On the convergence of SGLD

Rohan Hitchcock

28 November 2024

Throughout we consider a probability distribution π on $\mathbb{R}^d$. We assume that π is absolutely continuous with respect to the Lebesgue measure, and that it is non-zero everywhere and differentiable (so that $\nabla \log \pi(x)$ is well-defined). In the context of local learning coefficient estimation [Lau+24] π is the tempered posterior distribution

$$\pi(w) \propto \varphi(w) \prod_{i=1}^{n} p(Y_i \mid w)^{\beta} = \varphi(w) \exp(-\beta n L_n(w))$$

where $p(y \mid w)$ is a statistical model, $\varphi(w)$ the prior, $\{Y_1, \ldots, Y_n\}$ an iid dataset, $\beta > 0$ a fixed parameter and $L_n(w)$ the empirical negative log-likelihood.

1 Continuous-time Langevin dynamics

Given the distribution π, the corresponding *Langevin diffusion* is the stochastic differential equation (SDE)

$$dX_t = \tfrac{1}{2} \nabla \log \pi(X_t) dt + dW_t \tag{1.1}$$

where W_t is standard Brownian motion. A solution X_t to (1.1) is a continuous-time stochastic which satisfies a corresponding integral equation[1].

The SDE (1.1) is an example of a *time-homogeneous Itô diffusion* (see [Øks13, Chapter 7]). In particular this means that a solution X_t is a Markov process — a stochastic process where the behavior depends only on the current state (and not the history of the process). For any Markov process X_t we can define a family of probability measures

$$P_X^t(x, A) = \mathbf{P}(X_t \in A \mid X_0 = x)$$

which describe the distribution of X_t at time t, starting at x.

Definition 1. The *stationary distribution* of a Markov process X_t is a probability measure ν satisfying

$$\nu(A) = \int P_X^t(x, A) d\nu(x)$$

for all measurable sets A and times $t > 0$.

Theorem 2 (see [RT96, Theorem 2.1])**.** *Suppose that $\nabla \log \pi(x)$ is continuously differentiable and that for some $M, a, b \in \mathbb{R}$ we have*

$$\nabla \log \pi(x) \cdot x \leq a\|x\|^2 + b \qquad \text{for all } \|x\| > M \tag{1.2}$$

Then a solution to (1.1) X_t:

- *is non-explosive (does not reach infinity in finite time) and has a number of other nice properties.*

[1] For precise definitions see [Øks13] or my talk [Hit23].

- *has π as its stationary distribution and for all $x \in \mathbb{R}^d$*

$$\|P_X^t(x,\cdot) - \pi\| \to 0 \qquad as \qquad t \to \infty$$

where $\|\mu - \nu\| = \frac{1}{2}\sup_A |\mu(A) - \nu(A)|$ is the total variation norm.

The proof of [RT96, Theorem 2.1] seems standard, but relies on a lot of general results about Markov processes and Itô diffusions which require some work to understand. There are a lot of books about Markov processes:

- [Øks13] has a good introduction to diffusion processes but is not quite advanced enough.
- [EK05] and [RW00b; RW00a] are comprehensive but written as reference texts.
- [IW11] (particularly Chapter 5.4) has a good account of diffusion processes on manifolds not present in the above. It was written much earlier than the other books (by students of Itô!).
- [MT09] focuses on discrete-time Markov processes, which is useful when considering discretisations of (1.1).

1.1 The generator of a Markov process

We now take a brief digression to discuss the generator of a Markov process. Since we aren't giving any proofs we won't actually use what follows in this talk, but these ideas are essential in proving Theorem 2 and that SGLD samples from π under certain conditions in [TTV15, Theorem 7].

Let X_t be a time-homogeneous Markov process and assume $x \mapsto P_X^t(x, A)$ is measurable for all A (this is true of any diffusion process). Let $\mathcal{B}(\mathbb{R}^d)$ be the set of bounded measurable functions $\mathbb{R}^d \to \mathbb{R}$ (a Banach space). Associated to X_t, we define a family $(P_X^t)_{t\geq 0}$ of bounded linear operators on $\mathcal{B}(\mathbb{R}^d)$ by

$$(P_X^t f)(x) = \mathbf{E}[f(X_t) \mid X_0 = x] \text{ where } f \in \mathcal{B}(\mathbb{R}^d),\ x \in \mathbb{R}^d\ .$$

This is called the *transition semigroup* of X_t. Note that the measure $P_X^t(x,\cdot)$ can be identified with a linear functional $\mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$ by taking expectations. Reformulating $x \mapsto P_X^t(x,\cdot)$ by currying gives us the family of operators above. That $P_X^t f \in \mathcal{B}(\mathbb{R}^d)$ follows from the fact $P_X^t(x,\cdot)$ is a probability measure and the assumption that $x \mapsto P_X^t(x, A)$ is measurable.

