ROBUST STOCHASTIC GRADIENT POSTERIOR SAMPLING WITH LATTICE BASED DISCRETISATION

Zier Mensch
University of Amsterdam,
National Taiwan University

Lars Holdijk
University of Oxford,
Normal Computing

Samuel Duffield
Normal Computing

Maxwell Aifer
Normal Computing

Patrick J. Coles
Normal Computing

Max Welling
University of Amsterdam,
CuspAI

Miranda Cheng
University of Amsterdam
Academia Sinica Taiwan

ABSTRACT

Stochastic-gradient MCMC methods enable scalable Bayesian posterior sampling but often suffer from sensitivity to minibatch size and gradient noise. To address this, we propose Stochastic Gradient Lattice Random Walk (SGLRW), an extension of the Lattice Random Walk discretization. Unlike conventional Stochastic Gradient Langevin Dynamics (SGLD), SGLRW introduces stochastic noise only through the off-diagonal elements of the update covariance; this yields greater robustness to minibatch size while retaining asymptotic correctness. Furthermore, as comparison we analyze a natural analogue of SGLD utilizing gradient clipping. Experimental validation on Bayesian regression and classification demonstrates that SGLRW remains stable in regimes where SGLD fails, including in the presence of heavy-tailed gradient noise, and matches or improves predictive performance.

1 INTRODUCTION

Bayesian methods provide a principled framework for learning probabilistic models from data and natively capturing uncertainty by replacing the parameter point estimates in frequentist methods with a posterior distribution over parameters. By marginalizing over parameters, Bayesian methods act as a form of regularization and enable uncertainty quantification and robust model selection (Neal, 2012). In doing so, Bayesian models can potentially mitigate overfitting and miscalibration, which are prevalent in modern large-scale, overparameterized neural networks (Guo et al., 2017; Yang et al., 2023). Realizing these benefits in the modern hyperscaling era, however, requires posterior inference algorithms that scale to both dataset size and model complexity.

Within Bayesian methods, Markov chain Monte Carlo (MCMC) (Neal, 1993; Robert et al., 1999) remains the standard for posterior sampling, but it is also among the methods most affected by scalability and computational cost (Gelman et al., 1997). Alternatives, including variational inference (Blei et al., 2017), Laplace approximations (Tierney & Kadane, 1986), and single-pass methods (Gal & Ghahramani, 2016), are often less computationally demanding, but still introduce overhead in training and inference (Blei et al., 2017; Lakshminarayanan et al., 2017; Wilson & Izmailov, 2020). As a result, these methods have, in some settings, fallen out of favour relative to modern approaches for assessing model trustworthiness, such as explainable and interpretable models (Li et al., 2023).

One core issue of MCMC methods for Bayesian posterior inference is the theoretical requirement to evaluate the gradient of the posterior over the entire dataset at each iteration (Welling & Teh, 2011; Ma et al., 2015). With growing model complexity and dataset size, this is often prohibitively expensive. Stochastic-gradient variants of MCMC methods, such as Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011), alleviate this concern to some extent by allowing the gradient to be evaluated on a small minibatch of data at each iteration. However, these methods are still known to be sensitive to the minibatch size (Brosse et al., 2018). As a result, they do not scale to regimes where only a small minibatch is available at each evaluation step, or where only a small number of samples from the dataset can be stored in memory, as is becoming increasingly common.

In this work, we propose Stochastic Gradient Lattice Random Walk (SGLRW), a stochastic-gradient extension of the Lattice Random Walk (LRW) (Duffield et al., 2025) discretisation of overdamped Langevin dynamics. LRW replaces the Gaussian increments with bounded binary updates on a lattice. Unlike SGLD, the stochastic gradient noise in SGLRW enters only through the off-diagonal elements of the covariance matrix of the update and therefore remains robust to the minibatch size. This allows SGLRW to sample from the posterior distribution with the same asymptotic correctness as SGLD, but with improved stability for small minibatches, as shown in Figure 1.

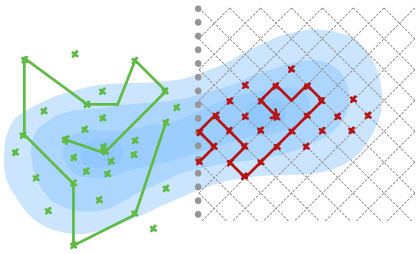

Figure 1: Comparing SGLD **(left)** and SGLRW **(right)** discretisations of Langevin dynamics, we can observe that the lattice based discretisation suppresses large parameter jumps that occur due to minibatch noise.

2 Background

As stated in the introduction, we consider the problem of minibatch-induced instability in stochastic gradient MCMC methods for Bayesian posterior sampling. Here, we recap the necessary background on Bayesian machine learning, posterior inference, and stochastic gradient methods.

Bayesian Machine Learning We consider the supervised learning setting where we have observed data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ and aim to infer a posterior distribution $p(\theta \mid \mathcal{D})$ over the parameter vector $\theta \in \mathbb{R}^d$. In contrast to frequentist approaches, which seek a single point estimate θ^* that maximises the likelihood $p(\mathcal{D} \mid \theta)$, Bayesian machine learning maintains a distribution over parameters, $p(\theta \mid \mathcal{D})$. This *posterior distribution* captures the uncertainty in our parameter estimates given the data.

The posterior distribution is given by Bayes' theorem as

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \propto p(\theta) \prod_{i=1}^{N} p(y_i \mid x_i, \theta), \tag{1}$$

where $p(\mathcal{D} \mid \theta)$ is the *likelihood* and $p(\theta)$ is the *prior*. Notably, we will often write the posterior distribution as $p(\theta \mid \mathcal{D}) \propto \exp[-U(\theta)]$ where the following is referred to as the negative log-posterior.

$$U(\theta) = -\log p(\theta) - \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta), \tag{2}$$

Predictions for a new input x^* are then obtained via the posterior predictive distribution

$$p(y^* \mid x^*, \mathcal{D}) = \int p(y^* \mid x^*, \theta)\, p(\theta \mid \mathcal{D})\, d\theta, \tag{3}$$

which marginalises over parameters θ and thus provides calibrated predictive distributions. Crucially, however, evaluating eq. 3 is generally infeasible, as it involves a high-dimensional integral over θ.

2.1 Bayesian Posterior Sampling

A common approach to Bayesian inference is to replace the integral in equation 3 with a Monte Carlo average. Most commonly, this is achieved by drawing parameter samples $\theta \sim p(\theta \mid \mathcal{D})$ from the posterior using a Markov chain Monte Carlo (MCMC) approach.

In this work, we focus on MCMC samplers that are expressed as discretisations of stochastic differential equations (SDEs) whose stationary distribution coincides with the target posterior (Ma et al., 2015). Among these, the most common choice is the overdamped Langevin diffusion,

$$d\theta_t = f(\theta_t)\, dt + \sqrt{2D(\theta_t)}\, dW_t, \tag{4}$$

where $D(\theta_t)$ is a symmetric positive semidefinite diffusion matrix and $f(\theta_t)$ is the drift. In the context of Bayesian posterior sampling, the drift is given by $f(\theta_t) = -\nabla U(\theta_t)$ and the diffusion matrix is typically set to $D(\theta_t) = I$, resulting in the following SDE:

$$d\theta_t = -\nabla U(\theta_t)\, dt + \sqrt{2}\, dW_t = (\nabla \log p(\theta_t) + \sum_{i=1}^{N} \nabla \log p(y_i \mid x_i, \theta_t))dt + \sqrt{2}\, dW_t. \tag{5}$$

2.1.1 STOCHASTIC GRADIENT METHODS.

As discussed in the introduction, in large-scale settings the requirement to evaluate the gradient of the posterior over the entire dataset at each iteration is limiting. As such, stochastic gradient MCMC (SG-MCMC) methods replace the full posterior gradient with an unbiased minibatch estimator:

$$\widehat{\nabla U}(\theta; \mathcal{B}) = -\nabla \log p(\theta) - \frac{N}{B} \sum_{i \in \mathcal{B}} \nabla \log p(y_i \mid x_i, \theta), \tag{6}$$

where $\mathcal{B} \subset \{1, \ldots, N\}$ is a minibatch index set of size B, and $\{(x_i, y_i)\}_{i \in \mathcal{B}} \subset \mathcal{D}$ are the datapoints.

Stochastic Gradient Langevin Dynamics Stochastic Gradient methods for Bayesian posterior sampling were first introduced by Welling & Teh (2011) for Langevin dynamics, and only later generalised to other MCMC samplers (Chen et al., 2014; Ma et al., 2015). Concretely, Welling & Teh (2011) obtain the following Stochastic Gradient Langevin Dynamics (SGLD) update rule:

Definition 2.1 (SGLD Update Rule) *Given a minibatch $\mathcal{B}$, applying the Euler-Maruyama discretisation to the Langevin SDE Equation (5) and replacing the full gradient with a minibatch estimate yields the SGLD update:*

$$\theta_{t+1} = \theta_t - \delta_t \, \widehat{\nabla U}(\theta_t; \mathcal{B}) + \sqrt{2\delta_t}\, \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I) \tag{7}$$

For the original full-gradient Langevin dynamics, Euler-Maruyama has weak order 1 convergence. However, with stochastic gradients, the convergence properties are more subtle (Vollmer et al., 2016). With a fixed step size $\delta_t = \delta$, SGLD converges to a stationary distribution that is biased relative to the true posterior, reflecting both discretisation error and the noise from minibatch gradients. To obtain asymptotically exact expectations, a decreasing step size schedule satisfying $\delta_t \to 0$, $\sum_t \delta_t = \infty$, and $\sum_t \delta_t^2 < \infty$ needs to be used (Teh et al., 2016).

In our analysis, we decompose the gradient as $\widehat{\nabla U}(\theta; \mathcal{B}) = \nabla U(\theta) + \zeta(\theta; \mathcal{B})$ and define $G(\theta)$ to be the minibatch-induced covariance, $\mathrm{Cov}_{\mathcal{B}}[\zeta(\theta; \mathcal{B}) \mid \theta] = G(\theta)$. Clearly we have $\mathbb{E}_{\mathcal{B}}[\zeta(\theta; \mathcal{B}) \mid \theta] = 0$.

3 RELATED WORK

Batch-Size Sensitivity. In practice, SGLD requires large minibatches for stability, limiting scalability (Baker et al., 2019). Variance-reduction and control-variate methods reduce noise but introduce memory overhead or require periodic full-data passes (Dubey et al., 2016; Baker et al., 2019; Li et al., 2020). Alternative approaches include adaptive subsampling (Korattikara et al., 2014), preconditioning (Li et al., 2016), and importance sampling (Li et al., 2020). Moreover, recent work characterizes stochastic gradient noise as heavy-tailed rather than Gaussian (Simsekli et al., 2019) and proposes fractional dynamics to retarget the distribution when such noise is present (Simsekli et al., 2020). While these methods primarily address the statistical properties of the noise, our solution improves robustness through its lattice-based discretisation.

Large-Scale Bayesian Inference. Bayesian uncertainty estimation is important for large neural models, including large language models, where calibration and robustness are critical (Yang et al., 2023). Scalable approximations such as Laplace methods (Daxberger et al., 2021; Yang et al., 2023; Chen & Garner, 2024; Sliwa et al.) and variational inference (Harrison et al., 2024; Wang et al., 2024; Xiang et al., 2025; Samplawski et al., 2025) trade accuracy for efficiency. Recent work shows that sampling-based methods can be applied to large models with appropriate algorithmic structure, as in SGLD-Gibbs (Kim & Hospedales, 2025).

4 STOCHASTIC GRADIENT LATTICE RANDOM WALK

With the background established, we now introduce the proposed method, Stochastic Gradient Lattice Random Walk (SGLRW). For this purpose, we first review the Lattice Random Walk (LRW) discretisation of Langevin dynamics and then introduce the proposed stochastic gradient extension.

4.1 LATTICE RANDOM WALK

The Lattice Random Walk (LRW) scheme, introduced in (Duffield et al., 2025), proposes an alternative to standard SDE discretisations, such as Euler-Maruyama, by substituting Gaussian noise with bounded binary increments, as shown in (Duffield et al., 2025). this has the benefit of being stable

under non-Lipschitz gradients, which are common in deep learning. The structure of LRW also lends itself to low-precision, stochastic hardware (Alaghi & Hayes, 2013) and thermodynamic hardware (Conte et al., 2019) which has recently gained interest for AI applications (Melanson et al., 2025).

