Adaptive Stepsizing for Stochastic Gradient Langevin Dynamics in Bayesian Neural Networks

Rajit Rajpal[1], **Benedict Leimkuhler**[1], **Yuanhao Jiang**[1]

[1]School of Mathematics, University of Edinburgh
{s2592586, B.Leimkuhler, yuanhao.jiang}@ed.ac.uk

Abstract

Bayesian neural networks (BNNs) require scalable sampling algorithms to approximate posterior distributions over parameters. Existing stochastic gradient Markov Chain Monte Carlo (SGMCMC) methods are highly sensitive to the choice of stepsize and adaptive variants such as pSGLD typically fail to sample the correct invariant measure without addition of a costly divergence correction term. In this work, we build on the recently proposed 'SamAdams' framework for timestep adaptation (Leimkuhler, Lohmann, and Whalley 2025), introducing an adaptive scheme: SA-SGLD, which employs time rescaling to modulate the stepsize according to a monitored quantity (typically the local gradient norm). SA-SGLD can automatically shrink stepsizes in regions of high curvature and expand them in flatter regions, improving both stability and mixing without introducing bias. We show that our method can achieve more accurate posterior sampling than SGLD on high-curvature 2D toy examples and in image classification with BNNs using sharp priors.

1 Introduction

Bayesian Neural Networks (BNNs) provide a framework for quantifying uncertainty in deep learning models by placing a posterior distribution over the weights, $p(\boldsymbol{\theta}|\mathcal{D})$. Algorithms like Stochastic Gradient Langevin Dynamics (SGLD) extend classical MCMC to the big-data setting by leveraging stochastic gradients. However, the loss landscape of deep neural networks is notoriously complex, characterized by pathological curvature and saddle points (Kim, Song, and Liang 2020).

Several methods have introduced adaptive step sizes or preconditioning to improve the convergence of SGMCMC on challenging loss landscapes, including geometry-based schemes such as SGRLD and SGRHMC (Patterson and Teh 2013; Ma, Chen, and Fox 2015), and practical variants like pSGLD (Li et al. 2015). However, as discussed in (Ma, Chen, and Fox 2015; Rensmeyer and Niggemann 2024) and Section 2.2, these methods are biased unless the dynamics is augmented by a computationally expensive divergence term.

Adaptive stepsizes can be viewed as an isotropic but dynamic preconditioning framework (Leroy et al. 2024). Building on the recent formulation of (Leimkuhler, Lohmann, and Whalley 2025), we revisit adaptive step size methods for SGLD in Bayesian sampling by introducing SA-SGLD. Importantly this scheme circumvents the computation of the divergence by use of statistical reweighting. We provide theoretical foundations and show using small examples and a Bayesian neural network that this method can improve performance compared to SGLD.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2 Background

2.1 Stochastic Gradient MCMC (SGMCMC)

SGMCMC is a family of sampling methods that seeks to perform efficient inference with noisy gradients as a result of using mini-batches of data. Suppose we have a dataset $\mathcal{D} = \{d_i\}_{i=1}^N$ and a model parameterized by $\boldsymbol{\theta}$. The likelihood is defined by $p(\mathcal{D}|\boldsymbol{\theta})$. In a Bayesian setting, we define the prior $p(\boldsymbol{\theta})$ and a posterior $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. This may be re-expressed as a Boltzmann-Gibbs distribution $p(\boldsymbol{\theta}|\mathcal{D}) \propto e^{-U(\boldsymbol{\theta})}$ where $U(\boldsymbol{\theta}) = \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ is the potential energy which forms the foundation for SDE-derived sampling procedures. Typically, one may use standard MCMC methods to sample from $p(\boldsymbol{\theta}|\mathcal{D})$, provided that we have access to gradients of $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$. However, for a large dataset $\mathcal{D}$, this is costly.

Therefore, we approximate $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta})$ with noisy mini-batch gradients. The most commonly used SGMCMC methods is Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh 2011). In continuous time, SGLD may be written as:

$$d\boldsymbol{\theta} = -\nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta})dt + \sqrt{2\beta^{-1}}d\boldsymbol{W}. \tag{1}$$

β^{-1} is the temperature to tune the stationary distribution $e^{-\beta U(\boldsymbol{\theta})}$. This SDE is discretized using Euler-Maruyama with decreasing stepsizes to remove the bias term asymptotically. However, fixed stepsizes are often used in practice.

2.2 Adaptive SGMCMC

Adaptive SGMCMC methods improve convergence by adapting to the local geometry of the parameter space. A prominent example is *Stochastic Gradient Riemannian Langevin Dynamics* (SGRLD) (Patterson and Teh 2013; Girolami and Calderhead 2011):

$$d\boldsymbol{\theta} = \left[-\boldsymbol{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\tilde{U}(\boldsymbol{\theta}) + \boldsymbol{\Gamma}(\boldsymbol{\theta})\right] dt + \sqrt{2\beta^{-1}}\boldsymbol{G}(\boldsymbol{\theta})^{-\frac{1}{2}}\, d\boldsymbol{W}, \tag{2}$$

where $\boldsymbol{G}(\boldsymbol{\theta})$ is a positive-definite metric (e.g. Fisher information) and $\Gamma_i(\boldsymbol{\theta}) = \sum_j \partial_j G^{-1}_{ij}(\boldsymbol{\theta})$ is the divergence correction term (Girolami and Calderhead 2011; Chung 2013). $\boldsymbol{G}(\boldsymbol{\theta})$ defines a local metric on parameter space, shaping both the drift and diffusion to follow the local curvature of the posterior, improving exploration and convergence. The divergence must be added to the drift to ensure that the sampler preserves the correct stationary distribution. It compensates for the geometric distortion induced by the metric.

