neal. Analysis of a nonreversible markov chain sampler. *Annals of Applied Probability*, pages 726–752, 2000.
- [Doshi *et al.*, 2023] Vishwaraj Doshi, Jie Hu, and Do Young Eun. Self-repellent random walks on general graphs—achieving minimal sampling variance via nonlinear markov chains. In *International Conference on Machine Learning*. PMLR, 2023.
- [Even, 2023] Mathieu Even. Stochastic gradient descent under markovian sampling schemes. In *International Conference on Machine Learning*, 2023.
- [Fort, 2015] Gersende Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80, 2015.
- [Gadat *et al.*, 2018] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12:461–529, 2018.
- [Harold J. Kushner, 1997] G. George Yin Harold J. Kushner. *Stochastic Approximation Algorithms and Applications*. Springer, 1997.
- [Hu *et al.*, 2022] Jie Hu, Vishwaraj Doshi, and Do Young Eun. Efficiency ordering of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2022.
- [Hu *et al.*, 2024] Jie Hu, Vishwaraj Doshi, and Do Young Eun. Accelerating distributed stochastic optimization via self-repellent random walks. In *International Conference on Learning Representations*, 2024.
- [Lee *et al.*, 2012] Chul-Ho Lee, Xin Xu, and Do Young Eun. Beyond Random Walk and Metropolis-Hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling. In *Proceedings of the ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS’12, pages 319–330, 2012.
- [Li *et al.*, 2015] Rong-Hua Li, Jeffrey Xu Yu, Lu Qin, Rui Mao, and Tan Jin. On random walk based graph sampling. In *2015 IEEE 31st international conference on data engineering*, pages 927–938. IEEE, 2015.
- [Li *et al.*, 2022] Tiejun Li, Tiannan Xiao, and Guoguo Yang. Revisiting the central limit theorems for the sgd-type methods. *arXiv preprint arXiv:2207.11755*, 2022.
- [Ma *et al.*, 2016] Yi-An Ma, Tianqi Chen, Lei Wu, and Emily B Fox. A unifying framework for devising efficient and irreversible mcmc samplers. *arXiv preprint arXiv:1608.05973*, 2016.
- [Moral and Doucet, 2010] Pierre Del Moral and Arnaud Doucet. Interacting Markov chain Monte Carlo methods for solving nonlinear measure-valued equations. *The Annals of Applied Probability*, 20(2):593 – 639, 2010.
- [Neal, 2004] Radford M Neal. Improving asymptotic variance of mcmc estimators: Non-reversible chains are better. *arXiv preprint math/0407281*, 2004.
- [Sun *et al.*, 2018] Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.
- [Thin *et al.*, 2020] Achille Thin, Nikita Kotelevskii, Christophe Andrieu, Alain Durmus, Eric Moulines, and Maxim Panov. Nonreversible mcmc from conditional invertible transforms: a complete recipe with convergence guarantees. *arXiv preprint arXiv:2012.15550*, 2020.
- [Turitsyn *et al.*, 2011] Konstantin S Turitsyn, Michael Chertkov, and Marija Vucelja. Irreversible monte carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, 2011.