Definition 3. The *generator* of $(P_t)_{t\geq 0}$ is an operator defined by the limit

$$(\mathcal{A}f)(x) \coloneqq \lim_{t\to 0} \frac{(P_X^t f)(x) - f(x)}{t}\ . \tag{1.3}$$

Let $\mathcal{D}(\mathcal{A})$ be the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ for which the limit (1.3) exists. We consider $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to \mathbb{R}^{\mathbb{R}^d}$.

For any time-homogeneous Itô diffusion we have an expression for its generator as a differential operator depending on the coefficients of the SDE [Øks13, Theorem 7.3.3]. For the Langevin diffusion we have

$$(\mathcal{A}f)(x) = \tfrac{1}{2}\nabla \log \pi(x) \cdot \nabla f(x) + \nabla^2 f(x)$$

where ∇^2 is the Laplacian.

When we restrict ourselves to nice classes of Markov processes, the generator uniquely characterises the process. This provides a nice way of defining diffusion processes on Riemannian manifolds, which come equipped with a notion of ∇ and ∇^2 (see [IW11, Chapter 5.4]).

The generator is practically useful for proving things, for example it can be used to characterise the stationary distribution of the process. If X_t is a Markov process and $\mathcal{A}$ is its generator

then there is a class of functions $\mathcal{C}$ such that the following statement holds. A probability measure ν is the stationary distribution of X_t if and only if

$$\nu(\mathcal{A}f) = 0$$

for all $f \in \mathcal{C}$. For example, when X_t is a non-explosive diffusion a suitable class $\mathcal{C}$ is the set of all twice continuously differentiable functions.

2 Discretisations of Langevin dynamics

Theorem 2 suggests that a discretisation of Langevin dynamics could be used to sample from $\pi(x)$. The obvious approach

$$x_{k+1} = x_k + \Delta x_k \qquad \text{where} \qquad \Delta x_k = \tfrac{\epsilon}{2}\nabla \log \pi(x_k) + \sqrt{\epsilon}\eta_k \tag{2.1}$$

where $\epsilon > 0$ is the step size and $(\eta_k)_{k\geq 0}$ is an iid sequence of standard normal random variables. This was first suggested as a sampling method in [Par81] and is now known as the *Unadjusted Langevin Algorithm* (ULA). Note that (2.1) is very similar to SGLD, except that we assume we have access to the true gradient rather than an estimator.

2.1 Issues with the naïve approach

Unfortunately ULA can fail for reasons relating to the tails of $\pi(x)$. The failure modes of ULA are discussed in detail in [RT96, Section 3]; here we present the results relevant for singular models. Consider the Markov chain $(x_k)_{k\geq 0}$ generated via (2.1) and for simplicity suppose $x_k \in \mathbb{R}$. Their analysis of ULA is organised around the limits

$$S_a^+ := \frac{\epsilon}{2} \lim_{x\to\infty} \nabla \log \pi(x) \cdot x^{-a} \tag{2.2}$$

$$S_a^- := \frac{\epsilon}{2} \lim_{x\to\infty} \nabla \log \pi(x) \cdot |x|^{-a} \tag{2.3}$$

each defined for fixed $a \in \mathbb{R}$. We paraphrase [RT96, Theorem 3.2] as follows.

Lemma 4. *Let K be any compact neighborhood of the origin. Then the number of times $(x_k)_{k\geq 0}$ visits K is finite with positive probability if either of the following conditions hold:*

- *For some $a > 1$ both $S_a^+ < 0$ and $S_a^- > 0$ exist.*
- *For $a = 1$ both $S_a^+ < -2$ and $S_a^- > 2$ exist.*

Proof. We give a sketch as follows, see [RT96, Theorem 3.2]. Suppose $x > 0$ is large, so that

$$S_a^+ \approx \frac{\epsilon}{2} \lim_{x\to\infty} \nabla \log \pi(x) \cdot x^{-a}$$