Neglecting $\boldsymbol{\Gamma}$–as is done in practice for many adaptive algorithms such as pSGLD (Li et al. 2015), AdamSGLD (Kim, Song, and Liang 2022), and SGRLD in alternative metrics (Yu et al. 2023)–will break the invariance of the target distribution. These methods replace $\boldsymbol{G}(\boldsymbol{\theta})$ by an exponentially averaged estimate (e.g. RMSProp-like preconditioning) to avoid computing second-order derivatives, which effectively downscales $\boldsymbol{\Gamma}$ by $(1-\alpha)$. As shown by Rensmeyer and Niggemann (2024), the resulting SDE no longer has the correct stationary density: in one dimension, omitting $\boldsymbol{\Gamma}$ yields $\pi(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathcal{D})\, G(\boldsymbol{\theta})^{-1}$, and in the general case this results in biased samples unless $G(\boldsymbol{\theta})$ is nearly constant over regions of high posterior mass. Thus, the divergence is essential for maintaining ergodicity and convergence to the true posterior (Ma, Chen, and Fox 2015).

Other adaptive methods include ad-hoc stepsize schedules such as Cyclical SGMCMC (Zhang et al. 2020) which applies a periodic, cosine-annealed stepsize schedule to alternate between exploration of new modes and local refinement without explicit curvature estimation. Adaptive Langevin (Shang et al. 2015) and SGNHT (Ding et al. 2014) introduce an adaptive friction term to adaptively damp the gradient noise.

2.3 SAM-ADAMS

SamAdams (Leimkuhler, Lohmann, and Whalley 2025) is a recently proposed adaptive method for kinetic Langevin dynamics, that adapts the stepsize based on the local geometry of the loss landscape while preserving stability of the dynamics. The key idea is to incorporate a state-dependent time-rescaling. This can be achieved by introducing a new physical time variable τ such that $\frac{dt}{d\tau} = \psi(\zeta_\tau)$ where $\psi(.)$ defines a Sundman time transformation. The choice of ζ can, in principle, be arbitrarily chosen. However, a smart choice would be an exponential average of some monitor function $g(\boldsymbol{\theta}_\tau)$ that captures local geometry. In (Leimkuhler, Lohmann, and Whalley 2025), $g(\boldsymbol{\theta}_\tau) = \|\nabla_\theta U(\boldsymbol{\theta})\|^2$ and $\psi(\zeta) = m\frac{\zeta^r + \frac{M}{m}}{\zeta^r + 1}$. This choice of $\psi(.)$ bounds the adaptive stepsize $m\Delta\tau \leq \Delta t \leq M\Delta\tau$. Overall, SamAdams may be written in continuous time as

$$d\boldsymbol{\theta} = \boldsymbol{p}dt \tag{3}$$

$$d\boldsymbol{p} = (-\nabla U(\boldsymbol{\theta}) - \gamma\boldsymbol{p})dt + \sqrt{2\beta^{-1}\gamma}d\boldsymbol{W}_t \tag{4}$$

$$dt = \psi(\zeta_\tau)d\tau \tag{5}$$

$$d\zeta_\tau = -\alpha\zeta_\tau d\tau + g(\boldsymbol{\theta}_\tau)d\tau \tag{6}$$

Unlike fixed or scheduled stepsizes, $\eta(\boldsymbol{\theta})$ automatically contracts in regions of high curvature and expands in flatter regions, thus improving both stability and mixing. The invariant measure $e^{-\beta(U(\boldsymbol{\theta})+\frac{1}{2}\boldsymbol{p}^T\boldsymbol{p})}$ is preserved as time is only rescaled. It is important to note that ζ here is a scalar and does not adapt each degree of freedom separately.

3 Methodology

The SamAdams algorithm in (Leimkuhler, Lohmann, and Whalley 2025) was introduced with exact gradients and in the kinetic Langevin setting. While that article mentions possible applications with noisy gradients the experiments in BNNs used a high accuracy approximate gradient. The algorithm presented here (SA-SGLD) is specifically viewed as a noisy gradient scheme for large scale training and is based on the following equations modifying SGLD:

$$\begin{aligned}
\zeta_{n+1} &= \rho\zeta_n + \frac{(1-\rho)}{\alpha}g(\boldsymbol{\theta}_n) \\
\Delta t_{n+1} &= \psi(\zeta_{n+1})\Delta\tau \\
\boldsymbol{\theta}_{n+1} &= -\nabla\tilde{U}(\boldsymbol{\theta}_n)\Delta t_{n+1} + \sqrt{2\beta^{-1}\Delta t_{n+1}}\varepsilon
\end{aligned}$$

where $\rho = \exp(-\alpha\Delta\tau)$, $g(\boldsymbol{\theta}_\tau) = \|\nabla_\theta\tilde{U}(\boldsymbol{\theta})\|^2 + \delta$, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. δ is a smaller regularization term to lower bound ζ from 0. SA-SGLD may be viewed as a middle ground between two extremes. On the one hand, full Riemannian methods such as SGRLD (Patterson and Teh 2013) incorporate a metric tensor $G(\boldsymbol{\theta})$ and a divergence correction term to preserve the stationary distribution, but the correction is prohibitively expensive in high dimensions. On the other hand, heuristic schemes like pSGLD (Li et al. 2015) adapt learning rates using optimization-inspired preconditioners (e.g. RMSProp), but neglect the correction term, which biases the stationary distribution. SamAdams achieves a compromise: it adapts stepsizes in a principled manner, avoiding the costly divergence computation, and retains the correct stationary distribution. This balance makes it both practical and robust for Bayesian deep learning. The additional computational cost is negligible as one only needs to compute the gradient norm and store a scalar.

4 Theoretical Results

The proofs for the following theorems are in Appendix B

Lemma 1 (Uniform moment bounds)**.** *Assume the following:*

1. *The monitor uses $g(\theta_n) = \|G_n\|^2 + \delta$, so that*
$$\zeta_{n+1} = \rho\zeta_n + \frac{1-\rho}{\alpha}g(\theta_n), \quad \rho = e^{-\alpha h}.$$
2. *The adaptive time step is $\Delta t_{n+1} = \psi(\zeta_{n+1})h$, where ψ is bounded and globally Lipschitz:*
$$\begin{aligned} 0 < m \leq \psi(\zeta) \leq M < \infty, \\ |\psi(x) - \psi(y)| \leq L_\psi|x-y|. \end{aligned}$$
3. *The potential U is L-smooth and dissipative:*
$$\begin{aligned} \|\nabla U(\theta) - \nabla U(\theta')\| \leq L\|\theta - \theta'\|, \\ \langle\theta, \nabla U(\theta)\rangle \geq a\|\theta\|^2 - b, \end{aligned}$$

for $L > 0$, $a > 0$, $b \geq 0$.

