Uniform Trajectory Bounds for Heavy-Tailed Stochastic Gradient Langevin Dynamics

Rameez Raja[1]

[1]University of Warwick
Correspondence to rameez.raja@warwick.ac.uk

Abstract

Two recent lines of work have made complementary progress on the generalisation theory of noisy stochastic optimisers, but a critical gap remains at their intersection. Dupuis et al. (2024) proved the first uniform high-probability bounds over entire Stochastic Gradient Langevin Dynamics (SGLD) trajectories — guaranteeing that worst-case generalisation error over all iterates is controlled simultaneously — but only for Gaussian noise. Dupuis and Şimşekli (2024) subsequently proved high-probability bounds for heavy-tailed stochastic differential equations (SDEs) via the fractional Fokker–Planck equation, but their bounds are for continuous-time SDEs, grow with time horizon T, and are not uniform over discrete SGLD trajectories, as the authors themselves acknowledge. The combination — uniform trajectory bounds for heavy-tailed discrete SGLD — remains an open problem. We propose bridging this gap by extending the PAC-Bayesian random set framework of Dupuis et al. (2024) to α-stable Lévy noise. The central challenge is that the KL divergence used as the information-theoretic penalty in the existing framework is infinite for α-stable posteriors when $\alpha < 2$. We propose replacing it with Rényi divergences, which remain finite in the heavy-tailed regime and recover KL when $\alpha \to 2$. We present a three-step proof strategy, identify the key technical obstacles, and conjecture the form of the resulting bound — showing it reduces to the Gaussian result of Dupuis et al. (2024) in the limit $\alpha \to 2^-$.

1. Introduction

Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) injects Gaussian noise into gradient updates and is widely used for approximate Bayesian inference and landscape exploration in deep learning:

$$W_{k+1} = W_k - \eta_{k+1}\hat{g}_{k+1} + \sqrt{2\beta^{-1}\eta_{k+1}}\,\epsilon_{k+1}, \qquad \epsilon_{k+1} \sim \mathcal{N}(0, I_d). \tag{1}$$

A fundamental question is: does SGLD generalise throughout training, not just at the final iterate? Prior work bounded the generalisation error only at the last step (Mou et al., 2018; Farghly and Rebeschini, 2021). Dupuis et al. (2024) resolved this for Gaussian SGLD, proving the first uniform high-probability bound over entire trajectories:

$$\mathbb{E}_A\left[\max_{1\le k\le T}\big(R(W_k) - \widehat{R}_S(W_k)\big)\right] \lesssim \mathcal{O}\left(\sqrt{\frac{1}{n\sigma^2}\sum_{k=1}^{T}\eta_k\,\mathbb{E}_A\|\hat{g}_k\|^2}\right). \tag{2}$$

This guarantees that good generalisation holds regardless of when training is stopped — an important practical property, since stopping criteria are rarely theoretically justified.

Table 1. Positioning relative to prior work. Our proposal fills the open quadrant.

Work	Noise type	Uniform over trajectory?	Discrete SGLD?
Dupuis et al. (2024)	Gaussian	Yes	Yes
Dupuis and Şimşekli (2024)	α-stable	No	No
This work (proposed)	α-stable	Yes	Yes

The heavy-tailed gap. Empirical and theoretical work has shown that gradient noise in real neural network training is frequently heavy-tailed, not Gaussian, particularly under large step sizes or small batch sizes (Şimşekli et al., 2019; Gürbüzbalaban et al., 2021). Dupuis and Şimşekli (2024) addressed this by proving high-probability bounds for heavy-tailed SDEs via the fractional Fokker–Planck equation. This was a major advance: their bounds avoid non-computable mutual information terms present in earlier fractal-based approaches, and they identify a phase transition showing heavy tails can be beneficial or harmful depending on problem structure. However, their bounds have three important limitations which the authors themselves explicitly acknowledge. First, the bounds apply to continuous-time SDEs, not discrete SGLD iterates. Second, and critically, the bounds suffer from a time-dependence issue — they grow with the time horizon T and are not uniform over the trajectory. Third, they do not address whether generalisation is controlled simultaneously over all iterates.

