Global Dynamics of Heavy-Tailed SGDs in Nonconvex Loss Landscape: Characterization and Control

Xingyu Wang and Chang-Han Rhee

Industrial Engineering and Management Sciences, Northwestern University
Evanston, IL, 60613, USA

July 1, 2024

Abstract

Stochastic gradient descent (SGD) and its variants in deep neural networks (DNNs) are fundamental to the advancements of modern artificial intelligence. However, our theoretical understanding lags far behind their empirical success. It is widely believed that SGD has a curious ability to avoid sharp local minima in the loss landscape, whereas sharp minima are believed to lead to poor generalization. To unravel this mystery and further enhance such capability of SGDs, it is imperative to go beyond the traditional local convergence analysis and obtain a comprehensive understanding of SGDs' global dynamics. In this paper, we develop a set of technical machinery based on the recent large deviations and metastability analysis in [34] and obtain sharp characterization of the global dynamics of heavy-tailed SGDs. In particular, we reveal a fascinating phenomenon in deep learning: by injecting and then truncating heavy-tailed noises during the training phase, SGD can almost completely avoid sharp minima and hence achieve better generalization performance for the test data. Simulation and deep learning experiments confirm our theoretical prediction that heavy-tailed SGD with gradient clipping finds local minima with a more flat geometry and achieves better generalization performance.

Contents

1 Introduction

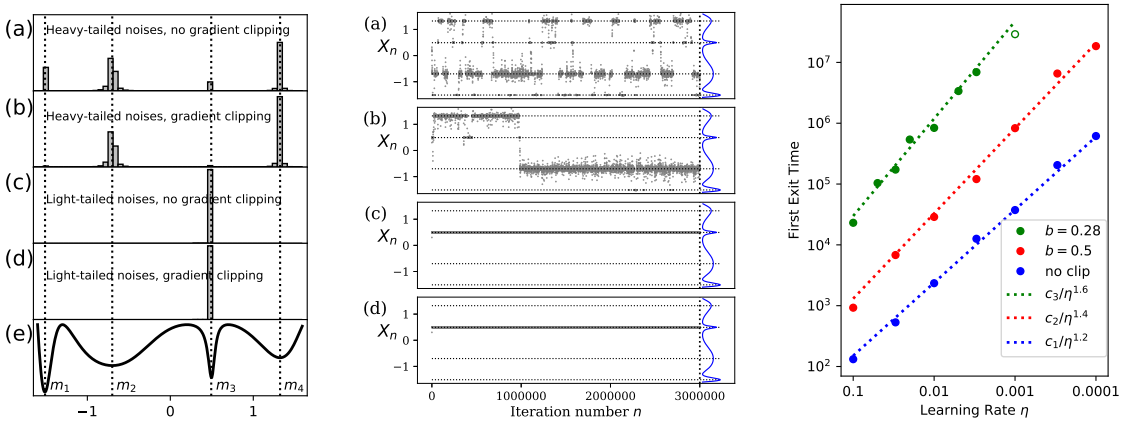

Figure 1.1: **(Left)** Histograms of the locations visited by SGD. With truncated heavy-tailed noises, SGD hardly ever visits the two sharp minima m_1 and m_3. The objective function f is plotted at the bottom, and dashed lines are added as references for the locations of local minima. **(Middle)** Typical trajectories of SGD in different cases: (a) Heavy-tailed noises, no gradient clipping; (b) Heavy-tailed noises, gradient clipping at $b = 0.5$; (c) Light-tailed noises, no gradient clipping; (d) Light-tailed noises, gradient clipping at $b = 0.5$. The objective function f is plotted at the right of each figure, and dashed lines are added as references for locations of the local minima. **(Right)** First Exit Time from $\Omega_2 = (-1.3, 0.2)$. Each dot represents the average of 20 samples of first exit time. Each dahsed line shows a polynomial function c_i/η^{β} where β is predicted by Theorem 2 and c_i is chosen to fit the dots. The non-solid green dot indicates that for some of the 20 samples of the termination threshold 5×10^7 was reached, and hence, it is an underestimation. Results in **(Left)** and **(Middle)** are obtained under learning rate $\eta = 0.001$ and initial value $X_0 = 0.3$.

Stochastic gradient descent (SGD) and its variants have seen unprecedented empirical successes in the training of deep neural networks. Specifically, the training of deep neural networks is typically posed as a non-convex optimization problem, and even without explicit regularization the solutions obtained by SGD often perform surprisingly well on test data. Such an unexpected generalization performance of SGD in deep neural networks are often attributed to SGD's ability to avoid *sharp local minima* in the loss landscape, which tends to lead to poor generalization [8, 15, 18, 13]. Despite significant efforts to explain such phenomena theoretically, understanding how SGD manages to avoid sharp local minima and end up with flat local minima within a realistic training time still remains as a central mystery of deep learning. Recently, the heavy-tailed dynamics of SGD received significant

attention, and it was suggested that the heavy tails in the stochastic gradients may be a key ingredient that facilitates SGD's escape from sharp local minima: for example, [30, 31] report the empirical evidence of heavy-tails in stochastic gradient noise in popular deep learning architectures (see also [9, 32, 4]) and show that SGD can escape sharp local minima in polynomial time under the presence of the heavy-tailed gradient noise. To be more specific, they view heavy-tailed SGDs as discrete approximations of Lévy driven Langevin equations and argue that the amount of time SGD trajectory spends in each local minimum is proportional to the width of the associated minimum according to the metastability theory [27, 11, 12] for such heavy-tailed processes.