4. *The stochastic gradients satisfy, for some $\sigma < \infty$:*

$$\mathbb{E}[G_n|\mathcal{F}_n] = \nabla U(\theta_n),$$
$$\mathbb{E}[\|G_n - \nabla U(\theta_n)\|^2|\mathcal{F}_n] \le \sigma^2(1 + \|\theta_n\|^2),$$

where $\mathcal{F}_n = \sigma(\theta_0, \zeta_0, \varepsilon_1, \ldots, \varepsilon_n, G_0, \ldots, G_{n-1})$.

Define the constants:

$$C_1 := 2L^2 + 2\sigma^2,$$
$$C_2 := 2\|\nabla U(0)\|^2 + 2\sigma^2,$$
$$C_3 := 2\beta^{-1}d.$$

If $h > 0$ is small enough that

$$\gamma(h) := 2amh - C_1M^2h^2 - 2\sigma Mh > 0, \tag{7}$$

then the iterates of

$$\theta_{n+1} = \theta_n - \Delta t_{n+1}G_n + \sqrt{2\beta^{-1}\Delta t_{n+1}}\,\varepsilon_{n+1}, \tag{8}$$
$$\varepsilon_{n+1} \sim \mathcal{N}(0, I_d), \tag{9}$$

satisfy the uniform moment bound

$$\sup_{n\ge 0}\mathbb{E}\|\theta_n\|^2 < \infty.$$

Theorem 1 (Ergodicity and $O(h)$ bias)**.** *Retain the assumptions of Theorem 1. Further assume:*

1. *For $p > 0$ large, $\sup_{n\ge 0}\mathbb{E}\|\theta_n\|^p < \infty$*
2. $\mathbb{E}[\|G_n - \nabla U(\theta_n)\|^4|\mathcal{F}_n] \le \sigma_4^2(1 + \|\theta_n\|^4)$.
3. *(θ_n, ζ_n) is ergodic Markov with invariant measure $\widetilde{\pi}_h$.*
4. *$U \in C^4(\mathbb{R}^d)$ with bounded derivatives, ∇U Lipschitz.*

Let $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathcal{L}\phi = f - \pi(f), \quad \mathcal{L} = -\nabla U \cdot \nabla + \beta^{-1}\Delta,$$

admits $\phi \in C^4(\mathbb{R}^d)$ with polynomial-growth derivatives:

$$\sup_\theta \frac{\|D^j\phi(\theta)\|}{1 + \|\theta\|^q} \le A_j, \quad j = 0, 1, 2, 3, 4.$$

Define the weighted time-average:

$$\mathcal{A}_n := \frac{\sum_{k=1}^n \Delta t_k f(\theta_k)}{\sum_{k=1}^n \Delta t_k}, \quad \Delta t_k = \psi(\zeta_k)h.$$

Let the ψ-weighted marginal be $\pi_h(f) := \frac{\widetilde{\pi}_h[\psi(\zeta)f(\theta)]}{\widetilde{\pi}_h[\psi(\zeta)]}$. Then, For every f:

$$\mathcal{A}_n \xrightarrow{n\to\infty} \text{a.s.}\pi_h(f).$$

There exists $C > 0$ such that

$$|\pi_h(f) - \pi(f)| \le Ch.$$

Moreover, for all $n \ge 1$:

$$\left|\mathbb{E}\mathcal{A}_n - \pi(f)\right| \le |\pi_h(f) - \pi(f)| + \frac{C'}{n} = O(h) + O(n^{-1}).$$

5 Experiments

5.1 Motivational Examples

We illustrate the benefit of SA-SGLD by mimicking two common challenges in Neural Network loss landscapes: multi-modal adaptation and high curvature regions, with 2D representative examples below.

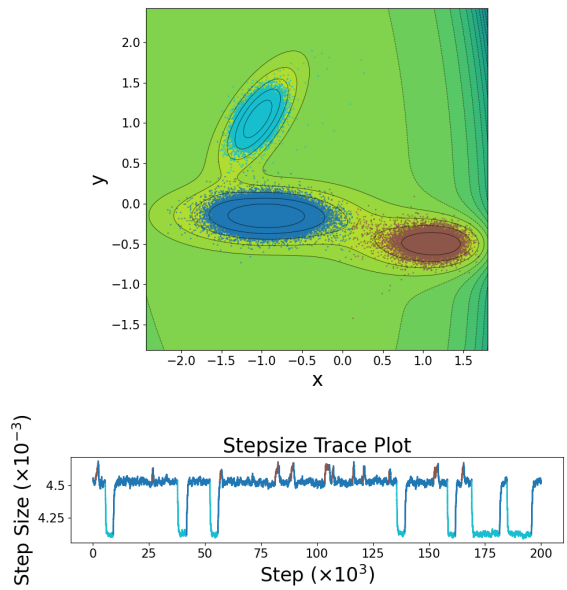


Figure 1: **Müller–Brown potential.** SA-SGLD adapts its step size to local curvature, enabling transitions across energy barriers—analogous to escaping local modes in complex BNN posteriors.

Müller–Brown potential. The Müller–Brown potential (Müller and Brown 1979) is a classical benchmark for studying transitions between metastable wells. As shown in Figure 1, the SAM-ADAMS-enhanced variant (SA-SGLD) adapts its step size according to local curvature. This dynamic adjustment reflects what we would expect in high-dimensional BNN posteriors—where different modes correspond to functionally distinct solutions separated by narrow energy barriers (Garipov et al. 2018).