The question of uniform trajectory bounds for heavy-tailed discrete SGLD therefore remains fully open. This is the gap we address.

Our contribution. We propose extending the PAC-Bayesian random set framework of Dupuis et al. (2024) to α-stable Lévy noise, filling the gap left by both prior works. Table 1 summarises how our proposal relates to each.

2. Background

PAC-Bayesian theory for random sets (Dupuis et al., 2024). Let $(E, \mathcal{E})$ be a measurable space of hypothesis sets, π a data-independent prior, and (ρ_S) a data-dependent posterior Markov kernel. For measurable $\Phi : E \times \mathcal{Z}^n \to \mathbb{R}$, the central bound of Dupuis et al. (2024) (Theorem 6) states: with probability at least $1 - \zeta$,

$$\mathbb{E}_{W\sim\rho_S}[\Phi(W,S)] \leq \mathrm{KL}(\rho_S\|\pi) + \log\frac{1}{\zeta} + \log \mathbb{E}_S\mathbb{E}_{W\sim\pi}\left[e^{\Phi(W,S)}\right]. \tag{3}$$

For SGLD, $W_S = \{W_1, \ldots, W_T\}$ is treated as a random closed set and $\Phi(W_S, S) = \sup_{w\in W_S}(R(w) - \widehat{R}_S(w))$ is the trajectory supremum. Controlling the information-theoretic (IT) term $\mathrm{KL}(\rho_S\|\pi)$ via the integrated gradient norm then yields (2).

Why (3) breaks for α-stable noise. For α-stable noise with $\alpha < 2$, the posterior ρ_S is an α-stable distribution. The KL divergence between two α-stable distributions with different parameters is infinite (Nolan, 2020) unless $\alpha = 2$. The IT penalty $\mathrm{KL}(\rho_S\|\pi)$ in (3) is therefore undefined, and the framework cannot be directly applied.

3. Proposed Extension

Setting: α-stable SGLD

We replace Gaussian noise in (1) with an α-stable Lévy increment:

$$W_{k+1} = W_k - \eta_{k+1}\hat{g}_{k+1} + \eta_{k+1}^{1/\alpha}\,\xi_{k+1}, \qquad \xi_{k+1} \sim \mathcal{S}_\alpha(\sigma, 0, 0), \tag{4}$$

where $\mathcal{S}_\alpha(\sigma, 0, 0)$ is the symmetric α-stable distribution with stability index $\alpha \in (1, 2)$ and scale $\sigma > 0$. Setting $\alpha = 2$ recovers (1).

Rényi Divergence as Replacement IT Term

The Rényi divergence of order $\beta > 0$, $\beta \neq 1$, is

$$D_\beta(P\|Q) = \frac{1}{\beta - 1} \log \int \left(\frac{dP}{dQ}\right)^\beta dQ.$$

Three properties make D_β the natural replacement for KL here. First, $\lim_{\beta \to 1} D_\beta(P\|Q) = \mathrm{KL}(P\|Q)$, so it strictly generalises KL. Second, since α-stable random variables have finite p-th moments only for $p < \alpha$ (Nolan, 2020), D_β is expected to remain finite for appropriately chosen $\beta = \beta(\alpha)$ depending on the tail index. Third, a Rényi-divergence analogue of (3) can be derived via the generalised Donsker–Varadhan variational formula (van Erven and Harremoës, 2014), providing the formal mechanism for Step 1 of our proof strategy below.