In LRW, at each iteration, the parameters are updated as

$$\Delta\theta_{t+1} = S_t, \qquad (S_t)_i \in \{-\sqrt{2\delta_t}, +\sqrt{2\delta_t}\}, \tag{8}$$

where each direction $(S_t)_i$ is sampled independently from the following state-dependent probability

$$\mathbb{P}\Big[(S_t)_i = \pm\sqrt{2\delta_t}\,\Big|\,\theta_t\Big] = \tfrac{1}{2} \pm \tfrac{1}{2}\sqrt{\tfrac{\delta_t}{2}}\, f_i(\theta_t). \tag{9}$$

These probabilities are valid whenever $\sqrt{\delta_t/2}\,|f_i(\theta_t)| \leq 1$, where we write $f_i(\theta)$ to denote the ith component of the drift vector field $f(\theta)$.

By construction, the first two conditional moments satisfy

$$\mathbb{E}[S_t \mid \theta_t] = \delta_t f(\theta_t), \qquad \mathbb{E}[S_t S_t^\top \mid \theta_t] = 2\delta_t I, \tag{10}$$

and LRW is shown to be weakly first-order consistent with the continuous-time Langevin dynamics in Equation (4) (Theorem 1 of (Duffield et al., 2025)). With the specific choice of $f(\theta) = -\nabla U(\theta)$, LRW thus provides a valid discretisation of the Langevin dynamics in Equation (5).

4.2 Stochastic Gradient Lattice Random Walk

We now come to the main contribution, the proposal of Stochastic Gradient Lattice Random Walk (SGLRW), which replaces the stochastic gradient update rule of SGLD with a lattice-based update:

Definition 4.1 (SGLRW Update Rule) *Given a minibatch $\mathcal{B}$, at the tth iteration, the Stochastic Gradient Lattice Random Walk updates the parameter vector as*

$$\theta_{t+1} = \theta_t + S_t. \tag{11}$$

where each coordinate $(S_t)_i \in \{-\sqrt{2\delta_t}, +\sqrt{2\delta_t}\}$ is sampled from the state-dependent probability

$$\mathbb{P}\Big[(S_t)_i = \pm\sqrt{2\delta_t}\,\Big|\,\theta_t\Big] = \tfrac{1}{2} \mp \tfrac{1}{2}\sqrt{\tfrac{\delta_t}{2}}\, \widehat{\partial_i U}(\theta_t;\mathcal{B}), \tag{12}$$

which is valid whenever $\sqrt{\delta_t/2}\,|\widehat{\partial_i U}(\theta_t;\mathcal{B})| \leq 1$.

We hypothesize, and analytically evaluate next, that due to the bounded structure, large fluctuations in stochastic gradients have a less severe impact on the update in the case of SGLRW than in the case of SGLD. This is illustrated in Figure 2, in a one-dimensional multimodal example.

Heavy-Tailed Noise Using the gradient as $\widehat{\nabla U}(\theta;\mathcal{B}) = \nabla U(\theta) + \zeta(\theta;\mathcal{B})$, we set $U(\theta)$ to be the negative log-probability of the multimodal Gaussian, and choose $\zeta(\theta;\mathcal{B})$ to follow a heavy-tailed α-stable distribution with $\alpha < 2$, for which second moments do not exist and was shown to closely resemble the minibatch gradient noise in the standard SGD setting by Simsekli et al. (2019).

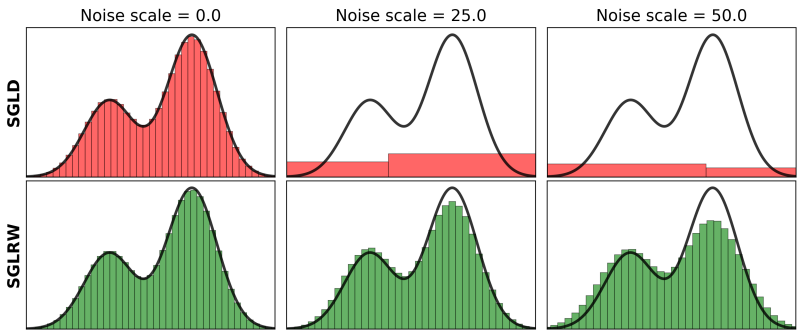

Figure 2: Multimodal univariate target with exact gradient corrupted by synthetic α-stable noise ($\alpha = 1.5$) of increasing scale. We observe that as the noise scale increases SGLD quickly fails while SGLRW remains stable.

As Figure 2 shows, in this regime, where stability depends critically on whether large stochastic fluctuations can induce rare but catastrophic updates, SGLD fails while SGLRW remains stable. In Appendix A.1 we also provide an analysis using Gaussian gradient noise (Figure 8).

4.3 Mean Squared Error Analysis

Having introduced SGLRW, we now present an analysis of the differences between SGLD and SGLRW that highlight the benefits of using SGLRW with small batch sizes. We follow a similar approach to the analysis of SGLD in Chen et al. (2015), focussing on the mean squared error (MSE) $\mathbb{E}(\hat{\phi} - \bar{\phi})^2$ between the true posterior expectation

$$\bar{\phi} := \int \phi(\theta)\, p(\theta \mid \mathcal{D})\, d\theta. \tag{13}$$

and the ergodic average

$$\hat{\phi} := \frac{1}{L} \sum_{n=1}^{L} \phi(\theta_{n\delta_t}), \tag{14}$$

over the discrete-time Markov chain $\{\theta_{n\delta_t}\}_{n\geq 0}$ generated by an SG-MCMC method, such as SGLD or SGLRW, with step size δ_t. Here, $\phi : \mathbb{R}^d \to \mathbb{R}^m$ represents a smooth test function, such as the posterior predictive distribution of a new data point x^*, as defined in Equation (3).

As a comparison of the MSE for SGLRW and SGLD, we present the following theorem:

Theorem 4.2 *Under Assumption A.1, we find that the MSE is bounded by three contributions*

$$\textit{MSE} \leq C\,(\mathcal{E}_{\textit{drift}} + \mathcal{E}_{\textit{disc}} + \mathcal{E}_{\textit{cov}})$$

for some C that depends on the target distribution. The covariance error term for SGLRW is never larger than that of SGLD

$$\mathcal{E}_{\textit{cov}}^{\textit{SGLRW}} \leq \mathcal{E}_{\textit{cov}}^{\textit{SGLD}}$$

while the other contributions are the same for both. Moreover, it is strictly smaller whenever $2\partial_i U \zeta_i + \zeta_i^2$ is non-vanishing for some direction i.

The first statement regarding the MSE upper bound, and the precise expression for each of the contributions, is the content of Theorem A.3, which is an extension of Theorem 3 of Chen et al. (2015) to take into account non-vanishing second-order contributions, as given by

$$\mathcal{E}_{\text{cov}} = \frac{\delta_t^2}{L} \sum_{n=1}^{L} \mathbb{E}\big[\|M_n\|_F^2\big]. \tag{15}$$

In the above, the scheme-dependent second-order error M_n induced by minibatching is defined as

$$M_n(\theta; \mathcal{B}_n) := \delta_t^{-2}\, \mathbb{E}_{\varepsilon_n}\Big[\Delta\theta_n^{\text{fb}}(\Delta\theta_n^{\text{fb}})^\top - \Delta\theta_n^{\text{mb}}(\Delta\theta_n^{\text{mb}})^\top \mid \theta_{(n-1)\delta_t} = \theta, \mathcal{B}_n\Big], \tag{16}$$

where $\Delta\theta_n^{\text{mb}}$ and $\Delta\theta_n^{\text{fb}}$ denote the one-step increments of the minibatch and full-batch updates.

To prove the bound in Theorem 4.2, we observe the lemma below which quantifies the difference between the second-order structure of SGLD and SGLRW (See Appendix A.3 for the proof).

Lemma 4.3 *The second moment error of the minibatch update for SGLRW satisfies*

$$M_{n,\text{SGLRW}}(\theta; \mathcal{B}_n) = \text{offdiag}\big(M_{n,\text{SGLD}}(\theta; \mathcal{B}_n)\big), \tag{17}$$

where $M_{n,\text{SGLD}}(\theta, \mathcal{B}_n)$ is the second-order error for SGLD given by

$$M_{n,\text{SGLD}} = \zeta\zeta^\top + \nabla U \zeta^\top + \zeta \nabla U^\top.$$

The lemma highlights that, in SGLRW, the lattice enforces fixed-magnitude coordinate updates, so the diagonal of the one-step second moment of the increment is deterministic. In contrast, for SGLD this diagonal depends on the stochastic gradient and is inflated by minibatch noise.

Combining the above, we readily obtain the error bound of Theorem 4.2. This shows that under some mild conditions, the SGLRW discretisation achieves a strictly tighter MSE bound than the SGLD, leading to a more robust implementation of minibatch gradient updates.

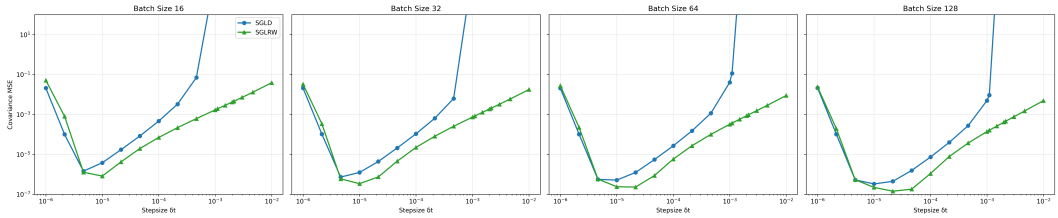

Figure 3: Mean-squared error (MSE) of the posterior covariance as a function of the step size δ_t, shown for different batch sizes for 50-dimensional Bayesian linear regression.

4.3.1 Validation and practical considerations

We experimentally validate this theoretical finding in Figure 3, where we compare the covariance MSE for SGLRW and SGLD for Bayesian linear and logistic regression. As δ_t increases, SGLD becomes unstable and the covariance error explodes, while SGLRW remains stable for all batch sizes. A further discussion of the experimental setup used to generate this insight is provided in Section 5.2.

Despite the theoretical guarantee in Lemma 4.3, we note that the restriction $\sqrt{\delta_t/2}\,|\widehat{\partial_i U}(\theta_t;\mathcal{B})| \leq 1$ can be challenging to arrange in practice without overly conservative step size tuning. To address this, in the implementation we clip the quantity $\sqrt{\delta_t/2}\,|\widehat{\partial_i U}(\theta_t;\mathcal{B})|$ to one. Although this clipping can lead to a bias, our empirical evaluations suggest that this does not pose any serious problem in practice, in the regimes of δ_t and minibatch sizes that one is interested in.

5 Experimental evaluation

We evaluate SGLRW on linear regression and a predictive classification posterior sampling. In App. A.5 we present a logistic regression task. Across all experiments, we vary the minibatch size B and base step size δ_t under decaying learning-rate schedules of the form $\delta_t(1+t)^{-0.55}$ (Welling & Teh, 2011). All experiments were implemented in `posteriors` (Duffield et al., 2024).

5.1 Strong Baseline: Clipped SGLD.

Similar to our analysis in Section 4.3, we compare SGLRW against SGLD in our empirical evaluation here. Additionally, we introduce *Clipped-SGLD* as an additional strong baseline. Gradient clipping is standard in large-scale SGD, where saturating the drift prevents rare large gradients from producing unstable updates. It is therefore natural to ask whether the same stabilisation can be applied to SGLD.

We define the Clipped-SGLD update rule as follows:

$$\theta_{t+1} = \theta_t - \mathrm{clip}(\delta_t\,\widehat{\nabla U}(\theta_t;\mathcal{B});R) + \sqrt{2\delta_t}\,\xi_t, \tag{18}$$

where $\mathrm{clip}(x;R)_i = \mathrm{sign}(x_i)\min\{|x_i|,R\}$ and $R = \sqrt{2\delta_t}$. This is a componentwise clipping operation, and deviates slightly from the standard definition of gradient clipping in SGD. In gradient clipping for standard SGD, the clipping is performed over the entire update vector $\Delta\theta_t$, while here we clip part of the update vector. Clipping the full update would result in the SDE having a different stationary distribution (App. A.4), while drift-truncated Euler schemes converge to exact Langevin dynamics in the small-step limit (Roberts & Tweedie, 1996; Hutzenthaler & Jentzen, 2015).

In Appendix A.1 we provide the same MSE and heavy-tailed noise analysis for Clipped-SGLD as previously discussed in Section 4.3 for SGLRW.

5.2 Bayesian Linear Regression

We first evaluate SGLRW using a linear–Gaussian model where the posterior admits a closed form. Using the closed-form solution, we can analytically compute the KL divergence between the true posterior and the empirical Gaussian fit to the samples. This allows us to provide a more rigorous evaluation of the empirical performance of SGLRW compared to SGLD and Clipped-SGLD.