Star potential. The second example, the "star" potential (Leimkuhler, Lohmann, and Whalley 2025), exhibits strong anisotropy and narrow funnel-shaped regions which may be induced in a BNN posterior by incorporating sparse priors. As shown in Figure 2, SGLD oversamples the outer ridges and struggles to penetrate the high-curvature funnels, whereas SA-SGLD adapts its step size to enter these regions smoothly, producing a more balanced exploration of the posterior geometry.

Further implementation details, parameter values, and potential definitions are provided in Appendix A.

5.2 Bayesian Neural Networks

(Leimkuhler, Lohmann, and Whalley 2025) demonstrated the effect of SAM-ADAMS on kinetic Langevin dynamics for neural network training at a very large batch size (10000/50000), effectively eliminating the role of stochastic gradient noise. Moreover, they simply assessed the dynamics of the collected samples rather than making evalua-

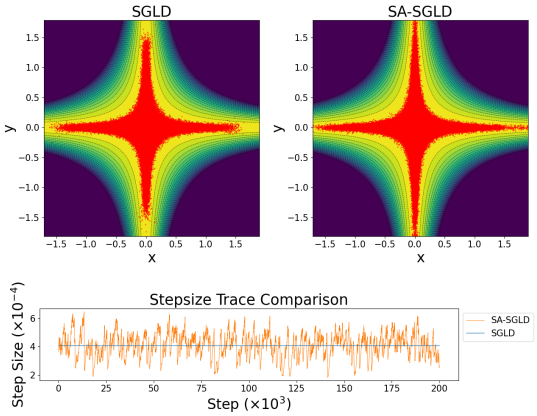


Figure 2: **Star potential.** SA-SGLD dynamically scales its step size, entering narrow high-curvature funnels that SGLD fails to explore.

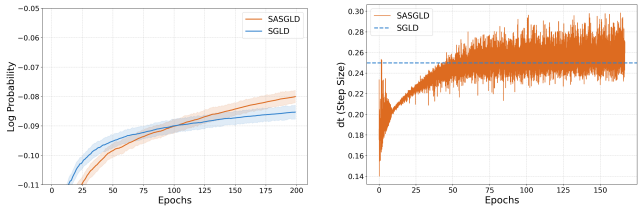
tions using the collected ensemble. As a result, we conduct experiments in a proper BNN setup with appropriate metrics (e.g ECE/NLL/Acc) after ensembling, Bayesian prior setups, and a small batch size. We use the MNIST dataset a 3 layer fully connected neural network (FCNN) at a batch size of 100. We aim to estimate the posterior predictive $p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$ with a Monte Carlo estimate $\frac{1}{n}\Sigma_{i=1}^{n} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$. According to (Fortuin et al. 2022), for FCNNs, uncorrelated heavy-tailed priors (e.g Student-t, Horseshoe) perform well. Such heavy-tailed priors induce sparsity in the parameter-space of the neural network and create funnel-like geometries in the posterior that SA-SGLD may be well-suited to adapting to. We use Cross Entropy Loss, and compare between a Horseshoe and Gaussian prior on the parameters. We run 200 epochs and collect samples with a thinning interval of 100 batches. This results in a total of 1000 models.

We evaluate performance with: negative log-likelihood (NLL), expected calibration error (ECE), and test accuracy. We selected standard values for $r = 0.25, s = 2$ and picked $m = 0.5, M = 2, \alpha = 1000, d\tau = 0.2$ for Table 1 and Fig 3. The first 100 epochs of the samples are discarded as burnin.

Method	NLL ↓	Acc (%) ↑	ECE (%) ↓
Gaussian Prior			
SGLD	$\mathbf{0.192} \pm 0.005$	95.23 ± 0.056	$\mathbf{5.72} \pm 0.27$
SA-SGLD	0.193 ± 0.004	$\mathbf{95.25} \pm 0.034$	5.76 ± 0.22
Horseshoe Prior			
SGLD	0.086 ± 0.004	98.03 ± 0.044	3.64 ± 0.11
SA-SGLD	$\mathbf{0.080} \pm 0.003$	$\mathbf{98.12} \pm 0.031$	$\mathbf{3.49} \pm 0.9$

Table 1: Comparison of predictive performance between SGLD and SA-SGLD for Gaussian and Horseshoe priors.

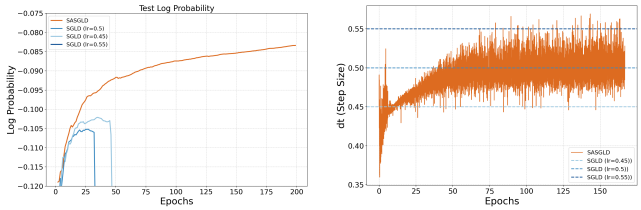
We find that for Gaussian priors, SA-SGLD does not offer much benefit. However, for a Horseshoe prior, there is a substantial benefit as indicated by improved NLL, Test Accuracy, and ECE. We suspect this is due to the high curvature induced by the Horseshoe prior. Furthermore, we observe in Fig 3 that as more ensembles are collected past the burnin period, the log-probability of SA-SGLD exceeds that of SGLD.



(a) Log Probability Curves **(b)** Stepsize Trace

Figure 3: SGLD vs SA-SGLD on sampling BNN with Horseshoe prior on MNIST data. Log Probability shown is computed with the entire ensemble until that epoch.

Fig 4 indicates that for $h \in [0.45, 0.55]$, SGLD's log probability diverges whereas SA-SGLD manages to retain performance even at a mean stepsize of 0.5. We only increased $d\tau$ to 0.35 and r to 0.5. SGLD ($h = 0.55$) diverged early and is not shown in the figure.



(a) Log Probability Curves **(b)** Stepsize Trace

Figure 4: SA-SGLD's robustness with large stepsizes.

6 Conclusion

We introduce a principled adaptive framework for stochastic gradient MCMC based on the SAM-ADAMS time-rescaling mechanism. By modulating stepsizes according to local gradient geometry, our approach preserves the correct invariant distribution while avoiding the expensive divergence corrections.