Then, under the update rule (2.1) the next expected position x' is approximately $x' \approx x + S_a^+$. If $a = 1$ then we have

$$x' \approx x(1 + S_a^+)$$

if $S_a^+ < -2$ then $x' < -x$. The same holds for the next step, replacing S_a^+ by S_a^- and so we get oscillation between increasingly extreme values of x. A similar argument holds for $a > 1$. That we eventually arrive at a sufficiently large value of x follows essentially from the fact that Brownian motion is ergodic. □

Example 5. Consider $\pi(x) \propto \exp(-\gamma x^{2b})$ for $b \in \mathbb{N}$, $\gamma > 0$. We have

$$\nabla \log \pi(x) = -2\gamma b x^{2b-1}$$

and

$$S_a^+ = \tfrac{\epsilon}{2} \lim_{x\to\infty} \nabla \log \pi(x) x^{-a} = -\epsilon\gamma b \lim_{x\to\infty} x^{2b-1-a} \,.$$

Note that $S_a^+ < 0$ if and only if $2b-1-a = 0$. Thus if $2b \geq 4$ (i.e. $\log \pi(x)$ has a non-degenerate critical point) then $a \geq 3$ and ULA will fail to sample from $\pi(x)$ by Lemma 4.

Example 5 shows a situation where using ULA to sample from distributions for which $\log \pi(x)$ has degenerate critical points will result in the sampling chain diverging to infinity. This happens because the discretisation errors versus (1.1) accumulate. Rather than arising directly from the degenerate critical point, this occurs because these distributions have lighter-than-Guassian tails. The gradient $\nabla \log \pi(x)$ grows very large in the tails, and this causes divergence.

[RT96] discuss the Metropolis Adjusted Langevin Algorithm (MALA), which combines Langevin dynamics with a proposal-acceptance mechanism as in the random walk Metropolis-Hastings algorithm. The ULA update (2.1) is used to generate a proposal, which is then accepted (the proposal is taken as a step) or rejected (a new proposal is generated) with a certain probability (see [RT96, p. 1.4.2]). The convergence of MALA is discussed in-detail in [RT96].

2.2 Stochastic Gradient Langevin Dynamics (SGLD)

Stochastic Gradient Langevin Dynamics (SGLD) uses the same update rule as ULA, except that a stochastic estimator of $\nabla \log \pi(x)$ is used in place of the true gradient. That is, a Markov chain is generated as

$$x_{k+1} = x_k + \Delta x_k \qquad \text{where} \qquad \Delta x_k = \frac{\epsilon_k}{2} g(x_k, U_k) + \sqrt{\epsilon_k}\eta_k \tag{2.4}$$

where $g(x_k, U_k)$ is an unbiased estimator of the gradient, $(U_k)_{k\geq 0}$ are iid random variables which contribute to the estimator $g(x_k, U_k)$, $(\epsilon_k)_{k\geq 0}$ is a sequence of step sizes (in the simplest case we would take $\epsilon_k = \epsilon_0$), and $(\eta_k)_{k\geq 0}$ is a sequence of iid standard normal random variables.

The typical example of $g(x_k, U_k)$ is using a minibatch estimate of the gradient when $\pi(w) \propto \varphi(w)\exp(-\beta n L_n(w))$. In this case

$$\nabla \log \pi(w) = \log \varphi(w) - n\beta \nabla L_n(w) \approx \nabla \log \varphi(w) - \beta \frac{n}{m} \sum_{i=1}^{m} \nabla \log p(Y_{u_i} \mid w) =: g(w, U)$$

where $U = (u_1, \ldots, u_m)$ are random indices into the dataset.

SGLD was first proposed in [WT11] and its properties are analysed in [TTV15; VZT15]. Here we focus on [TTV15], which gives conditions for convergence to $\pi(x)$ in the situation where $\epsilon_k \downarrow 0$. In [VZT15] the behavior at finite k for step size is considered under the same conditions.

Assumption 1 ([TTV15, Assumption 1])**.** We assume that the sequence of step sizes $(\epsilon_k)_{k\geq 0}$ is decreasing and $\epsilon_k \to 0$ as $k \to \infty$ (i.e. $\epsilon_k \downarrow 0$) and that $T_k \to \infty$ as $k \to \infty$, where $T_k = \sum_{i=0}^{k} \epsilon_i$.