Concretely, the linear model we consider is given by

$$y = X\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2 I), \tag{19}$$

where $X \in \mathbb{R}^{N\times d}$ is the design matrix and $\varepsilon \in \mathbb{R}^N$ is the noise vector. With a Gaussian prior $p(\theta) = \mathcal{N}(0, \tau^{-1}I)$, the resulting posterior $\mathcal{N}(\mu, \Sigma)$ is therefore given by

$$\Sigma^{-1} = \tfrac{1}{\sigma^2}X^\top X + \tau I, \qquad \mu = \tfrac{1}{\sigma^2}\Sigma X^\top y. \tag{20}$$

Setup Synthetic data are generated with $N = 1000$ and $d = 20$, using $\theta^* \sim \mathcal{N}(0, I)$, $\sigma^2 = 1.5$, and $\tau = 10^{-2}$. Each method is run with 2,000 parallel particles for 10,000 iterations, using matched minibatch sizes and the same decaying learning-rate schedule. As stated, the performance is quantified using the analytic Kullback–Leibler divergence between the true posterior $\mathcal{N}(\mu, \Sigma)$ and the empirical Gaussian fit to the samples, computed from their estimated mean and covariance.

Table 1: KL-divergence between the posterior and the empirical Gaussian fit of the samples, shown for different samplers, minibatch size B, and base lr δ_t. The Monte Carlo reference KL is 0.055201. **Bold** indicates the lowest KL.

Hyperparameters		**KL Divergence**		
B	δ_t	**SGLD**	**SGLRW**	**Clipped SGLD**
8	10^{-3}	19.889	**6.060**	18.184
8	10^{-4}	0.483	**0.202**	0.777
16	10^{-3}	7.155	**2.317**	8.530
16	10^{-4}	0.175	**0.070**	0.204
32	10^{-3}	2.441	**0.729**	3.540
32	10^{-4}	0.087	**0.064**	0.091
64	10^{-3}	0.838	**0.165**	1.114
64	10^{-4}	0.063	**0.055**	0.067
128	10^{-3}	0.315	**0.074**	0.351
128	10^{-4}	**0.054**	**0.054**	0.061
256	10^{-3}	0.140	**0.065**	0.141
256	10^{-4}	**0.051**	0.056	0.062
512	10^{-3}	0.086	**0.065**	0.088
512	10^{-4}	**0.052**	0.054	0.062
1000	10^{-3}	**0.058**	**0.058**	0.060
1000	10^{-4}	**0.052**	0.054	0.061

5.2.1 Results

The KL curves reveal two consistent effects, portrayed in Table 1: (i) *step size sensitivity:* SGLRW remains stable and continues to decrease KL under larger learning rates δ_t where SGLD diverges. (ii) *batch efficiency:* for comparable KL at matched δ_t, SGLRW achieves the same accuracy with approximately half the minibatch size, indicating greater robustness to stochastic-gradient noise.

These trends are accompanied by empirical covariance behaviour across methods, illustrated in Figure 4. Here we observe that the error in the diagonal terms of the estimated covariance matrices is lower for SGLRW than SGLD and Clipped-SGLD, and in general less impacted by the batch size.

5.3 Sentiment Classification With LLMs

Having considered tasks with well-understood posterior distributions, we now turn to a more realistic problem where issues arise due to the model size impacting the possible size of the minibatch: language modelling using LLMs. Specifically, we evaluate SGLRW on a sentiment classification task using the IMDB dataset (Maas et al., 2011), following a setup similar to Harrison et al. (2024). The dataset consists of 50,000 strongly polarized movie reviews, split evenly into training and test sets. To study the effect of data scale, we additionally consider subsampled training sets of varying sizes.

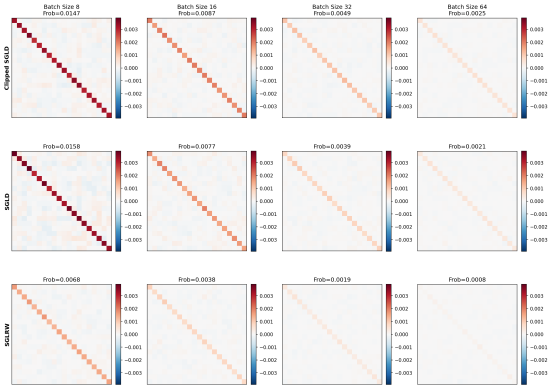

Figure 4: Covariance difference matrices $\Sigma_{\text{est}} - \Sigma_{\text{true}}$ for Bayesian linear regression at stepsize $\delta_t = 10^{-3}$, shown across increasing minibatch sizes B. **Top:** Clipped SGLD. **Middle:** SGLD. **Bottom:** SGLRW. Each panel visualizes the deviation of the empirical posterior covariance from the analytic posterior covariance; the Frobenius norm (Frob) reports the total error magnitude.

Setup For each experiment, we extract fixed sequence embeddings from a pretrained OPT language model (Zhang et al., 2022) with 350M parameters by taking the final-layer representation of the last token. These embeddings are held fixed, and Bayesian posterior sampling is performed over the parameters of a two-layer binary classification head. This isolates the behaviour of the sampling algorithms from learning the data representation.

Each method is run with 15 chains for 10,000 iterations, discarding the first 5,000 iterations. For each training-set size, we vary the minibatch size to probe batch-size sensitivity while keeping learning-rate schedules and other hyperparameters matched across methods. Performance is evaluated on the test set using classification accuracy, negative log-likelihood (NLL), and expected calibration error (ECE).

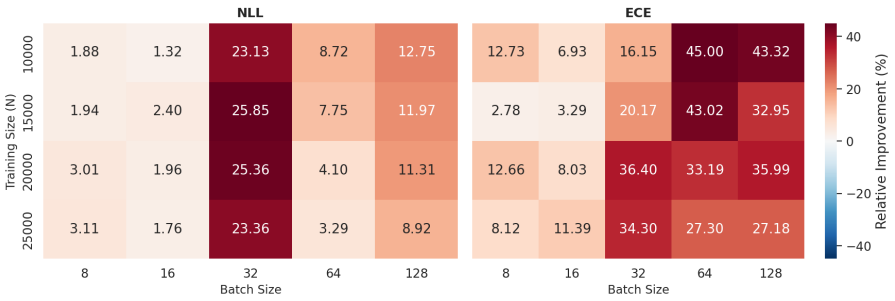

Figure 5: Relative improvement of SGLRW over clipped SGLD at increased lr scale ($\eta_0 = 1.5 \times 10^{-4}$). Heatmaps show percentage differences in negative log-likelihood (**left**) and expected calibration error (**right**)

5.3.1 Results

As highlighted, SGLRW has consistently been less sensitive to the choice of learning-rate schedule than standard SGLD; it can handle substantially larger step sizes while still maintaining stability across all training-set sizes and minibatch configurations. As such, we first compare the performance of SGLRW against standard SGLD and Clipped-SGLD with small initial step sizes. Following this, we explore the other end of the spectrum, where we compare the performance of SGLRW against Clipped-SGLD at step sizes for which standard SGLD is unstable.

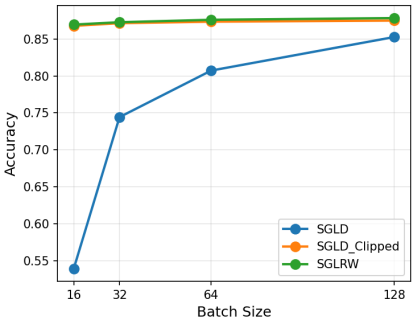

Figure 6: Predictive accuracy as a function of batch size for training-set size ($N = 25{,}000$), using base lr $\eta_0 = 7.5 \times 10^{-6}$.

Comparison at small step sizes. Fig. 6 shows predictive accuracy and NLL with respect to minibatch size for a large training-set size ($N = 25{,}000$) for a run with $\eta_0 = 7.5 \times 10^{-6}$. At small minibatch sizes, SGLD exhibits a degradation in both accuracy and NLL, whereas SGLRW remains stable across the sweep. As the minibatch size increases, the accuracy of SGLD improves, while differences in NLL persist at moderate batch sizes. Similar to SGLRW, Clipped-SGLD also outperforms SGLD in this regime, again highlighting the strength of the baseline.

Clipped-SGLD versus SGLRW at larger step sizes. We next compare Clipped-SGLD and SGLRW in regimes where standard SGLD is unstable, focusing on step sizes beyond the conservative regime. Across all batch sizes considered in Figure 5, SGLRW consistently outperforms Clipped-SGLD in terms of predictive quality, with this behaviour remaining robust across training-set sizes.

The relative advantage of SGLRW becomes most pronounced in small-to-moderate minibatch regimes. While accuracy differences remain minor, SGLRW consistently attains lower negative log-likelihood and improved calibration relative to Clipped-SGLD. A representative comparison at an increased learning-rate scale is shown in Figure 5, with complete results across learning-rate schedules reported in Appendix A.6 (Figures 9, 10, and 11).

6 Conclusion

This work introduced Stochastic Gradient Lattice Random Walk (SGLRW), a robust discretisation of Langevin dynamics for Bayesian inference. By replacing traditional Gaussian increments with coordinate-wise bounded updates, SGLRW significantly reduces sensitivity to minibatch size and stochastic gradient noise, a common failure point for standard SGLD. Our theoretical analysis demonstrated that SGLRW achieves strictly tighter mean squared error (MSE) bounds than SGLD by confining minibatch-induced noise to the off-diagonal elements of the update covariance.

Empirically, SGLRW showed superior stability and predictive performance across diverse tasks, from linear regression to LLM-based sentiment classification. This is in comparison to both standard SGLD as well as a strong baseline in the form of Clipped-SGLD. Notably, it remains stable under various conditions where SGLD diverges and maintains high calibration even with small minibatches.

Beyond its algorithmic advantages, the structure of SGLRW makes it uniquely suited for implementation on energy-efficient, low-precision, and stochastic hardware (Duffield et al., 2025), which is becoming increasingly important as the impact of AI on energy consumption and sustainability becomes a major concern (Aifer et al., 2025).

REFERENCES

Maxwell Aifer, Zach Belateche, Suraj Bramhavar, Kerem Y Camsari, Patrick J Coles, Gavin Crooks, Douglas J Durian, Andrea J Liu, Anastasia Marchenkova, Antonio J Martinez, et al. Solving the compute crisis with physics-based asics. *arXiv preprint arXiv:2507.10463*, 2025.

Armin Alaghi and John P Hayes. Survey of stochastic computing. *ACM Transactions on Embedded computing systems (TECS)*, 12(2s):1–19, 2013.

Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, 29(3):599–615, 2019.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. *Advances in neural information processing systems*, 28, 2015.

Haolin Chen and Philip N Garner. Bayesian parameter-efficient fine-tuning for overcoming catastrophic forgetting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691. PMLR, 2014.

Tom Conte, Erik DeBenedictis, Natesh Ganesh, Todd Hylton, John Paul Strachan, R Stanley Williams, Alexander Alemi, Lee Altenberg, Gavin Crooks, James Crutchfield, et al. Thermodynamic computing. *arXiv preprint arXiv:1911.01968*, 2019.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient langevin dynamics. *Advances in neural information processing systems*, 29, 2016.

Samuel Duffield, Kaelan Donatella, Johnathan Chiu, Phoebe Klett, and Daniel Simpson. Scalable Bayesian learning with posteriors. *arXiv preprint arXiv:2406.00104*, 2024.

Samuel Duffield, Maxwell Aifer, Denis Melanson, Zach Belateche, and Patrick J Coles. Lattice random walk discretisations of stochastic differential equations. *arXiv preprint arXiv:2508.20883*, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Andrew Gelman, Walter R Gilks, and Gareth O Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. *arXiv preprint arXiv:2404.11599*, 2024.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Martin Hutzenthaler and Arnulf Jentzen. *Numerical approximations of stochastic differential equations with non-globally Lipschitz continuous coefficients*, volume 236. American Mathematical Society, 2015.

Minyoung Kim and Timothy Hospedales. Lift: Learning to fine-tune via bayesian parameter efficient meta fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in mcmc land: Cutting the metropolis-hastings budget. In *International conference on machine learning*, pp. 181–189. PMLR, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Ruilin Li, Xin Wang, Hongyuan Zha, and Molei Tao. Improving sampling accuracy of stochastic gradient mcmc methods via non-uniform subsampling of gradients. *arXiv preprint arXiv:2002.08949*, 2020.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Denis Melanson, Mohammad Abu Khater, Maxwell Aifer, Kaelan Donatella, Max Hunter Gordon, Thomas Ahle, Gavin Crooks, Antonio J Martinez, Faris Sbahi, and Patrick J Coles. Thermodynamic computing system for ai applications. *Nature Communications*, 16(1):3757, 2025.