7 Limitations and Future Work

Extending SAM-ADAMS to a broader class of monitor functions or matrix-valued adaptations is a promising avenue, assuming complexity concerns can be addressed. Future work may include extending the method to other SGMCMC variants (e.g. SGHMC, SGNHT), exploring multidimensional adaptation, studying it as an optimizer in the zero-temperature limit, and applying it to wider range of probabilistic modeling tasks.

References

Chung, K. L. 2013. *Lectures from Markov processes to Brownian motion*, volume 249. Springer Science & Business Media.

Ding, N.; Fang, Y.; Babbush, R.; Chen, C.; Skeel, R. D.; and Neven, H. 2014. Bayesian Sampling Using Stochastic Gradient Thermostats. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3203–3211.

Fortuin, V.; Garriga-Alonso, A.; Ober, S. W.; Wenzel, F.; Ratsch, G.; Turner, R. E.; van der Wilk, M.; and Aitchison, L. 2022. Bayesian Neural Network Priors Revisited. In *International Conference on Learning Representations (ICLR)*.

Garipov, T.; Izmailov, P.; Podoprikhin, D.; Vetrov, D. P.; and Wilson, A. G. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31.

Girolami, M.; and Calderhead, B. 2011. Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2): 123–214.

Kim, S.; Song, Q.; and Liang, F. 2020. Stochastic Gradient Langevin Dynamics Algorithms with Adaptive Drifts. arXiv:2009.09535.

Kim, S.; Song, Q.; and Liang, F. 2022. Stochastic gradient Langevin dynamics with adaptive drifts. *Journal of Statistical Computation and Simulation*, 92(2): 318–336. PMID: 35559269.

Kushner, H. J.; and Yin, G. G. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition.

Leimkuhler, B.; Lohmann, R.; and Whalley, P. 2025. A Langevin sampling algorithm inspired by the Adam optimizer. arXiv:2504.18911.

Leroy, A.; Leimkuhler, B.; Latz, J.; and Higham, D. 2024. Adaptive Stepsize Algorithms For Langevin Dynamics. *SIAM Journal on Scientific Computing*, 46(6): A3574–A3598.

Li, C.; Chen, C.; Carlson, D.; and Carin, L. 2015. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. arXiv:1512.07666.

Ma, Y.-A.; Chen, T.; and Fox, E. B. 2015. A Complete Recipe for Stochastic Gradient MCMC. arXiv:1506.04696.

Mattingly, J. C.; Stuart, A. M.; and Tretyakov, M. V. 2010. Convergence of Numerical Time-Averaging and Stationary Measures via Poisson Equations. *SIAM Journal on Numerical Analysis*, 48(2): 552–577.

Meyn, S.; Tweedie, R. L.; and Glynn, P. W. 2009. *Convergence*, 311–312. Cambridge Mathematical Library. Cambridge: Cambridge University Press.

Müller, K.; and Brown, L. D. 1979. Location of Saddle Points and Minimum Energy Paths by a Constrained Simplex Optimization Procedure. *Theoretica chimica acta*, 53(1): 75–93.

Patterson, S.; and Teh, Y. W. 2013. Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Raginsky, M.; Rakhlin, A.; and Telgarsky, M. 2017. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, 1674–1703. PMLR.

Rensmeyer, T.; and Niggemann, O. 2024. On the Convergence of Locally Adaptive and Scalable Diffusion-Based Sampling Methods for Deep Bayesian Neural Network Posteriors. arXiv:2403.08609.

Sato, I.; and Nakagawa, H. 2014. Approximation Analysis of Stochastic Gradient Langevin Dynamics Using Fokker-Planck Equation and Ito Process. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 982–990. Beijing, China: PMLR.

Shang, X.; Zhu, Z.; Leimkuhler, B.; and Storkey, A. J. 2015. Covariance-Controlled Adaptive Langevin Thermostat for Large-Scale Bayesian Sampling. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Vollmer, S. J.; Zygalakis, K. C.; and Teh, Y. W. 2016. Exploration of the (non-)asymptotic Bias and Variance of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17(1): 5504–5548.

Welling, M.; and Teh, Y. W. 2011. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 681–688. Omnipress.

Yu, H.; Hartmann, M.; Williams, B.; and Klami, A. 2023. Scalable Stochastic Gradient Riemannian Langevin Dynamics in Non-Diagonal Metrics. *Transactions on Machine Learning Research*.

Zhang, R.; Li, C.; Zhang, J.; Chen, C.; and Wilson, A. G. 2020. Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning. In *International Conference on Learning Representations*.

A Experimental Details

Müller–Brown potential. The Müller–Brown potential (Müller and Brown 1979) is defined as

$$U(x,y) = \frac{1}{20}\sum_{i=1}^{4} C_i \exp\{l_i\},$$
$$l_i = a_i(x-u_i)^2 + b_i(x-u_i)(y-v_i) + c_i(y-v_i)^2,$$

where $C_i, a_i, b_i, c_i, u_i, v_i$, for $i = 1, \ldots, 4$, are the potential parameters. We used

$$\begin{aligned} C_i &= [-267.0, -285.0, -275.0, 2.5], \\ a_i &= [-0.9, -0.9, -9.5, 0.6], \\ b_i &= [0.0, 0.0, 10.0, 0.1], \\ c_i &= [-9.0, -9.0, -5.5, 0.1], \\ u_i &= [1.35, -0.95, -1.05, -1.0], \\ v_i &= [-0.5, -0.15, 1.05, 0.9]. \end{aligned}$$

The adaptive parameters for SA-SGLD were $\alpha = 0.07$, $r = 0.25$, and $s = 2$.

Star potential. The "star" potential (Leimkuhler, Lohmann, and Whalley 2025) is defined as

$$U(x, y) = x^2 + 1000x^2y^2 + y^2,$$

with adaptive parameters $\alpha = 0.5$, $r = 0.5$, and $s = 2$.
Each simulation used 10^6 samples with a fixed base step size for SGLD and a dynamically evolving step size for SA-SGLD. The corresponding trace plots visualize how the adaptive mechanism responds to local curvature across modes.