Assumption 2 ([TTV15, Assumption 4])**.** We assume that the distribution $\pi(x)$ and the unbiased gradient estimator $g(x, U)$ jointly satisfy the following. We assume that there exists a twice-differentiable function $V : \mathbb{R}^d \to [1, \infty)$ such that $V(x) \to \infty$ as $\|x\| \to \infty$ and $V(x)$ has bounded second derivatives.

(1) There exist $p_H \geq 2$ and $C_1 > 0$ such that

$$\mathbf{E}\left[\|g(x, U) - \nabla \log \pi(x)\|^{2p_H}\right] \leq C_1 V(\theta)^{p_H} \qquad \forall x \in \mathbb{R}^d$$

(2) There exists $C_2 > 0$

$$\|\nabla V(x)\|^2 + \|\nabla \log \pi(x)\|^2 \leq C_2 V(x) \qquad \forall x \in \mathbb{R}^d$$

(3) There exist $a, b > 0$ such that

$$\tfrac{1}{2}\nabla V(x) \cdot \nabla \log \pi(x) \leq -aV(x) + b \qquad \forall x \in \mathbb{R}^d$$

Theorem 6 ([TTV15, Theorem 7]). *Suppose that Assumption 1 and Assumption 2 hold. Define*

$$\pi_k(f) = \frac{1}{T_k}\sum_{i=1}^{k} \epsilon_i f(x_i) \qquad \text{for } f : \mathbb{R}^d \to \mathbb{R}$$

where $T_k = \sum_{i=1}^{k}$ *and* $(x_k)_{k\geq 0}$ *is sampling chain generated by SGLD* (2.4) *(note that this defines a probability distribution). Then* π_k *converges in distribution to* π *almost surely.*

Remark. On Theorem 6:

- Convergence in distribution tells us that $\pi_k(f) \to \pi(f)$ as $k \to \infty$ only for *bounded* functions. In [TTV15] this is shown to hold for a larger class of functions determined by the exponent p_H and function V in Assumption 2.
- We say that π_k converges in distribution to π *almost surely* because the sampling chain depends on the random variables $(U_k)_{k\geq 0}$ and $(\eta_k)_{k\geq 0}$.
- [CDC15] proves essentially the same result using a different method which applies to a wider range of samplers. They explain in [CDC15, Appendix 1] that these assumptions entail Assumption 2.

2.2.1 Issues with singular models

For singular models there is a problematic interaction between the fact that V must have bounded second derivatives and condition (2) in Assumption 2. We give a negative result for deep linear networks in Lemma 8, and a similar result can be shown for class of distributions discussed in Example 5.

Lemma 7. *Suppose that* $V : R^d \to \mathbb{R}$ *has bounded second derivatives. Then there exists a polynomial function* $P : \mathbb{R}^d \to \mathbb{R}$ *of at most degree two such that* $V(x) \leq P(x)$ *for all* $x \in \mathbb{R}^d$.

Proof. We prove this in the case of $d = 1$, with the higher dimensional case being identical. There exists $A \in \mathbb{R}$ such that

$$V''(x) \leq A$$

for all $x \in \mathbb{R}$. Integrating both sides twice gives us $V''(x) \leq \frac{1}{2}Ax^2 + Bx + C$. □

Lemma 8. *Let* $F(y, w)$ *be a deep linear network, where* y *is the input variable and* w *is weights. Let* $Y_1, \ldots, Y_n$ *be a dataset drawn iid from a distribution* $q(y)$. *We consider the distribution*

$$\pi(w) \propto \exp(-nL(w)) \qquad \text{where} \qquad L(w) = \int \|F(y, w) - F(y, w_0)\|^2 q(y)dy$$

for some fixed weight w_0. *Suppose that Assumption 2 holds of* $\pi(w)$. *Then* $F(y, w)$ *has at most two layers.*

Proof. Let $P = \prod_{\ell=1}^{L} P_\ell$ be the deep linear network, where $P_1, \ldots, P_L$ are compatible matrices. That is, we take $w = (P_1, \ldots, P_L)$ and $F(y, w) = Py$ for all $y \in \mathbb{R}^N$.