Radford M Neal. Probabilistic inference using markov chain monte carlo methods. 1993.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. 1996.

Colin Samplawski, Adam D Cobb, Manoj Acharya, Ramneet Kaur, and Susmit Jha. Scalable bayesian low-rank adaptation of large language models via stochastic variational subspace inference. *arXiv preprint arXiv:2506.21408*, 2025.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837. PMLR, 2019.

Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped Langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pp. 8970–8980. PMLR, 2020.

Joanna Sliwa, Frank Schneider, Philipp Hennig, and José Miguel Hernández-Lobato. Mitigating forgetting in low rank adaptation. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

Sebastian J Vollmer, Konstantinos C Zygalakis, and Yee Whye Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *Advances in Neural Information Processing Systems*, 37:67758–67794, 2024.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

Haotian Xiang, Jinwen Xu, and Qin Lu. Fine-tuning llms with variational bayesian last layer for high-dimensional bayesian optimzation. *arXiv preprint arXiv:2510.01471*, 2025.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Appendix

A.1 Analysis plots including Clipped SGLD

For completeness, we also include Clipped SGLD in the plots for the analysis.

For the covariance MSE analysis (Figure 3), Clipped SGLD already mitigates the sharp error explosion observed for SGLD at larger stepsizes (Figure 7). However, across batch sizes and learning rates, SGLRW consistently attains comparable or lower covariance MSE throughout the stable regime.

For the heavy-tailed noise robustness analysis (Figure 2), both Clipped SGLD and SGLRW prevent the severe instability exhibited by SGLD under heavy-tailed gradient noise. Across noise scales, SGLRW maintains a closer qualitative agreement with the target distribution than Clipped SGLD.

A.2 Proof of Theorem 4.2

In this appendix we prove Theorem 4.2. The argument follows the Poisson-equation framework of Chen et al. (2015); we restate the required notation so the appendix is self-contained.

Generator and Kolmogorov operators. Consider a continuous-time Itô diffusion on $\mathbb{R}^d$ with infinitesimal generator

$$\mathcal{L}g(\theta) = f(\theta) \cdot \nabla g(\theta) + \frac{1}{2}\, \sigma(\theta)\sigma(\theta)^\top : \nabla^2 g(\theta), \tag{21}$$

with f the drift of the SDE in equation 4. Let $(e^{t\mathcal{L}})_{t\geq 0}$ denote the associated Kolmogorov (backward) semigroup, so that for any suitable test function g,

$$\mathbb{E}[g(\theta_t) \mid \theta_0 = \theta] = (e^{t\mathcal{L}} g)(\theta). \tag{22}$$

Since $e^{t\mathcal{L}}$ is generally intractable, we consider a time-δ_t numerical update with one-step Markov operator P_{δ_t} defined by

$$\mathbb{E}[g(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}] = (P_{\delta_t} g)(\theta_{(n-1)\delta_t}). \tag{23}$$

A one-step scheme is a weak order-K local integrator if, for all sufficiently smooth g,

$$(P_{\delta_t} g)(\theta) = (e^{\delta_t \mathcal{L}} g)(\theta) + O(\delta_t^{K+1}). \tag{24}$$

Stochastic gradients and random one-step operators. In SGLD and SGLRW, the exact drift f is replaced by a minibatch approximation $\widehat{f}$. Let $\widetilde{P}_{\delta_t}^{(n)}$ denote the resulting (random) one-step Markov operator at iteration n, i.e.

$$\mathbb{E}[g(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}] = (\widetilde{P}_{\delta_t}^{(n)} g)(\theta_{(n-1)\delta_t}). \tag{25}$$

For minibatch $\mathcal{B}$, define

$$\zeta(\theta; \mathcal{B}) := f(\theta) - \widehat{f}(\theta; \mathcal{B}). \tag{26}$$

and the associated first-order differential operator

$$(\Delta V_n g)(\theta; \mathcal{B}) := \zeta(\theta; \mathcal{B}_n) \cdot \nabla g(\theta). \tag{27}$$

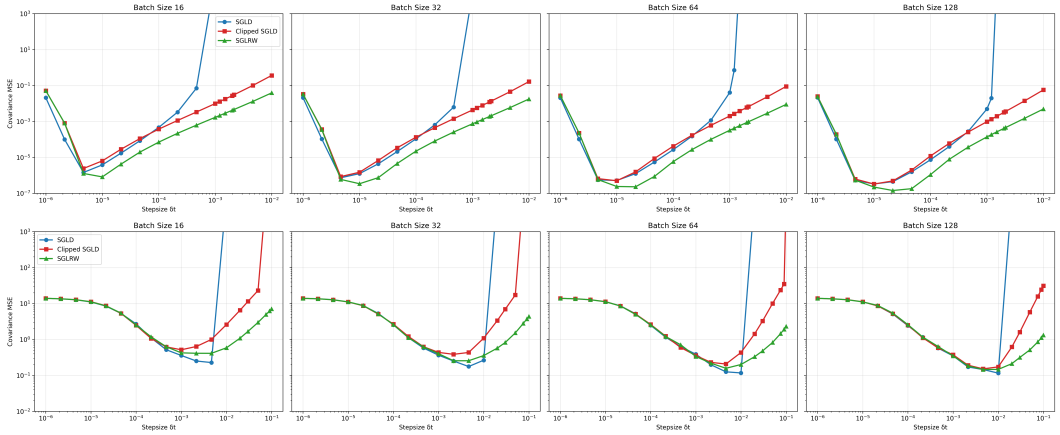

Figure 7: Mean-squared error (MSE) of the posterior covariance as a function of the stepsize δ_t, shown for different batch sizes. **Top:** 50-dimensional Bayesian linear regression. **Bottom:** Bayesian logistic regression on the breast cancer dataset (Wolberg et al., 1993).

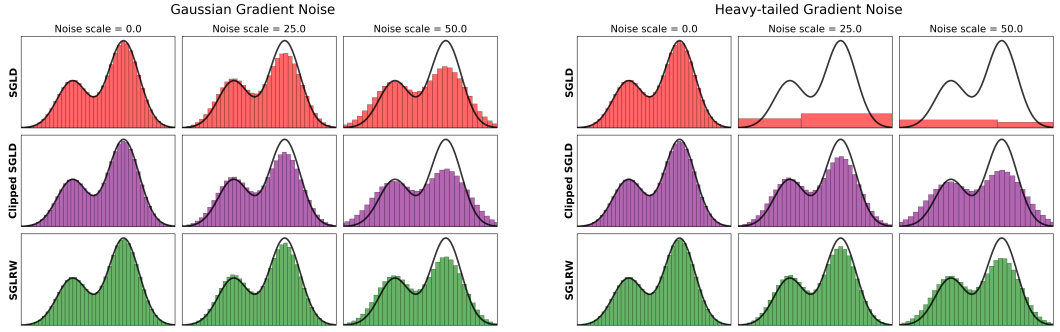

Figure 8: Multimodal univariate target with exact gradient corrupted by synthetic noise of increasing scale. **Left:** Gaussian noise ($\alpha = 2$). **Right:** heavy-tailed noise drawn from an α-stable distribution with $\alpha = 1.5$.

capturing the error due to the minibatch drift error. Note that for overdamped Langevin diffusion, we have $f = -\nabla U$, $\widehat{f}(\theta;\mathcal{B}) = -\widehat{\nabla U}(\theta;\mathcal{B})$, so that $\zeta(\theta;\mathcal{B}) = \widehat{\nabla U}(\theta;\mathcal{B}) - \nabla U(\theta)$.

Beyond the first-order drift perturbation, minibatching can induce a second-order correction through the conditional second moment of the increment. We define the second-order differential operator

$$(\Delta A_n g)(\theta) := \frac{1}{2}\, M_n(\theta) : \nabla^2 g(\theta), \tag{28}$$

with

$$M_n(\theta, \mathcal{B}_n) := \delta_t^{-2}\, \mathbb{E}_{\varepsilon_n}\big[\Delta\theta_n^{\mathrm{fb}}(\Delta\theta_n^{\mathrm{fb}})^\top - \Delta\theta_n^{\mathrm{mb}}(\Delta\theta_n^{\mathrm{mb}})^\top \;\big|\; \theta_{(n-1)\delta_t} = \theta, \mathcal{B}_n\big]. \tag{29}$$

Here $\Delta\theta_n^{\mathrm{mb}}$ and $\Delta\theta_n^{\mathrm{fb}}$ denote the one-step increments of the minibatch and full-batch updates, respectively, and $\mathbb{E}_{\varepsilon_n}[\cdot]$ denotes expectation with respect to the internal randomness of the update at step n (e.g. injected Gaussian noise or lattice path sampling).

For weak order-one integrators, the corresponding one-step Markov operator admits the weak expansion

$$\widetilde{P}_{\delta_t}^{(n)} g(\theta) = g(\theta) + \delta_t(\mathcal{L} - \Delta V_n)g(\theta) + \delta_t^2\, \mathcal{R}_n g(\theta), \tag{30}$$

where $\mathcal{R}_n$ denotes the local weak remainder operator. We assume that, for the Poisson solution ψ, the remainder satisfies the Lyapunov-weighted bound, in the sense that there exists a constant $p_0 > 0$ such that

$$|\mathcal{R}_n \psi(\theta)| \le C\, \mathcal{V}(\theta)^{p_0}, \qquad \text{uniformly in } n \text{ and } \delta_t \in (0, 1]. \tag{31}$$

Poisson equation. Given a smooth observable $\phi : \mathbb{R}^d \to \mathbb{R}$, define its stationary expectation under the invariant distribution π by

$$\bar{\phi} := \int \phi(\theta)\, \pi(\theta)\, d\theta. \tag{32}$$

We analyze ergodic averages via the Poisson equation

$$\mathcal{L}\psi = \phi - \bar{\phi}, \tag{33}$$

and express finite-time errors in terms of the corresponding solution ψ.

Lyapunov-Poisson regularity.

Assumption A.1 *Let ψ solve the Poisson equation $\mathcal{L}\psi = \phi - \bar{\phi}$. Assume there exists a function $\mathcal{V} : \mathbb{R}^d \to [1, \infty)$ such that:*

*(i) (*Derivative control*) There exist constants $C_k, p_k > 0$ for $k = 0, 1, 2, 3, 4$ such that*

$$\|D^k \psi(\theta)\| \le C_k\, \mathcal{V}(\theta)^{p_k}. \tag{34}$$

*(ii) (*Uniform moments along the chain*) For all $p \le p^*$,*

$$\sup_n \mathbb{E}[\mathcal{V}(\theta_{n\delta_t})^p] < \infty. \tag{35}$$

*(iii) (*Growth compatibility*) There exist constants p^*, $C > 0$, such that for all $p \le p^*$ and all $s \in (0, 1)$,*

$$\mathcal{V}^p(s\theta + (1-s)\vartheta) \le C\big(\mathcal{V}^p(\theta) + \mathcal{V}^p(\vartheta)\big). \tag{36}$$

*(iv) (*Uniform second-moment bound*) The second-order coefficient field satisfies*

$$\sup_n \mathbb{E}\big[\|M_n(\theta_{(n-1)\delta_t})\|_F^2\big] < \infty. \tag{37}$$

*(v) (*Increment moment control*) There exist constants $C > 0$ and exponents $q_2, q_4 > 0$ such that, for all n,*

$$\mathbb{E}_{\varepsilon_n, \mathcal{B}_n}\big[\|\Delta\theta_n\|^{2k} \,\big|\, \theta_{(n-1)\delta_t}\big] \le C\, \delta_t^k\, \mathcal{V}(\theta_{(n-1)\delta_t})^{q_{2k}}, \qquad k \in \{1, 2\}, \tag{38}$$

where $\Delta\theta_n := \theta_{n\delta_t} - \theta_{(n-1)\delta_t}$.

(vi) *(*Third-moment tensor control*) There exist constants $C > 0$ and an exponent $q_3 > 0$ such that, for all n,*

$$\left\| \mathbb{E}_{\varepsilon_n, \mathcal{B}_n} \left[(\Delta\theta_n^{\mathrm{mb}})^{\otimes 3} \mid \theta_{(n-1)\delta_t} \right] - \mathbb{E}_{\varepsilon_n} \left[(\Delta\theta_n^{\mathrm{fb}})^{\otimes 3} \mid \theta_{(n-1)\delta_t} \right] \right\| \leq C\, \delta_t^2\, \mathcal{V}(\theta_{(n-1)\delta_t})^{q_3}. \tag{39}$$

(vii) *(*Unbiased minibatch estimator*) The minibatch drift estimator is conditionally unbiased, in the sense that*

$$\mathbb{E}_{\mathcal{B}_n} \left[\zeta(\theta_{(n-1)\delta_t}; \mathcal{B}_n) \,\middle|\, \theta_{(n-1)\delta_t} \right] = 0 \qquad \text{a.s. for all } n, \tag{40}$$

where $\zeta(\theta; \mathcal{B}) := f(\theta) - \widehat{f}(\theta; \mathcal{B})$ denotes the minibatch drift error.