Bayesian Neural Network. All experiments were run on a NVIDIA V100 40GB GPU. Following the setup from (Yu et al. 2023) and (Li et al. 2015), we use the MNIST dataset a 3 layer fully connected neural network (FCNN) 784-N-N-10, where $N = 1200$. We do not use data augmentation to avoid the cold posterior effect to sample at temperature $T = 1$. All results reported as mean $\pm$ 95% confidence interval over 5 runs. We note that the ECE may be ambiguous to interpret as it is not a proper scoring rule as a random classifier could have an ECE of 0. NLL on the other hand is a proper scoring rule.

B Theoretical Results

Our analysis builds upon foundational work in stochastic approximation and Markov chain theory. The moment bounds in Theorem 1 employ Lyapunov drift techniques inspired by (Raginsky, Rakhlin, and Telgarsky 2017)'s non-asymptotic analysis of non-convex learning, while the ergodicity framework follows the Markov chain stability theory of (Meyn, Tweedie, and Glynn 2009). For the bias analysis in Theorem 2, we adapt the Poisson equation approach developed by (Mattingly, Stuart, and Tretyakov 2010) for numerical time-averaging, combined with the non-asymptotic analysis of stochastic gradient Langevin dynamics by (Vollmer, Zygalakis, and Teh 2016). The local error analysis draws from (Sato and Nakagawa 2014)'s Fokker-Planck approximation framework, and the handling of adaptive step-sizes utilizes the stochastic approximation theory of (Kushner and Yin 2003).
We note that our proof does not yield optimal convergence rates and that inequalities can be tightened further. We used simplifying assumptions for theoretical convenience such as (θ_n, ζ_n) being ergodic Markov with unique invariant measure. This is reasonable to assume because $\zeta > 0$ and $m \leq \psi(.) \leq M$, however it must still be shown rigorously. We also assumed $\sup_{n\geq 0} \mathbb{E}\|\theta_n\|^p < \infty$ for arbitrary p while only proving it for $p = 2$ in Theorem 1. This must also be proven rigorously and we leave it to future work. The regularization parameter $\delta > 0$ ensures ζ_n is bounded away from zero, allowing ψ to be globally Lipschitz on that domain. All convergence constants depend on δ but remain finite for any fixed $\delta > 0$.

Lemma 1 (Uniform moment bounds)**.** *Assume the following:*

1. *The Sundman step is fixed:* $\Delta\tau_n \equiv h > 0$.
2. *The monitor uses* $g(\theta_n) = \|G_n\|^2 + \delta$*, so that*

$$\zeta_{n+1} = \rho\zeta_n + \frac{1-\rho}{\alpha}\|G_n\|^2, \quad \rho = e^{-\alpha h}.$$

3. *The adaptive time step is* $\Delta t_{n+1} = \psi(\zeta_{n+1})h$*, where* ψ *is bounded and globally Lipschitz:*

$$\begin{aligned} 0 < m \leq \psi(\zeta) \leq M < \infty, \\ |\psi(x) - \psi(y)| \leq L_\psi |x - y|. \end{aligned}$$

   for $\zeta \geq \frac{(1-\rho)\delta}{\alpha}$.
4. *The potential* U *is* L*-smooth and dissipative:*

$$\begin{aligned} \|\nabla U(\theta) - \nabla U(\theta')\| \leq L\|\theta - \theta'\|, \\ \langle \theta, \nabla U(\theta)\rangle \geq a\|\theta\|^2 - b, \end{aligned}$$

   for $L > 0$*,* $a > 0$*,* $b \geq 0$.
5. *The stochastic gradients satisfy, for some* $\sigma < \infty$*:*

$$\mathbb{E}[\|G_n - \nabla U(\theta_n)\|^2 | \mathcal{F}_n] \leq \sigma^2(1 + \|\theta_n\|^2),$$

   where $\mathcal{F}_n = \sigma(\theta_0, \zeta_0, \varepsilon_1, \ldots, \varepsilon_n, G_0, \ldots, G_{n-1})$.

Define the constants:

$$\begin{aligned} C_1 &:= 2L^2 + 2\sigma^2, \\ C_2 &:= 2\|\nabla U(0)\|^2 + 2\sigma^2, \\ C_3 &:= 2\beta^{-1}d. \end{aligned}$$

If $h > 0$ *is small enough that*

$$\gamma(h) := 2amh - C_1M^2h^2 - 2\sigma Mh > 0, \tag{10}$$

then the iterates of

$$\theta_{n+1} = \theta_n - \Delta t_{n+1}G_n + \sqrt{2\beta^{-1}\Delta t_{n+1}}\,\varepsilon_{n+1}, \tag{11}$$

$$\varepsilon_{n+1} \sim \mathcal{N}(0, I_d), \tag{12}$$

satisfy the uniform moment bound

$$\sup_{n\geq 0} \mathbb{E}\|\theta_n\|^2 < \infty.$$

Proof. Since $\Delta t_{n+1} = \psi(\zeta_{n+1})h$ depends on G_n, it is not $\mathcal{F}_n$–measurable. All conditional expectations keep Δt_{n+1} inside the expectation.
Expanding (11):

$$\begin{aligned} \|\theta_{n+1}\|^2 = \|\theta_n\|^2 - 2\Delta t_{n+1}\langle\theta_n, G_n\rangle \\ + \Delta t_{n+1}^2\|G_n\|^2 + 2\beta^{-1}\Delta t_{n+1}\|\varepsilon_{n+1}\|^2 + R_n, \end{aligned}$$

where R_n contains cross terms. Taking conditional expectations using $\mathbb{E}[\|\varepsilon_{n+1}\|^2|\mathcal{F}_n] = d$ and $\mathbb{E}[\varepsilon_{n+1}|\mathcal{F}_n] = 0$:

$$\mathbb{E}[\|\theta_{n+1}\|^2|\mathcal{F}_n] = \|\theta_n\|^2 - 2\,\mathbb{E}[\Delta t_{n+1}\langle\theta_n, G_n\rangle|\mathcal{F}_n] + \mathbb{E}[\Delta t_{n+1}^2\|G_n\|^2|\mathcal{F}_n] + 2\beta^{-1}d\,\mathbb{E}[\Delta t_{n+1}|\mathcal{F}_n]. \tag{13}$$