We treat each parameter in weight space as a different polynomial variable. That is, we have $P_\ell = (w_{ij}^\ell)$ where w_{ij}^ℓ are distinct polynomial variables. For a matrix Q by $\deg(Q)$ we mean to take the maximum degree of its entries. Then we have $\deg(P) = L$ since each entry is a sum of monomials of the form $\prod_{\ell=1}^{L} w_{i_\ell j_\ell}^{(\ell)}$. This implies that (assuming absolute continuity of $q(y)$)

$$L(w) = \int \|Py - P^{(0)}y\|^2 q(y)dy$$

has degree $\deg L(w) = 2L$, where $P^{(0)}$ is the true parameter matrix. It follows that $\deg \|\nabla L(w)\|^2 = 2L - 2$.

From (2) in Assumption 2 and Lemma 7 we have

$$\|\nabla L(w)\|^2 \leq h(w)$$

where $h(w)$ is a degree two polynomial. This implies that $\deg \|\nabla L(w)\| \leq 2$ and so $L \leq 2$. □

3 Discussion

So where does this leave us? On one hand, the assumptions needed to prove that SGLD can sample from a distribution $\pi(x)$ using the method in [TTV15], and [RT96] tells us we shouldn't expect SGLD to be well-behaved for singular distributions (or at least for distributions with light tails, which coincide with singular distributions in certain simple classes). On the other, we have empirical evidence [Lau+24] that SGLD *can* sample well enough to accurately estimate the local learning coefficient of deep linear networks.

Option 1: the tails are actually not too light

The tails of distributions arising from deep linear networks and neural networks are not too light and Example 5 does not apply. In light of Lemma 8 this still leaves open the question of why SGLD works in deep linear networks, as we show that Assumption 2 cannot apply to non-trivial deep linear networks. Given that the loss function of the deep linear network discussed in Lemma 8 is a polynomial, this seems implausible.

Option 2: gradient noise stabilizes SGLD

Consider the gradient noise $\mathcal{E}(x) = g(x, U) - \nabla \log \pi(x)$. There are experiments [Şim+19; SSG19; BL23] which seem to suggest the distribution of $\mathcal{E}(x)$ is heavy-tailed. These works assume a *symmetric* heavy tailed distribution but don't seem to investigate the claim of symmetry.

We know that $\mathbf{E}\mathcal{E}(x) = 0$, but this does not rule out that $\mathcal{E}(x)$ is asymmetric in the sense of heavy negatively-skewed tail (in each coordinate). That is, extreme *under*estimates of the gradient are 'more likely than normal'. Going further, if $\nabla \log \pi(x)$ being large *causes* $\mathcal{E}(x)$ to be more likely to produce an extreme negative value then this could help stabilize SGLD.

Option 3: SGLD chains *do* diverge (eventually)

Every SGLD chain will eventually diverge almost surely, but at a finite number of steps we can still get useful approximations to the posterior distribution. Perhaps singularities cause a kind of metastability in the SGLD chain which means expected divergence time is very large. Probabilistically, the correct tool to analyse the finite-time behavior of such 'metastable' Markov processes would seem to be *quasi-stationary distributions* [CMS13].

3.1 Addendum

At the end of writing this talk I found several papers [RRT17; Xu+20; TLR18; ZLC18] coming at SGLD though the lens of statistical optimization. They are concerned with proving that SGLD finds a global minimum of empirical or population risk, or proving it escapes from local minima. Without regularity assumptions, it is a little difficult to see how this is compatible with Example 5. I have only read the papers briefly, but from what I can see:

- [TLR18] explicitly assumes a non-degenerate global minimum or a Morse condition (see the start of Section 2.1 and Theorem 3).
- [ZLC18] assumes the objective function or its gradients are bounded, but it looks like this analysis is concerned with escape from a local neighborhood. This may be compatible with Example 5.
- I'm unsure exactly what (if any) regularity assumptions are made in [RRT17; Xu+20], though in these papers the spectral gap of the generator (Definition 3) is non-zero. I believe this constitutes a kind of regularity condition, though I am not sure what justification they provide for this.

References

[BL23] Barak Battash and Ofir Lindenbaum. 'Revisiting the Noise Model of Stochastic Gradient Descent'. arXiv:2303.02749 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2303.02749 (visited on 27/11/2024).

[CDC15] Changyou Chen, Nan Ding and Lawrence Carin. 'On the convergence of stochastic gradient MCMC algorithms with high-order integrators'. In: *Advances in neural information processing systems* 28 (2015).

[CMS13] Pierre Collet, Servet Martínez and Jaime San Martín. *Quasi-Stationary Distributions: Markov Chains, Diffusions and Dynamical Systems*. en. Probability and Its Applications. Springer Berlin Heidelberg, 2013. URL: https://link.springer.com/10.1007/978-3-642-33131-2 (visited on 20/11/2024).