Note that the derivative bounds equation 34 in Assumption A.1 can be verified by constructing a Lyapunov function $\mathcal{V} : \mathbb{R}^d \to [1, \infty)$ which tends to infinity as $\theta \to \infty$, is twice continuously differentiable with bounded second derivatives, and satisfies the following conditions, as shown in the Appendix C of Chen et al. (2015):

(a) (*Lyapunov drift condition*) There exist constants $\alpha, \beta > 0$ such that the exact drift field f satisfies

$$\langle \nabla \mathcal{V}(\theta), f(\theta) \rangle \leq -\alpha\, \mathcal{V}(\theta) + \beta. \tag{41}$$

(b) (*Minibatch-induced drift fluctuations*) There exists $p_H \geq 2$ such that

$$\mathbb{E}_{\varepsilon_n, \mathcal{B}_n} \left[\|\zeta(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\|^2 \mid \theta_{(n-1)\delta_t} = \theta \right] \leq C\, \mathcal{V}(\theta)^{p_H}, \tag{42}$$

together with the growth condition

$$\|\nabla \mathcal{V}(\theta)\|^2 + \|f(\theta)\|^2 \leq C\, \mathcal{V}(\theta). \tag{43}$$

Throughout, we write $\theta_{(n-1)\delta_t} = \theta$ and recall that for overdamped Langevin diffusion we have $\zeta(\theta; \mathcal{B}) = \widehat{\nabla U}(\theta; \mathcal{B}) - \nabla U(\theta)$. Assumptions A.1(v)–(vi) hold for SGLD and SGLRW for varying requirements on the minibatch noise ζ.

1. **SGLD.**
   - *Second moment ($k = 1$):* From $\Delta\theta_n^{\mathrm{mb}} = -\delta_t(\nabla U(\theta) + \zeta) + \sqrt{2\delta_t}\, \xi_n$, we have

   $$\mathbb{E}_{\varepsilon_n, \mathcal{B}_n} \left[\|\Delta\theta_n^{\mathrm{mb}}\|^2 \mid \theta \right] \leq 2\delta_t^2\, \mathbb{E}_{\mathcal{B}_n} \left[\|\nabla U(\theta) + \zeta\|^2 \mid \theta \right] + 4d\, \delta_t. \tag{44}$$

   Thus, assuming

   $$\mathbb{E}_{\mathcal{B}_n} \left[\|\zeta\|^2 \mid \theta \right] \leq C\, \mathcal{V}(\theta)^{q_2}, \qquad \|\nabla U(\theta)\|^2 \leq C\, \mathcal{V}(\theta), \tag{45}$$

   yields

   $$\mathbb{E}_{\mathcal{B}_n, \varepsilon_n} \left[\|\Delta\theta_n^{\mathrm{mb}}\|^2 \mid \theta \right] = O(\delta_t). \tag{46}$$

   - *Fourth moment ($k = 2$):* Similarly,

   $$\mathbb{E}_{\varepsilon_n, \mathcal{B}_n} \left[\|\Delta\theta_n^{\mathrm{mb}}\|^4 \mid \theta \right] \leq C\delta_t^4\, \mathbb{E}_{\mathcal{B}_n} \left[\|\nabla U(\theta) + \zeta\|^4 \mid \theta \right] + C\delta_t^2. \tag{47}$$

   Hence, if

   $$\mathbb{E}_{\mathcal{B}_n} \left[\|\zeta\|^4 \mid \theta \right] \leq C\, \mathcal{V}(\theta)^{q_4}, \tag{48}$$

   then

   $$\mathbb{E}_{\varepsilon_n, \mathcal{B}_n} \left[\|\Delta\theta_n^{\mathrm{mb}}\|^4 \mid \theta \right] = O(\delta_t^2). \tag{49}$$

   - *Third-moment control:* From the expansion in equation 100, the difference in equation 39 arises from cubic drift–noise interaction terms and scales as $O(\delta_t^3)$. Assuming

   $$\mathbb{E}_{\mathcal{B}_n} \left[\|\zeta\|^3 \mid \theta \right] \leq C\, \mathcal{V}(\theta)^{q_3}, \tag{50}$$

   this difference is bounded by $C\, \delta_t^2\, \mathcal{V}(\theta)^{q_3}$ for $\delta_t \in (0, 1]$.
2. **SGLRW.**

- *Increment control ($k = 1, 2$):* As $\|\Delta\theta_n^{\mathrm{mb}}\|^2 = 2d\,\delta_t$ and $\|\Delta\theta_n^{\mathrm{mb}}\|^4 = 4d^2\,\delta_t^2$, the bound equation 38 holds without any assumption on ζ.
- *Third-moment control:* As shown in equation 114, the third-order conditional moments agree between minibatch and full-batch dynamics for index configurations with at least two equal indices. For distinct indices (i, j, k), the difference arises from coordinate coupling through the shared minibatch noise and is $O(\delta_t^3)$. Consequently, equation 39 holds provided

$$\mathbb{E}_{\mathcal{B}_n}\left[\|\zeta\|^3 \mid \theta\right] \le C\,\mathcal{V}(\theta)^{q_3}. \tag{51}$$

Lemma A.2 *Under Assumption A.1, the random one-step Markov operator $\widetilde{P}_{\delta_t}^{(n)}$ associated with a weak order-one integrator admits the expansion*

$$\widetilde{P}_{\delta_t}^{(n)}\psi(\theta) = \psi(\theta) + \delta_t(\mathcal{L} - \Delta V_n)\psi(\theta) - \delta_t^2\,\Delta A_n\psi(\theta) + \delta_t^2\,\mathcal{R}_n\psi(\theta), \tag{52}$$

where ΔV_n is the (minibatch-dependent) first-order drift perturbation, ΔA_n is the second-order covariance perturbation, and $|\mathcal{R}_n\psi(\theta)| \le C\,\mathcal{V}(\theta)^{p_0}$ uniformly in n and $\delta_t \in (0, 1]$.

By Taylor's theorem with integral remainder applied to the minibatch increment,

$$\begin{aligned}\psi(\theta + \Delta\theta_n^{\mathrm{mb}}) = \psi(\theta) + \nabla\psi(\theta)\cdot\Delta\theta_n^{\mathrm{mb}} + \frac{1}{2}\,\Delta\theta_n^{\mathrm{mb}}(\Delta\theta_n^{\mathrm{mb}})^\top : \nabla^2\psi(\theta)\\ + \frac{1}{6}\,\nabla^3\psi(\theta) : (\Delta\theta_n^{\mathrm{mb}})^{\otimes 3} + R_4(\theta, \Delta\theta_n^{\mathrm{mb}}),\end{aligned} \tag{53}$$

where

$$R_4(\theta, \Delta\theta) = \frac{1}{6}\int_0^1 (1-s)^3\nabla^4\psi(\theta + s\Delta\theta) : \Delta\theta^{\otimes 4}\,ds. \tag{54}$$

For a given minibatch $\mathcal{B}_n$, taking the conditional expectation yields

$$\begin{aligned}\widetilde{P}_{\delta_t}^{(n)}\psi(\theta) = \psi(\theta) + \nabla\psi(\theta)\cdot\mathbb{E}_{\varepsilon_n}[\Delta\theta_n^{\mathrm{mb}} \mid \theta] + \frac{1}{2}\,\mathbb{E}_{\varepsilon_n}[\Delta\theta_n^{\mathrm{mb}}(\Delta\theta_n^{\mathrm{mb}})^\top \mid \theta] : \nabla^2\psi(\theta)\\ + \frac{1}{6}\,\nabla^3\psi(\theta) : \mathbb{E}_{\varepsilon_n}[(\Delta\theta_n^{\mathrm{mb}})^{\otimes 3} \mid \theta] + \mathbb{E}_{\varepsilon_n}[R_4(\theta, \Delta\theta_n^{\mathrm{mb}}) \mid \theta].\end{aligned} \tag{55}$$

By definition of the increment-based drift difference,

$$\zeta(\theta; \mathcal{B}_n) = f(\theta) - \widehat{f}(\theta; \mathcal{B}), \tag{56}$$

so

$$\nabla\psi(\theta)\cdot\mathbb{E}_{\varepsilon_n}[\Delta\theta_n^{\mathrm{mb}} \mid \theta] = \delta_t\,\nabla\psi(\theta)\cdot\widehat{f}(\theta; \mathcal{B}) = \delta_t\,f(\theta)\cdot\nabla\psi(\theta) - \delta_t\,\Delta V_n\psi(\theta). \tag{57}$$

Similarly, by definition of ΔA_n,

$$\frac{1}{2}\,\mathbb{E}_{\varepsilon_n}[\Delta\theta_n^{\mathrm{mb}}(\Delta\theta_n^{\mathrm{mb}})^\top \mid \theta] : \nabla^2\psi(\theta) = \frac{1}{2}\,\mathbb{E}_{\varepsilon_n}[\Delta\theta_n^{\mathrm{fb}}(\Delta\theta_n^{\mathrm{fb}})^\top \mid \theta] : \nabla^2\psi(\theta) - \delta_t^2\,\Delta A_n\psi(\theta). \tag{58}$$

We now treat the cubic term by adding and subtracting the full-batch third moment:

$$\begin{aligned}\frac{1}{6}\,\nabla^3\psi(\theta) : \mathbb{E}_{\varepsilon_n}[(\Delta\theta_n^{\mathrm{mb}})^{\otimes 3} \mid \theta] = \frac{1}{6}\,\nabla^3\psi(\theta) : \mathbb{E}_{\varepsilon_n}[(\Delta\theta_n^{\mathrm{fb}})^{\otimes 3} \mid \theta]\\ + \frac{1}{6}\,\nabla^3\psi(\theta) : \Big(\mathbb{E}_{\varepsilon_n}[(\Delta\theta_n^{\mathrm{mb}})^{\otimes 3} \mid \theta] - \mathbb{E}_{\varepsilon_n}[(\Delta\theta_n^{\mathrm{fb}})^{\otimes 3} \mid \theta]\Big).\end{aligned} \tag{59}$$

Since the full-batch scheme is weak order one, its contribution is absorbed into the $O(\delta_t^2)$ remainder of the full-batch expansion. By Assumption A.1(vi), the difference of third-order moments is $O(\delta_t^2\,\mathcal{V}(\theta)^{q_3})$, so the net cubic contribution is $O(\delta_t^2\,\mathcal{V}(\theta)^{p_0})$.

For the fourth-order remainder, using $|T : x^{\otimes 4}| \le \|T\|\,\|x\|^4$ and Assumption A.1(i) with $k = 4$, we obtain

$$|R_4(\theta, \Delta\theta_n^{\mathrm{mb}})| \le C\,\|\Delta\theta_n^{\mathrm{mb}}\|^4\int_0^1 \mathcal{V}(\theta + s\Delta\theta_n^{\mathrm{mb}})^{p_4}\,ds. \tag{60}$$

By Assumption A.1(iii), $\mathcal{V}(\theta + s\Delta\theta_n^{\mathrm{mb}})^{p_4} \le C(\mathcal{V}(\theta)^{p_4} + \mathcal{V}(\theta + \Delta\theta_n^{\mathrm{mb}})^{p_4})$ for $s \in (0,1)$. Taking conditional expectation, applying the increment fourth-moment bound in Assumption A.1(v), and using the uniform moment bound in Assumption A.1(ii), yields

$$\mathbb{E}_{\varepsilon_n}[|R_4(\theta, \Delta\theta_n^{\mathrm{mb}})| \mid \theta, \mathcal{B}_n] \le C\,\delta_t^2\,\mathcal{V}(\theta)^{p_0}, \tag{61}$$

uniformly in n and $\delta_t \in (0,1]$.

Collecting all $O(\delta_t^2)$ contributions into $\mathcal{R}_n$ yields the stated expansion.