Decompose:

$$\begin{aligned}&\mathbb{E}[\Delta t_{n+1}\langle\theta_n, G_n\rangle|\mathcal{F}_n]\\ &= \mathbb{E}[\Delta t_{n+1}\langle\theta_n, \nabla U(\theta_n)\rangle|\mathcal{F}_n]\\ &\quad + \mathbb{E}[\Delta t_{n+1}\langle\theta_n, G_n - \nabla U(\theta_n)\rangle|\mathcal{F}_n].\end{aligned}$$

By dissipativity and $mh \le \Delta t_{n+1} \le Mh$:

$$-2\,\mathbb{E}[\Delta t_{n+1}\langle\theta_n, \nabla U(\theta_n)\rangle|\mathcal{F}_n] \le -2amh\,\|\theta_n\|^2 + 2bMh.$$

For the stochastic term, Cauchy–Schwarz gives:

$$\begin{aligned}&\left|\mathbb{E}[\Delta t_{n+1}\langle\theta_n, G_n - \nabla U(\theta_n)\rangle|\mathcal{F}_n]\right|\\ &\le Mh\,\|\theta_n\|\,\mathbb{E}[\|G_n - \nabla U(\theta_n)\||\mathcal{F}_n]\\ &\le Mh\,\|\theta_n\|\sqrt{\mathbb{E}[\|G_n - \nabla U(\theta_n)\|^2|\mathcal{F}_n]}\\ &\le Mh\sigma\,\|\theta_n\|\sqrt{1+\|\theta_n\|^2}.\end{aligned}$$

Using Young's inequality $xy \le \frac{x^2}{2} + \frac{y^2}{2}$ with $x = \|\theta_n\|$:

$$\|\theta_n\|\sqrt{1+\|\theta_n\|^2} \le \|\theta_n\|^2 + \frac{1}{2}.$$

Thus:

$$\begin{aligned}&-2\,\mathbb{E}[\Delta t_{n+1}\langle\theta_n, G_n\rangle|\mathcal{F}_n]\\ &\le -(2amh - 2\sigma Mh)\|\theta_n\|^2 + (2bMh + \sigma Mh).\end{aligned} \tag{14}$$

From L-smoothness and the noise assumption:

$$\begin{aligned}\mathbb{E}[\|G_n\|^2|\mathcal{F}_n] &\le 2\mathbb{E}[\|\nabla U(\theta_n)\|^2|\mathcal{F}_n]\\ &\quad + 2\mathbb{E}[\|G_n - \nabla U(\theta_n)\|^2|\mathcal{F}_n]\\ &\le 2L^2\|\theta_n\|^2 + 2\|\nabla U(0)\|^2 + 2\sigma^2(1+\|\theta_n\|^2)\\ &= C_1\|\theta_n\|^2 + C_2.\end{aligned}$$

Since $\Delta t_{n+1}^2 \le M^2h^2$:

$$\begin{aligned}&\mathbb{E}[\Delta t_{n+1}^2\|G_n\|^2|\mathcal{F}_n] \le M^2h^2(C_1\|\theta_n\|^2 + C_2),\\ &2\beta^{-1}d\,\mathbb{E}[\Delta t_{n+1}|\mathcal{F}_n] \le C_3Mh.\end{aligned}$$

Substituting into (13):

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}\|^2|\mathcal{F}_n] &\le \|\theta_n\|^2 - (2amh - 2\sigma Mh)\|\theta_n\|^2 + C_3Mh\\ &\quad + (2bMh + \sigma Mh) + M^2h^2(C_1\|\theta_n\|^2 + C_2)\\ &= (1 - 2amh + C_1M^2h^2 + 2\sigma Mh)\|\theta_n\|^2\\ &\quad + (C_2M^2h^2 + (2b + C_3 + \sigma)Mh).\end{aligned}$$

Let $\kappa(h) := C_2M^2h^2 + (2b + C_3 + \sigma)Mh$. If $h > 0$ is small enough that $\gamma(h) > 0$, taking expectations:

$$x_{n+1} \le (1 - \gamma(h))x_n + \kappa(h), \quad x_n := \mathbb{E}\|\theta_n\|^2.$$

Since $0 < 1 - \gamma(h) < 1$:

$$\sup_{n\ge 0} x_n \le \max\{x_0, \kappa(h)/\gamma(h)\} < \infty.$$

□

Theorem 1 (Ergodicity and $O(h)$ bias)**.** *Retain Lemma 1 assumptions. Further assume:*

1. $\mathbb{E}[G_n|\mathcal{F}_n] = \nabla U(\theta_n)$.
2. *For $p > 0$ large:* $\sup_{n\ge 0}\mathbb{E}\|\theta_n\|^p < \infty$,
3. $\mathbb{E}[\|G_n - \nabla U(\theta_n)\|^4|\mathcal{F}_n] \le \sigma_4^2(1 + \|\theta_n\|^4)$.
4. *(θ_n, ζ_n) is ergodic Markov with invariant measure $\widetilde{\pi}_h$.*
5. *$U \in C^4(\mathbb{R}^d)$ with bounded derivatives, ∇U Lipschitz.*

Let $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathcal{L}\phi = f - \pi(f), \quad \mathcal{L} = -\nabla U \cdot \nabla + \beta^{-1}\Delta,$$

admits $\phi \in C^4(\mathbb{R}^d)$ with polynomial-growth derivatives:

$$\sup_\theta \frac{\|D^j\phi(\theta)\|}{1 + \|\theta\|^q} \le A_j, \quad j = 0, 1, 2, 3, 4.$$

Define the weighted time-average:

$$\mathcal{A}_n := \frac{\sum_{k=1}^n \Delta t_k f(\theta_k)}{\sum_{k=1}^n \Delta t_k}, \quad \Delta t_k = \psi(\zeta_k)h.$$