Conjectured Bound

Conjecture 1 (Informal). *Under appropriate regularity conditions on ℓ and the α-stable SGLD dynamics (4), with high probability:*

$$\mathbb{E}_A\left[\max_{1 \le k \le T}\left(R(W_k) - \widehat{R}_S(W_k)\right)\right] \lesssim \mathcal{O}\left(\left(\frac{1}{n}\sum_{k=1}^{T} \eta_k^{1+1/\alpha}\, \mathbb{E}_A\|\hat{g}_k\|^\alpha\right)^{1/\alpha}\right). \tag{5}$$

Remark 1. *The bound (5) has the same structural form as (2): a sum of step-size-weighted gradient norms, scaled by $1/n$, raised to a power. When $\alpha = 2$, the α-th moment $\mathbb{E}\|\hat{g}_k\|^\alpha$ reduces to the second moment $\mathbb{E}\|\hat{g}_k\|^2$, consistent with the Gaussian case. The step-size exponent $1 + 1/\alpha$ reduces to $3/2$ at $\alpha = 2$, compared to the exponent 1 in (2); resolving this discrepancy precisely is one of the open tasks in the proof. As $\alpha \to 1^+$, the sum grows faster (larger step-size exponents and heavier moment conditions), consistent with heavier tails being harder to control.*

4. Proof Strategy and Technical Obstacles

Step 1: Rényi PAC-Bayes for random sets. Extend (3) by replacing $\mathrm{KL}(\rho_S\|\pi)$ with $D_\beta(\rho_S\|\pi)$. The generalised Donsker–Varadhan formula (van Erven and Harremoës, 2014) provides the necessary mechanism. We expect this step to be relatively clean, conditional on the finiteness result in Step 2.

Step 2: Finiteness of $D_\beta(\rho_S\|\pi)$ for α-stable posteriors. (Core technical challenge.) The posterior ρ_S over the trajectory $\{W_1, \ldots, W_T\}$ is determined by the Lévy–Itô decomposition of the α-stable SGLD iterates. We must show $D_\beta(\rho_S\|\pi) < \infty$ for a suitable reference prior π and some $\beta = \beta(\alpha) > 1$. Preliminary analysis suggests $\beta < \alpha/(\alpha - 1)$ is sufficient, mirroring the classical result that α-stable random variables have finite p-th moments only for $p < \alpha$ (Nolan, 2020), which constrains the order at which the Rényi divergence remains finite.

Step 3: Controlling the log-exp term. The term $\log \mathbb{E}_S \mathbb{E}_{W\sim\pi}[e^{\Phi(W,S)}]$ in (3) requires concentration of the supremum generalisation error. In the Gaussian case, Hoeffding's lemma suffices because sub-Gaussian tails imply bounded moment generating functions. For α-stable trajectories, the absence of second moments requires instead sub-Weibull concentration inequalities (Kuchibhotla and Chakrabortty, 2022), which provide analogous exponential-type bounds under only α-th moment conditions.

5. Discussion

Why uniformity matters and how we achieve it. The time-dependence issue in Dupuis and Şimşekli (2024) arises because the fractional Fokker–Planck approach analyses the distribution at a fixed time T, giving a bound that necessarily grows with T. The PAC-Bayesian random set approach is structurally different: the entire trajectory $\{W_1, \ldots, W_T\}$ is treated as a single random object, so the resulting bound is uniform over all iterates by construction. This is not merely a technical refinement — it directly answers the question of whether generalisation is controlled throughout training, not just at a chosen stopping point.

Limitations and next steps. This work is a theoretical proposal. The three open tasks are: (1) formally proving $D_\beta(\rho_S \| \pi) < \infty$ for α-stable SGLD posteriors; (2) establishing the Rényi PAC-Bayes bound for random sets; (3) completing Step 3 via sub-Weibull concentration. Empirical validation of the α-dependence on synthetic datasets is planned as a companion study.

References

Benjamin Dupuis and Umut Şimşekli. Generalization bounds for heavy-tailed SDEs through the fractional Fokker–Planck equation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 12087–12137, 2024.

Benjamin Dupuis, Paul Viallard, George Deligiannidis, and Umut Şimşekli. Uniform generalization bounds on data-dependent hypothesis sets via PAC-Bayesian theory on random sets. *Journal of Machine Learning Research*, 25:1–55, 2024.

Tyler Farghly and Patrick Rebeschini. Time-independent generalization bounds for SGLD in non-convex settings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Manuel Gil, Fady Alajaji, and Tamás Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, 2021.

Arun K. Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory (COLT)*, 2018.

John P. Nolan. *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer, 2020.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011.