Theorem A.3 *Under the Assumption A.1 and that the drift perturbations are unbiased. Then there exists $C > 0$, independent of (L, δ_t), such that*

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \le C\left(\frac{1}{L^2}\sum_{n=1}^{L}\mathbb{E}\big[\|\zeta_n\|^2\big] + \frac{1}{L\delta_t} + \frac{1}{L^2\delta_t^2} + \delta_t^2 + \frac{\delta_t^2}{L}\sum_{n=1}^{L}\mathbb{E}\big[\|M_n\|^2\big]\right). \tag{62}$$

Let ψ solve the Poisson equation $\mathcal{L}\psi = \phi - \bar{\phi}$. By Lemma A.2, for each n,

$$(\widetilde{P}_{\delta_t}^{(n)}\psi)(\theta) = \psi(\theta) + \delta_t(\mathcal{L} - \Delta V_n)\psi(\theta) - \delta_t^2\,\Delta A_n\psi(\theta) + \delta_t^2\,\mathcal{R}_n\psi(\theta), \tag{63}$$

with $|\mathcal{R}_n\psi(\theta)| \le C\,\mathcal{V}(\theta)^{p_0}$ uniformly in n and $\delta_t \in (0,1]$.

Evaluating equation 63 at $\theta = \theta_{(n-1)\delta_t}$, at minibatch $\mathcal{B}_n$, and using $\mathbb{E}_{\varepsilon_n}[\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}, \mathcal{B}_n] = (\widetilde{P}_{\delta_t}^{(n)}\psi)(\theta_{(n-1)\delta_t})$ gives

$$\mathbb{E}_{\varepsilon_n}[\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}] = \psi(\theta_{(n-1)\delta_t}) + \delta_t(\mathcal{L} - \Delta V_n)\psi(\theta_{(n-1)\delta_t}) - \delta_t^2\,\Delta A_n\psi(\theta_{(n-1)\delta_t}) + \delta_t^2\,\mathcal{R}_n\psi(\theta_{(n-1)\delta_t}). \tag{64}$$

Summing equation 64 over $n = 1, \dots, L$ and using $\mathcal{L}\psi = \phi - \bar{\phi}$ yields

$$\begin{aligned}\hat{\phi} - \bar{\phi} = {} & \frac{1}{L\delta_t}\Big(\psi(\theta_{L\delta_t}) - \psi(\theta_0)\Big) - \frac{1}{L\delta_t}\sum_{n=1}^{L}\Big(\mathbb{E}_{\varepsilon_n}[\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}] - \psi(\theta_{n\delta_t})\Big) \\ & + \frac{1}{L}\sum_{n=1}^{L}\Delta V_n\psi(\theta_{(n-1)\delta_t}) + \frac{\delta_t}{L}\sum_{n=1}^{L}\Delta A_n\psi(\theta_{(n-1)\delta_t}) - \frac{\delta_t}{L}\sum_{n=1}^{L}\mathcal{R}_n\psi(\theta_{(n-1)\delta_t}).\end{aligned} \tag{65}$$

Taking squares and using $(a+b+c+d+e)^2 \le 5(a^2+b^2+c^2+d^2+e^2)$ gives

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \le C\,(A_1 + A_2 + A_3 + A_4 + A_5), \tag{66}$$

where

$$A_1 := \mathbb{E}\left[\frac{(\mathbb{E}\psi(\theta_{L\delta_t}) - \psi(\theta_0))^2}{L^2\delta_t^2}\right], \tag{67}$$

$$A_2 := \mathbb{E}\left[\frac{1}{L^2\delta_t^2}\left(\sum_{n=1}^{L}\big(\mathbb{E}_{\varepsilon_n}[\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}] - \psi(\theta_{n\delta_t})\big)\right)^2\right], \tag{68}$$

$$A_3 := \mathbb{E}\left[\frac{1}{L^2}\left(\sum_{n=1}^{L}\Delta V_n\psi(\theta_{(n-1)\delta_t})\right)^2\right], \tag{69}$$

$$A_4 := \mathbb{E}\left[\left(\frac{\delta_t}{L}\sum_{n=1}^{L}\Delta A_n\psi(\theta_{(n-1)\delta_t})\right)^2\right], \tag{70}$$

$$A_5 := \mathbb{E}\left[\left(\frac{\delta_t}{L}\sum_{n=1}^{L}\mathcal{R}_n\psi(\theta_{(n-1)\delta_t})\right)^2\right]. \tag{71}$$

(i) A_1**.** By Lyapunov control of ψ (Assumption A.1(i–ii) with $k = 0$), $\sup_n \mathbb{E}[\psi(\theta_{n\delta_t})^2] < \infty$, hence

$$A_1 \le \frac{C}{L^2\delta_t^2}. \tag{72}$$

(ii) A_2**.** Let

$$Z_n := \mathbb{E}_{\varepsilon_n}[\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}] - \psi(\theta_{n\delta_t}), \tag{73}$$

so that $\{Z_n\}_{n\geq 1}$ is a martingale difference sequence and

$$A_2 = \frac{1}{L^2\delta_t^2}\sum_{n=1}^{L}\mathbb{E}\big[\mathrm{Var}(\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t})\big]\,. \tag{74}$$

Write $\theta := \theta_{(n-1)\delta_t}$ and $\Delta\theta_n := \theta_{n\delta_t} - \theta$. By the fundamental theorem of calculus,

$$\psi(\theta + \Delta\theta_n) - \psi(\theta) = \int_0^1 \nabla\psi(\theta + s\Delta\theta_n)^\top \Delta\theta_n\, ds. \tag{75}$$

Therefore,

$$\begin{aligned}\mathrm{Var}(\psi(\theta_{n\delta_t}) \mid \theta) &\leq \mathbb{E}\big[(\psi(\theta+\Delta\theta_n) - \psi(\theta))^2 \mid \theta\big] \\ &\leq \mathbb{E}\left[\|\Delta\theta_n\|^2 \int_0^1 \|\nabla\psi(\theta + s\Delta\theta_n)\|^2\, ds \;\middle|\; \theta\right],\end{aligned} \tag{76}$$

where we used Cauchy-Schwarz.
By Assumption A.1(i), $\|\nabla\psi(x)\|^2 \leq C\,\mathcal{V}(x)^{2p_1}$, and by Assumption A.1(iii),

$$\mathcal{V}(\theta + s\Delta\theta_n)^{2p_1} \leq C\big(\mathcal{V}(\theta)^{2p_1} + \mathcal{V}(\theta+\Delta\theta_n)^{2p_1}\big), \qquad s \in (0,1). \tag{77}$$

Hence,

$$\mathrm{Var}(\psi(\theta_{n\delta_t}) \mid \theta) \leq C\,\mathbb{E}\Big[\|\Delta\theta_n\|^2\big(\mathcal{V}(\theta)^{2p_1} + \mathcal{V}(\theta_{n\delta_t})^{2p_1}\big) \mid \theta\Big]. \tag{78}$$

Taking conditional expectation with respect to $\mathcal{B}_n$ and using the tower property yields

$$\mathbb{E}[\mathrm{Var}(\psi(\theta_{n\delta_t}) \mid \theta) \mid \theta] \leq C\,\mathbb{E}\Big[\|\Delta\theta_n\|^2\big(\mathcal{V}(\theta)^{2p_1} + \mathcal{V}(\theta_{n\delta_t})^{2p_1}\big) \mid \theta\Big]. \tag{79}$$

By Assumption A.1(v), the increment satisfies the conditional second-moment bound

$$\mathbb{E}\big[\|\Delta\theta_n\|^2 \mid \theta_{(n-1)\delta_t}\big] \leq C\,\delta_t\,\mathcal{V}(\theta_{(n-1)\delta_t})^q, \tag{80}$$

for some $q > 0$. Taking total expectations and using Assumption A.1(ii) to control moments of $\mathcal{V}(\theta_{n\delta_t})$ yields

$$\mathbb{E}\big[\mathrm{Var}(\psi(\theta_{n\delta_t}) \mid \theta_{(n-1)\delta_t}, \mathcal{B}_n)\big] \leq C\,\delta_t. \tag{81}$$

Substituting this bound into the definition of A_2 gives

$$A_2 \leq \frac{1}{L^2\delta_t^2}\sum_{n=1}^{L} C\,\delta_t = \frac{C}{L\delta_t}. \tag{82}$$

(iii) A_3**.** Set $X_n := \Delta V_n \psi(\theta_{(n-1)\delta_t})$. By the assumed unbiasedness we have $\mathbb{E}[X_n \mid \theta_{(n-1)\delta_t}] = 0$, so $\{X_n\}$ is a martingale difference sequence and therefore cross-terms vanish:

$$\mathbb{E}\Big[\Big(\sum_{n=1}^{L} X_n\Big)^2\Big] = \sum_{n=1}^{L}\mathbb{E}[X_n^2]. \tag{83}$$

Moreover,

$$|X_n| = |\zeta(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\cdot\nabla\psi(\theta_{(n-1)\delta_t})| \leq \|\zeta(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\|\,\|\nabla\psi(\theta_{(n-1)\delta_t})\|, \tag{84}$$

so by Assumption A.1(i-ii),

$$\mathbb{E}[X_n^2] \leq C\,\mathbb{E}\big[\|\zeta(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\|^2\big]\,, \tag{85}$$

absorbing $\|\nabla\psi\|^2$ into the constant using the Lyapunov moment bounds. Hence

$$A_3 = \frac{1}{L^2}\sum_{n=1}^{L}\mathbb{E}[X_n^2] \leq \frac{C}{L^2}\sum_{n=1}^{L}\mathbb{E}\big[\|\zeta(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\|^2\big]\,. \tag{86}$$

(iv) A_4**.** By Cauchy-Schwarz,

$$A_4 \le \frac{\delta_t^2}{L} \sum_{n=1}^{L} \mathbb{E}\big[(\Delta A_n \psi(\theta_{(n-1)\delta_t}))^2\big]. \tag{87}$$

Using $\Delta A_n f = \frac{1}{2} M_n : \nabla^2 f$ and the Hilbert-Schmidt inequality,

$$|\Delta A_n \psi(\theta)| \le \frac{1}{2} \|M_n(\theta)\|_F \, \|\nabla^2 \psi(\theta)\|_F \le C \, \|M_n(\theta)\| \, \mathcal{V}(\theta)^{p_2}, \tag{88}$$

where we used Assumption A.1(i) with $k = 2$. Taking expectations and using Assumptions A.1(ii, iv), we obtain

$$\mathbb{E}\big[(\Delta A_n \psi(\theta_{(n-1)\delta_t}))^2\big] \le C \, \mathbb{E}\big[\|M_n(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\|^2\big]. \tag{89}$$

Therefore

$$A_4 \le \frac{C \, \delta_t^2}{L} \sum_{n=1}^{L} \mathbb{E}\big[\|M_n(\theta_{(n-1)\delta_t}; \mathcal{B}_n)\|^2\big]. \tag{90}$$

(v) A_5**.** By Cauchy-Schwarz and the remainder bound from Lemma A.2,

$$A_5 \le \frac{\delta_t^2}{L} \sum_{n=1}^{L} \mathbb{E}\big[(\mathcal{R}_n \psi(\theta_{(n-1)\delta_t}))^2\big] \le \frac{C \, \delta_t^2}{L} \sum_{n=1}^{L} \mathbb{E}\big[\mathcal{V}(\theta_{(n-1)\delta_t})^{2p_0}\big] \le C \, \delta_t^2, \tag{91}$$

using Assumption A.1(ii).

Combining the bounds on $A_1, \dots, A_5$ yields equation 62.

A.3 Conditional Covariance and Moments of SGLD and SGLRW Updates

We decompose the stochastic gradient as

$$\widehat{\nabla U}(\theta; \mathcal{B}) = \nabla U(\theta) + \zeta(\theta; \mathcal{B}), \qquad \mathbb{E}_{\mathcal{B}}[\zeta(\theta; \mathcal{B}) \mid \theta] = 0, \qquad \mathrm{Cov}_{\mathcal{B}}[\zeta(\theta; \mathcal{B}) \mid \theta] = G(\theta), \tag{92}$$

where $G(\theta)$ quantifies the minibatch-induced gradient covariance.

Recall the definition the second-order minibatch contribution M_n used in Theorem A.3:

$$M_n(\theta, \mathcal{B}_n) := \delta_t^{-2} \, \mathbb{E}_{\varepsilon_n}\big[\Delta\theta_n^{\mathrm{mb}} (\Delta\theta_n^{\mathrm{mb}})^\top - \Delta\theta_n^{\mathrm{fb}} (\Delta\theta_n^{\mathrm{fb}})^\top \;\big|\; \theta_{(n-1)\delta_t} = \theta, \mathcal{B}_n\big], \tag{93}$$

where ε_n denotes the internal randomness of the integrator.