[EK05] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes: characterization and convergence*. eng. Wiley series in probability and mathematical statistics. Wiley-Interscience, 2005.

[Hit23] Rohan Hitchcock. 'Stochastic integration and stochastic differential equations'. 2023. URL: https://rohanhitchcock.com/stat-mech/stochastic-integration-and-sdes.pdf.

[IW11] Nobuyuki Ikeda and Shinzō Watanabe. *Stochastic differential Equations and diffusion processes*. eng. 2. ed., transferred to digital print. North-Holland mathematical Library 24. North-Holland [u.a.], 2011.

[Lau+24] Edmund Lau et al. 'The Local Learning Coefficient: A Singularity-Aware Complexity Measure'. arXiv:2308.12108 [stat]. Sept. 2024. URL: http://arxiv.org/abs/2308.12108 (visited on 22/10/2024).

[MT09] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. 2nd ed. Communications and control engineering series. Cambridge University Press, 2009.

[Øks13] Bernt Øksendal. *Stochastic differential equations: an introduction with applications*. 6th ed. Universitext. OCLC: ocn166267310. Springer, 2013.

[Par81] G. Parisi. 'Correlation functions and computer simulations'. en. In: *Nuclear Physics B* 180.3 (May 1981), pp. 378–384. URL: https://linkinghub.elsevier.com/retrieve/pii/0550321381900560 (visited on 26/11/2024).

[RRT17] Maxim Raginsky, Alexander Rakhlin and Matus Telgarsky. 'Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis'. arXiv:1702.03849 [cs]. June 2017. URL: http://arxiv.org/abs/1702.03849 (visited on 27/11/2024).

[RT96] Gareth O. Roberts and Richard L. Tweedie. 'Exponential convergence of Langevin distributions and their discrete approximations'. In: *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 2.4 (1996). Publisher: Bernoulli Society for Mathematical Statistics and Probability, pp. 341–363.

[RW00a] L Chris G Rogers and David Williams. *Diffusions, Markov processes, and martingales: Itô calculus*. 2nd ed. Vol. 2. Cambridge university press, 2000.

[RW00b] Leonard CG Rogers and David Williams. *Diffusions, markov processes, and martingales: Foundations*. 2nd ed. Vol. 1. Cambridge university press, 2000.

[Şim+19] Umut Şimşekli et al. 'On the Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks'. In: *arXiv:1912.00018 [cs, math, stat]* (Nov. 2019). arXiv: 1912.00018. URL: http://arxiv.org/abs/1912.00018 (visited on 01/04/2021).

[SSG19] Umut Simsekli, Levent Sagun and Mert Gurbuzbalaban. 'A tail-index analysis of stochastic gradient noise in deep neural networks'. In: *Proceedings of the 36th international conference on machine learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of machine learning research. PMLR, June 2019, pp. 5827–5837. URL: https://proceedings.mlr.press/v97/simsekli19a.html.

[TLR18] Belinda Tzen, Tengyuan Liang and Maxim Raginsky. 'Local Optimality and Generalization Guarantees for the Langevin Algorithm via Empirical Metastability'. arXiv:1802.06439 [cs]. June 2018. URL: http://arxiv.org/abs/1802.06439 (visited on 27/11/2024).

[TTV15] Yee Whye Teh, Alexandre Thiéry and Sebastian Vollmer. 'Consistency and fluctuations for stochastic gradient Langevin dynamics'. arXiv:1409.0578 [stat]. June 2015. URL: http://arxiv.org/abs/1409.0578 (visited on 14/08/2024).

[VZT15] Sebastian J. Vollmer, Konstantinos C. Zygalakis and and Yee Whye Teh. '(Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics'. arXiv:1501.00438 [stat]. Sept. 2015. URL: http://arxiv.org/abs/1501.00438 (visited on 24/10/2024).

[WT11] 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)*. 2011, pp. 681–688.

[Xu+20] Pan Xu et al. 'Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization'. arXiv:1707.06618 [stat]. Oct. 2020. URL: http://arxiv.org/abs/1707.06618 (visited on 27/11/2024).

[ZLC18] Yuchen Zhang, Percy Liang and Moses Charikar. 'A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics'. arXiv:1702.05575 [cs]. Apr. 2018. URL: http://arxiv.org/abs/1702.05575 (visited on 27/11/2024).