Let $\pi_h(f) := \frac{\widetilde{\pi}_h[\psi(\zeta)f(\theta)]}{\widetilde{\pi}_h[\psi(\zeta)]}$ be the ψ-weighted marginal. Then, For every f:

$$\mathcal{A}_n \xrightarrow{n\to\infty} \text{a.s.}\pi_h(f).$$

There exists $C > 0$ such that

$$|\pi_h(f) - \pi(f)| \le Ch.$$

Moreover, for all $n \ge 1$:

$$\left|\mathbb{E}\mathcal{A}_n - \pi(f)\right| \le |\pi_h(f) - \pi(f)| + \frac{C'}{n} = O(h) + O(n^{-1}).$$

Proof. By strengthened moments, $\sup_n \mathbb{E}\|\theta_n\|^p < \infty$. Since (θ_n, ζ_n) is ergodic Markov with invariant measure $\widetilde{\pi}_h$ and f has polynomial growth, $(\theta, \zeta) \mapsto \psi(\zeta)f(\theta)$ is $\widetilde{\pi}_h$-integrable.
By the ergodic theorem for weighted averages and $T_n = \sum_{k=1}^n \Delta t_k \ge mhn \to \infty$ a.s.:

$$\mathcal{A}_n = \frac{\sum_{k=1}^n \Delta t_k f(\theta_k)}{\sum_{k=1}^n \Delta t_k} \xrightarrow{n\to\infty} \text{a.s.}\pi_h(f).$$

Let ϕ solve $\mathcal{L}\phi = f - \pi(f)$. Fourth-order Taylor:

$$\begin{aligned}\phi(\theta_{n+1}) =&\phi(\theta_n) + D\phi(\theta_n)\cdot(\theta_{n+1} - \theta_n)\\ &+ \tfrac{1}{2}(\theta_{n+1} - \theta_n)^\top D^2\phi(\theta_n)(\theta_{n+1} - \theta_n)\\ &+ \tfrac{1}{6}D^3\phi(\theta_n)[\theta_{n+1} - \theta_n]^{\otimes 3}\\ &+ \tfrac{1}{24}D^4\phi(\tilde{\theta}_n)[\theta_{n+1} - \theta_n]^{\otimes 4}.\end{aligned}$$

Substituting $\theta_{n+1} - \theta_n = -\Delta t_{n+1}G_n + \sqrt{2\beta^{-1}\Delta t_{n+1}}\,\varepsilon_{n+1}$ and taking $\mathbb{E}[\cdot|\mathcal{F}_n]$:

$$\begin{aligned}\mathbb{E}[\phi(\theta_{n+1})|\mathcal{F}_n] =&\phi(\theta_n) - \mathbb{E}[\Delta t_{n+1}D\phi(\theta_n)\cdot G_n|\mathcal{F}_n]\\ &+ \beta^{-1}\mathbb{E}[\Delta t_{n+1}\Delta\phi(\theta_n)|\mathcal{F}_n]\\ &+ \tfrac{1}{2}\mathbb{E}[\Delta t_{n+1}^2(G_n^\top D^2\phi(\theta_n)G_n)|\mathcal{F}_n]\\ &+ \mathbb{E}[R_n|\mathcal{F}_n],\end{aligned}$$

where terms with odd powers of ε_{n+1} vanish.

By properties of Gaussian noise, the terms $\mathbb{E}[\Delta t_{n+1}^{3/2}\langle D^2\phi(\theta_n)G_n, \varepsilon_{n+1}\rangle|\mathcal{F}_n]$ and related cross terms vanish.
Decompose:

$$\mathbb{E}[\Delta t_{n+1} D\phi(\theta_n)\cdot G_n|\mathcal{F}_n] = \mathbb{E}[\Delta t_{n+1}|\mathcal{F}_n]D\phi(\theta_n)\cdot\nabla U(\theta_n)+B_n,$$

where the bias term B_n arises from correlation between Δt_{n+1} and $G_n - \nabla U(\theta_n)$. By Lipschitz property of ψ and moment bounds:

$$|B_n| \leq C_2 h^2(1 + \|\theta_n\|^{q+3}).$$

For third and fourth order terms in the Taylor expansion:

$$|\mathbb{E}[R_n|\mathcal{F}_n]| \leq C_3 h^2(1 + \|\theta_n\|^{q+4}),$$

using $\|\theta_{n+1} - \theta_n\| \leq Mh\|G_n\| + \sqrt{2\beta^{-1}Mh}\|\varepsilon_{n+1}\|$ and polynomial growth of derivatives.
Rearranging the expansion, in stationarity:

$$\mathbb{E}_{\widetilde{\pi}_h}[\mathbb{E}[\Delta t_{n+1}|\mathcal{F}_n](f(\theta_n)-\pi(f))] = \mathbb{E}_{\widetilde{\pi}_h}[B_n]+O(h^2\mathbb{E}[\|\theta_n\|^{q+4}]).$$

Since

$$\mathbb{E}[\Delta t_{n+1}|\mathcal{F}_n](f - \pi(f)) = h\psi(\zeta_n)(f - \pi(f)) + O(h^2),$$

we obtain:

$$h \cdot \mathbb{E}_{\widetilde{\pi}_h}[\psi(\zeta_n)(f - \pi(f))] = O(h^2),$$

thus

$$|\pi_h(f) - \pi(f)| = O(h).$$

Summing from $k = 0$ to $n - 1$:

$$\sum_{k=0}^{n-1}\mathbb{E}[\Delta t_{k+1}|\mathcal{F}_k](f(\theta_k)-\pi(f)) = \phi(\theta_0)-\phi(\theta_n)+\sum_{k=0}^{n-1}E_k,$$

where $|\mathbb{E}[E_k]| \leq C_5 h^2$. Dividing by $T_n \geq mhn$:

$$|\mathbb{E}\mathcal{A}_n - \pi(f)| \leq Ch + \frac{C'}{n}.$$

□