SGLD. The standard SGLD update is $\Delta\theta_t = -\delta_t(\nabla U(\theta_t) + \zeta_t) + \sqrt{2\delta_t}\xi_t$ with $\xi_t \sim \mathcal{N}(0, I)$. Using the independence of ξ_t and ζ_t, and $\mathbb{E}[\zeta_t \mid \theta_t] = 0$, the first and second moments are:

$$\mathbb{E}[\Delta\theta_t \mid \theta_t] = -\delta_t \nabla U(\theta_t), \tag{94}$$

$$\mathbb{E}[\Delta\theta_t \Delta\theta_t^\top \mid \theta_t] = 2\delta_t I + \delta_t^2 \left(\nabla U(\theta_t) \nabla U(\theta_t)^\top + G(\theta_t)\right). \tag{95}$$

We next derive the second-order minibatch contribution M_n. Fix θ and a minibatch $\mathcal{B}$, define the full-batch and minibatch increments (sharing the same Gaussian ξ)

$$\Delta\theta^{\mathrm{fb}} = -\delta_t \nabla U(\theta) + \sqrt{2\delta_t}\, \xi, \qquad \Delta\theta^{\mathrm{mb}} = -\delta_t(\nabla U(\theta) + \zeta) + \sqrt{2\delta_t}\, \xi = \Delta\theta^{\mathrm{fb}} - \delta_t \zeta. \tag{96}$$

Expanding the outer products gives

$$\Delta\theta^{\mathrm{mb}} \Delta\theta^{\mathrm{mb}\top} - \Delta\theta^{\mathrm{fb}} \Delta\theta^{\mathrm{fb}\top} = -\delta_t \, \Delta\theta^{\mathrm{fb}} \zeta^\top - \delta_t \, \zeta (\Delta\theta^{\mathrm{fb}})^\top + \delta_t^2 \, \zeta\zeta^\top. \tag{97}$$

Taking conditional expectation over the internal randomness ξ and using $\mathbb{E}_\xi[\Delta\theta^{\mathrm{fb}} \mid \theta] = -\delta_t \nabla U(\theta)$ yields

$$\mathbb{E}_\xi\big[\Delta\theta^{\mathrm{mb}} \Delta\theta^{\mathrm{mb}\top} - \Delta\theta^{\mathrm{fb}} \Delta\theta^{\mathrm{fb}\top} \;\big|\; \theta, \mathcal{B}\big] = \delta_t^2 \Big(\zeta\zeta^\top + \nabla U(\theta) \zeta^\top + \zeta \nabla U(\theta)^\top\Big). \tag{98}$$

Therefore, with M_n defined as in equation 28,

$$M_{n,\mathrm{SGLD}}(\theta;\mathcal{B}) = \zeta(\theta;\mathcal{B})\zeta(\theta;\mathcal{B})^\top + \nabla U(\theta)\,\zeta(\theta;\mathcal{B})^\top + \zeta(\theta;\mathcal{B})\,\nabla U(\theta)^\top. \tag{99}$$

Averaging additionally over minibatches gives $\mathbb{E}_{\mathcal{B}}[M_{n,\mathrm{SGLD}}(\theta,\mathcal{B}) \mid \theta] = G(\theta)$.

For the third moment, let $u_t = -\delta_t(\nabla U + \zeta_t)$ and $w_t = \sqrt{2\delta_t}\xi_t$. Expanding $\mathbb{E}[(u_t + w_t)^{\otimes 3} \mid \theta_t]$, terms with odd powers of ξ_t vanish. The remaining terms are $\mathbb{E}[u_t^{\otimes 3}]$ and the cross-terms $\mathbb{E}[u_t \otimes w_t \otimes w_t]$ (and permutations). The leading $O(\delta_t^2)$ error comes from the cross-terms:

$$\mathbb{E}[u_{t,i}w_{t,j}w_{t,k} \mid \theta_t] = \mathbb{E}[-\delta_t(\partial_i U + \zeta_i)\cdot 2\delta_t\delta_{jk}] = -2\delta_t^2\partial_i U\delta_{jk}.$$

Because $\mathbb{E}[\zeta_i] = 0$, the $O(\delta_t^2)$ part of the tensor is identical for full-batch and minibatch schemes. The third-order moments of the noise appear in the $\mathbb{E}[u_t^{\otimes 3}]$ expansion. By expanding the cubic terms and using $\mathbb{E}[\zeta_t] = 0$, the tensor entries decompose as follows:

$$\mathbb{E}[\Delta\theta_i\,\Delta\theta_j\,\Delta\theta_k \mid \theta_t] = \begin{cases} -6\delta_t^2\,\partial_i U - \delta_t^3\Big((\partial_i U)^3 + 3\,\partial_i U\,G_{ii} + \mathbb{E}[\zeta_i^3 \mid \theta_t]\Big), & i=j=k, \\ -2\delta_t^2\,\partial_k U - \delta_t^3\Big((\partial_i U)^2\,\partial_k U + \partial_k U\,G_{ii} + 2\,\partial_i U\,G_{ik} + \mathbb{E}[\zeta_i^2\zeta_k \mid \theta_t]\Big), & i=j\neq k, \\ -\delta_t^3\Big(\partial_i U\,\partial_j U\,\partial_k U + \sum_{\mathrm{cyc}(i,j,k)}\partial_i U G_{jk} + \mathbb{E}[\zeta_i\zeta_j\zeta_k \mid \theta_t]\Big), & i,j,k \text{ all distinct}, \end{cases} \tag{100}$$

and permutations.

SGLRW. For the lattice random walk (LRW) discretisation, each coordinate i takes a binary step $\Delta\theta_{t,i} \in \{\pm\sqrt{2\delta_t}\}$ with probabilities

$$\mathbb{P}\Big[\Delta\theta_{t,i} = \sqrt{2\delta_t}\,\Big|\,\theta_t,\zeta_t\Big] = \tfrac{1}{2} - \tfrac{1}{2}\sqrt{\tfrac{\delta_t}{2}}\,[\partial_i U(\theta_t) + \zeta_{t,i}], \tag{101}$$

$$\mathbb{P}\Big[\Delta\theta_{t,i} = -\sqrt{2\delta_t}\,\Big|\,\theta_t,\zeta_t\Big] = \tfrac{1}{2} + \tfrac{1}{2}\sqrt{\tfrac{\delta_t}{2}}\,[\partial_i U(\theta_t) + \zeta_{t,i}]. \tag{102}$$

Conditionally on (θ_t,ζ_t) the coordinates are independent. A short calculation gives

$$\mathbb{E}[\Delta\theta_t \mid \theta_t,\zeta_t] = -\delta_t(\nabla U(\theta_t) + \zeta_t), \tag{103}$$

$$\mathbb{E}[\Delta\theta_t^2 \mid \theta_t,\zeta_t] = 2\delta_t I. \tag{104}$$

Averaging over ζ_t yields $\mathbb{E}[\Delta\theta_t \mid \theta_t] = -\delta_t\nabla U(\theta_t)$ and $\mathbb{E}[\Delta\theta_{t,i}^2 \mid \theta_t] = 2\delta_t$. For off–diagonal elements $(i \neq j)$,

$$\mathbb{E}[\Delta\theta_{t,i}\Delta\theta_{t,j} \mid \theta_t] = \mathbb{E}[\mathbb{E}[\Delta\theta_{t,i} \mid \zeta_t]\mathbb{E}[\Delta\theta_{t,j} \mid \zeta_t] \mid \theta_t] = \delta_t^2\big[\partial_i U(\theta_t)\partial_j U(\theta_t) + G_{ij}(\theta_t)\big], \tag{105}$$

hence we have

$$\mathbb{E}[\Delta\theta_t\Delta\theta_t^\top \mid \theta_t] = 2\delta_t I + \delta_t^2\,\mathrm{offdiag}\big(\nabla U(\theta_t)\nabla U(\theta_t)^\top + G(\theta_t)\big). \tag{106}$$

We next compute the second-order minibatch contribution M_n. Fix θ and a minibatch $\mathcal{B}$. For SGLRW, since $\Delta\theta_i^2 \equiv 2\delta_t$ deterministically, the diagonal entries of $\mathbb{E}_\varepsilon[\Delta\theta\Delta\theta^\top \mid \theta,\mathcal{B}]$ coincide with their full-batch counterparts, and therefore

$$\big(M_{n,\mathrm{SGLRW}}(\theta,\mathcal{B})\big)_{ii} = 0, \qquad i = 1,\dots,d. \tag{107}$$

For $i \neq j$, conditional independence (given $(\theta,\mathcal{B})$) implies

$$\begin{aligned}\mathbb{E}_\varepsilon[\Delta\theta_i\Delta\theta_j \mid \theta,\mathcal{B}] &= \mathbb{E}_\varepsilon[\Delta\theta_i \mid \theta,\mathcal{B}]\,\mathbb{E}_\varepsilon[\Delta\theta_j \mid \theta,\mathcal{B}] \\ &= \delta_t^2\big(\partial_i U(\theta) + \zeta_i\big)\big(\partial_j U(\theta) + \zeta_j\big),\end{aligned} \tag{108}$$

where we used equation 103. The corresponding full-batch term is $\mathbb{E}_\varepsilon[\Delta\theta_i^{\mathrm{fb}}\Delta\theta_j^{\mathrm{fb}} \mid \theta] = \delta_t^2\,\partial_i U(\theta)\partial_j U(\theta)$ for $i \neq j$. Subtracting and rescaling therefore gives, for $i \neq j$,

$$\big(M_{n,\mathrm{SGLRW}}(\theta;\mathcal{B})\big)_{ij} = \partial_i U(\theta)\,\zeta_j + \zeta_i\,\partial_j U(\theta) + \zeta_i\zeta_j. \tag{109}$$

Equivalently,

$$M_{n,\mathrm{SGLRW}}(\theta;\mathcal{B}) = \mathrm{offdiag}\big(\zeta\zeta^\top + \nabla U(\theta)\zeta^\top + \zeta\nabla U(\theta)^\top\big) = \mathrm{offdiag}\big(M_{n,\mathrm{SGLD}}(\theta;\mathcal{B})\big). \tag{110}$$

Averaging additionally over minibatches gives $\mathbb{E}_{\mathcal{B}}[M_{n,\mathrm{SGLRW}}(\theta,\mathcal{B}) \mid \theta] = \mathrm{offdiag}(G(\theta))$.

Since $\Delta\theta_{t,i} \in \{\pm\sqrt{2\delta_t}\}$, we have the identity $\Delta\theta_{t,i}^3 = 2\delta_t\,\Delta\theta_{t,i}$. Higher-order moments are computed via the law of iterated expectations. For the case where at least two indices are equal:

$$\mathbb{E}[\Delta\theta_{t,i}^3 \mid \theta_t] = 2\delta_t \mathbb{E}[\Delta\theta_{t,i} \mid \theta_t] = -2\delta_t^2\,\partial_i U(\theta_t), \tag{111}$$

$$\mathbb{E}[\Delta\theta_i^2 \Delta\theta_k \mid \theta_t] = \mathbb{E}[2\delta_t \cdot \mathbb{E}[\Delta\theta_k \mid \theta_t, \zeta_t] \mid \theta_t] = \mathbb{E}[2\delta_t(-\delta_t(\partial_k U + \zeta_k))] = -2\delta_t^2 \partial_k U(\theta_t). \tag{112}$$

For fully distinct indices, the coordinates are coupled by the noise ζ_t:

$$\begin{aligned}\mathbb{E}[\Delta\theta_i\Delta\theta_j\Delta\theta_k \mid \theta_t] &= \mathbb{E}[(-\delta_t(\partial_i U + \zeta_i))(-\delta_t(\partial_j U + \zeta_j))(-\delta_t(\partial_k U + \zeta_k)) \mid \theta_t] \\ &= -\delta_t^3 \left(\partial_i U \partial_j U \partial_k U + \sum_{\mathrm{cyc}(i,j,k)} \partial_i U G_{jk} + \mathbb{E}[\zeta_i\zeta_j\zeta_k \mid \theta_t] \right). \end{aligned} \tag{113}$$

The resulting third-moment tensor entries are:

$$\mathbb{E}[\Delta\theta_{t,i}\Delta\theta_{t,j}\Delta\theta_{t,k} \mid \theta_t] = \begin{cases} -2\delta_t^2\,\partial_i U(\theta_t), & i=j=k, \\ -2\delta_t^2\,\partial_k U(\theta_t), & i=j\neq k, \\ -\delta_t^3\left(\partial_i U\partial_j U\partial_k U + \sum_{\mathrm{cyc}(i,j,k)} \partial_i U G_{jk} + \mathbb{E}[\zeta_i\zeta_j\zeta_k \mid \theta_t]\right), & i,j,k \text{ all distinct}, \end{cases} \tag{114}$$

and permutations.

A.4 Full-Increment Covariance Analysis.

We analyze *full-increment clipping* for (SG)LD and show that it yields an incorrect diffusion limit.

We consider the SGLD increment

$$\Delta\theta_t = -\delta_t\,\widehat{\nabla U}(\theta_t) + \sqrt{2\delta_t}\,\xi_t, \qquad \xi_t \sim \mathcal{N}(0, I), \tag{115}$$

and set the clipping radius $R_t = \sqrt{2\delta_t}$. Introduce the rescaled increment

$$Z_t := \frac{\Delta\theta_t}{\sqrt{2\delta_t}} = \xi_t - \sqrt{\frac{\delta_t}{2}}\,\widehat{\nabla U}(\theta_t). \tag{116}$$

Then we define the clipped increment by $\widetilde{\Delta\theta}_t := \sqrt{2\delta_t}\,\mathrm{clip}(Z_t)$, where $(\mathrm{clip}(x;R)_i = \mathrm{sign}(x_i)\min\{|x_i|, R\}$. Since clip is bounded and Lipschitz, and since $Z_t \to \xi$ in distribution (and in second moment) conditionally on θ_t whenever $\widehat{\nabla U}(\theta_t)$ has finite second moment, we obtain

$$\lim_{\delta_t\to 0} \frac{1}{\delta_t}\,\mathrm{Cov}\big[\widetilde{\Delta\theta}_t \mid \theta_t\big] = \lim_{\delta_t\to 0} 2\,\mathrm{Cov}\big(\mathrm{clip}(Z_t) \mid \theta_t\big) = 2\,\mathrm{Cov}\big(\mathrm{clip}(\xi)\big). \tag{117}$$

Thus the diffusion limit is determined by the covariance of $\mathrm{clip}(\xi)$. By independence across coordinates, this yields

$$\mathrm{Cov}\big(\mathrm{clip}(\xi)\big) = sI, \qquad s = \mathbb{E}[\min(\xi_1^2, 1)] = 1 - \sqrt{\frac{2}{\pi}}e^{-1/2} \approx 0.516, \tag{118}$$

which can be shown by explicitly computing $\mathbb{E}[\min(\xi_1^2, 1)]$. Therefore we have

$$\lim_{\delta_t\to 0} \frac{1}{\delta_t}\,\mathrm{Cov}\big[\widetilde{\Delta\theta}_t \mid \theta_t\big] = 2s\,I. \tag{119}$$

The limiting covariance equals $2sI$ with $s < 1$, whereas the Langevin diffusion requires covariance $2I$. Hence full-increment clipping yields an incorrect diffusion limit and breaks convergence to the target distribution even as $\delta_t \to 0$.

A.5 Additional Experiment: UCI Bayesian Logistic Regression

We now compare the sensitivity of SGLRW, SGLD and Clipped-SGLD on a non-Gaussian posterior sampling task, specifically logistic regression with the breast cancer dataset (Wolberg et al., 1993). The UCI breast cancer dataset consists of 569 samples, 30 features and 2 classes, which results in a 31-dimensional posterior distribution.

Table 2: Inferred Kullback–Leibler divergence for the logistic regression problem. The KL divergence is measured between Gaussian distributions fitted to the empirical mean and covariance of a gold-standard reference sample and those obtained by each algorithm under the specified hyperparameters. **Bold** indicates the lowest KL divergence for a given (B, δ_t) pair.

Hyperparameters		**KL Divergence**		
B	δ_t	**SGLD**	**SGLRW**	**Clipped SGLD**
1	10^{0}	inf	**8.3504**	9.3560
1	10^{-1}	16.6812	**6.0144**	6.4732
1	10^{-2}	10.3472	**3.9594**	4.2549
2	10^{0}	inf	**7.6706**	9.1187
2	10^{-1}	8.9809	**5.0197**	5.5856
2	10^{-2}	5.4982	**2.6152**	2.8710
4	10^{0}	27.3098	**6.9768**	8.8915
4	10^{-1}	3.7046	**3.6632**	4.2974
4	10^{-2}	3.0155	**1.4698**	1.6333
8	10^{0}	19.0951	**6.0397**	8.4212
8	10^{-1}	**1.3149**	1.9429	2.4981
8	10^{-2}	1.9578	**1.0051**	1.0553
16	10^{0}	12.0667	**4.5631**	7.0883
16	10^{-1}	**0.4993**	0.9423	1.2400
16	10^{-2}	1.5158	0.8059	**0.8027**
32	10^{0}	6.7814	**2.8940**	4.9928
32	10^{-1}	**0.2506**	0.4538	0.5611
32	10^{-2}	1.3235	0.6845	**0.6629**
64	10^{0}	3.4559	**1.7535**	3.0879
64	10^{-1}	**0.1769**	0.2153	0.2417
64	10^{-2}	1.2267	0.6490	**0.6169**

Setup Since the true posterior is not available analytically, we compare to a gold-standard sample generated with NUTS (Hoffman et al., 2014) via Pyro (Bingham et al., 2019). Throughout, we use a standard Gaussian prior on all parameters. In all cases, we ran 5,000 parallel chains for 1,000 steps, retaining only the final sample of each chain. All runs are averaged over 5 seeds.

A.5.1 Results

Comparing the inferred KL divergence of the three different methods in Table 2, we see that SGLRW consistently outperforms SGLD and Clipped-SGLD across learning-rate settings, similar to what was observed in the linear regression experiment. However, in contrast to the linear regression experiments where Clipped-SGLD performed roughly similarly to SGLD and significantly worse than SGLRW, for the logistic regression experiment considered here Clipped-SGLD shows itself as a strong baseline. In the small-batch limit, where we observe a complete failure of SGLD, Clipped-SGLD performs only slightly worse than SGLRW, and occasionally better as the batch size increases.

A.6 Full Experimental Details: Sentiment Classification

This appendix provides complete experimental details and full results for the sentiment classification experiments presented in Section 5.3. We report results for all evaluated learning-rate schedules, training-set sizes, and minibatch sizes, using fixed embeddings extracted from a pretrained OPT-350M model.

Experimental grid. For each method (clipped SGLD and SGLRW), we evaluate training set sizes $N \in \{10{,}000, 15{,}000, 20{,}000, 25{,}000\}$ and minibatch sizes $B \in \{8, 16, 32, 64, 128\}$. For each configuration, we perform three independent runs, each consisting of 15 independent chains of length 10,000, discarding the first 5,000 iterations as burn-in.

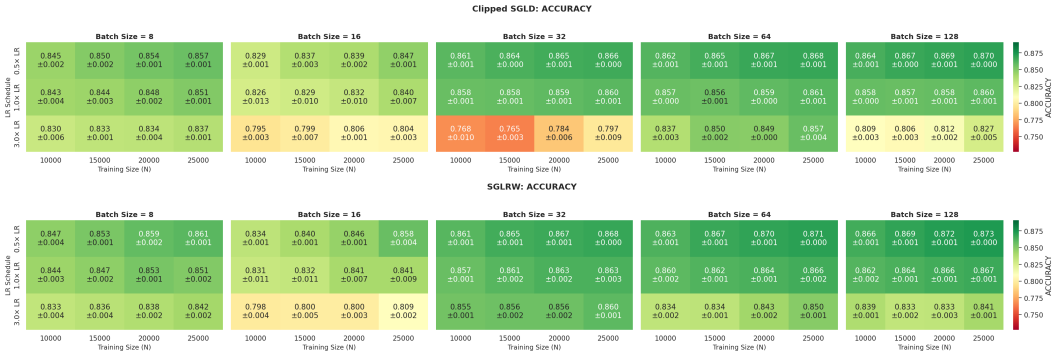

Figure 9: Predictive accuracy heatmaps for sentiment classification using fixed OPT-350M embeddings. **Top:** Clipped SGLD. **Bottom:** SGLRW. Each panel corresponds to a minibatch size B, with columns showing training-set size N and rows indicating learning-rate scale. Values report mean test accuracy across chains, with standard deviation shown below each entry.

Learning-rate schedules. For the experiments we employ a decaying learning-rate schedule of the form

$$\delta_t = s \cdot \eta_0 \, (t+1)^{-0.55}, \tag{120}$$

where $\eta_0 = 5 \times 10^{-5}$ and $s \in \{0.5, 1.0, 3.0\}$ denotes a multiplicative scale factor. We refer to these as the conservative ($0.5\times$), baseline ($1.0\times$), and increased ($3.0\times$) learning-rate scales, respectively.

Metrics. Predictive metrics are computed on the held-out test set using posterior predictive probabilities obtained by averaging predicted probabilities across retained MCMC samples. Classification accuracy is computed by thresholding the averaged probability at 0.5. The negative log-likelihood (NLL) is computed as

$$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \big[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \big], \tag{121}$$

where $y_i \in \{0, 1\}$ is the true label for test example i, N is the number of test examples, and p_i denotes the averaged posterior predictive probability for sample i. Expected calibration error (ECE) is computed using $K = 10$ equal-width bins over $[0, 1]$,

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left| \mathrm{acc}(B_k) - \mathrm{conf}(B_k) \right|, \tag{122}$$

where $\mathrm{acc}(B_k)$ and $\mathrm{conf}(B_k)$ denote the empirical accuracy and mean predicted probability within bin B_k. Reported values correspond to means across chains; variability across chains is shown where indicated.

Reading the heatmaps. Figures 9, 10, and 11 report absolute predictive accuracy, negative log-likelihood (NLL), and expected calibration error (ECE) for clipped SGLD and SGLRW across the full experimental grid. These figures complement the relative-improvement summaries shown in the main text and allow inspection of absolute performance across regimes.

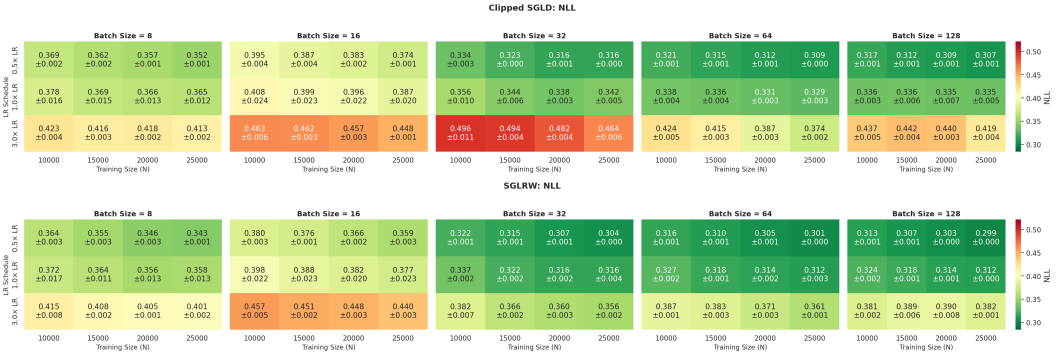

Figure 10: NLL heatmaps for sentiment classification using fixed OPT-350M embeddings. **Top:** Clipped SGLD. **Bottom:** SGLRW. Each panel corresponds to a minibatch size B, with columns showing training-set size N and rows indicating learning-rate scale. Values report mean NLL values across chains, with standard deviation shown below each entry.

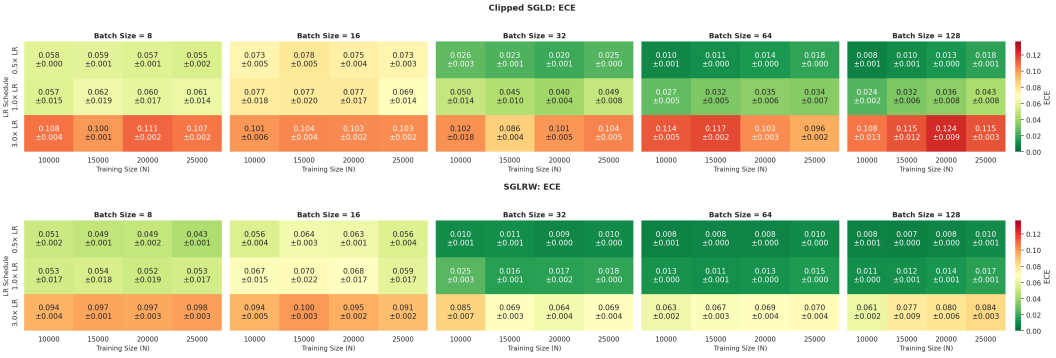

Figure 11: ECE heatmaps for sentiment classification using fixed OPT-350M embeddings. **Top:** Clipped SGLD. **Bottom:** SGLRW. Each panel corresponds to a minibatch size B, with columns showing training-set size N and rows indicating learning-rate scale. Values report mean ECE values across chains, with standard deviation shown below each entry.