In this paper, we study the global dynamics and long-run behavior of heavy-tailed SGD and its practical variant in depth. In particular, we consider an adaptive version of SGD, where the stochastic gradient is truncated above a fixed threshold. Such truncation scheme is often called *gradient clipping* and employed as default in various contexts [3, 20, 6, 26, 36, 5]. We uncover a rich mathematical structure in the global dynamics of SGD under this scheme and prove that the asymptotic behavior of such SGD is fundamentally different from that of the pure form of SGD: in particular, under a suitable structural condition on the geometry of the loss landscape, gradient clipping *completely eliminates sharp minima from the trajectory of SGDs*. This intriguing phenomenon leads to a new training strategy in deep learning for finding local minima that achieve better generalization performance. More precisely, the main contributions of this article can be summarized as follows.

- **Theoretical Contributions: Characterization of Global Dynamics.** We establish a scaling limit of the heavy-tailed dynamical systems over a multi-well potential in $\mathbb{R}^1$ at the process level. The scaling limit is a Markov jump process whose state space consists of the local minima of the potential. In particular, our findings systematically characterize a curious phenomenon that the truncated heavy-tailed processes avoid narrow local minima altoghether in the limit. As a direct application, we prove an ergodic theorem, which shows that the fraction of time such processes spend in the narrow attraction field converges to zero as the step-size tends to zero.

- **Algorithmic Contributions: Control of SGDs using Truncated Heavy Tails.** Inspired by the sharp characterization of the global behavior of heavy-tailed dynamical systems in $\mathbb{R}^1$, we propose a new training strategy in deep learning that improves the generalization performance of SGD. Specifically, by injecting and then truncating heavy-tailed noise in SGD, the new training strategy manages to find local minima with a more flat geometry and better generalization performance. We test the proposed algorithm with deep learning tasks and demonstrate its superiority with an ablation study. This also suggests that the key phenomenon we characterize in our theory— elimination of sharp local minima—manifests in real-world tasks.

Throughout this paper, we focus on the class of heavy tails captured by the notion of regular variation. Let $(Z_i)_{i\geq 1}$ be a sequence of iid random variables such that $\mathbf{E}Z_1 = 0$ and $\mathbf{P}(|Z_1| > x)$ is regularly varying with index $-\alpha$ as $x \to \infty$ for some $\alpha > 1$. That is, there exists some slowly varying function ϕ such that $\mathbf{P}(|Z_1| > x) = \phi(x)x^{-\alpha}$. Let $\varphi_b(\cdot) : x \mapsto \frac{x}{|x|}\max\{b, |x|\}$ be the projection operator from $\mathbb{R}$ onto $[-b, b]$, where $b > 0$ is a truncation threshold. For any $\eta > 0$, $x \in \mathbb{R}$, and $b \in (0, \infty)$, we define $\big(X_j^{\eta|b}(x)\big)_{j\geq 0}$ with the following recursion:

$$X_0^{\eta|b}(x) = x; \qquad X_{j+1}^{\eta|b}(x) = X_j^{\eta|b}(x) + \varphi_b\Big(- \eta U'\big(X_j^{\eta|b}(x)\big) + \eta\sigma\big(X_j^{\eta|b}(x)\big)Z_{j+1}\Big) \quad \forall j \geq 0. \tag{1.1}$$

In other words, $\big(X_j^{\eta|b}(x)\big)_{j\geq 0}$ solves a class of stochastic difference equations driven by truncated heavy-tailed perturbations. Here, $U'(\cdot)$ is the gradient field of some potential function $U \in \mathcal{C}^1(\mathbb{R})$, x dictates the initial condition of the stochastic difference equation, η is interpreted as the step-length, and the distance traveled at each step is truncated under the threshold level b. Generalizating to the scenario where the truncation threshold b is set as ∞ (i.e., when we remove the truncation mechanism),

we define $X_j^{\eta|\infty}(x) = X_j^{\eta}(x)$ as the solution of the following stochastic difference equation

$$X_0^{\eta}(x) = x; \qquad X_{j+1}^{\eta}(x) = X_j^{\eta}(x) - \eta U'\big(X_j^{\eta}(x)\big) + \eta\sigma\big(X_j^{\eta}(x)\big)Z_{j+1} \quad \forall j \geq 0. \tag{1.2}$$

Below, we describe the main contributions in more detail.

Characterization of the Global Dynamics of Heavy-Tailed Systems: Our first contribution is to provide a sharp characterization of the global behavior of $X_j^{\eta|b}(x)$'s when traversing a multi-well potential U; see Figure 3.1 for an illustration of a potential U and its attraction fields. Under suitable conditions, Theorem 3.2 establishes that the stochastic process $X_j^{\eta|b}(x)$ converges to a Markov jump process that visits only the widest local minima with proper scaling. By considering an arbitrarily large truncation threshold $b \approx \infty$, we also recover the sample-path convergence of the untruncated dynamics $X_j^{\eta}(x)$ in Theorem 3.3. The modes of convergence are in finite dimensional distributions and weakly w.r.t. the L^p norm in $\mathbb{D}[0,\infty)$. See Section 3.2 for precise definitions and statements.

As a consequence of the sharp characterization of the global dynamics in Theorem 3.2, we also obtain Corollary 3.4, which proves an ergodic theorem for the fraction of the times $X_j^{\eta|b}(x)$ spends in narrow attraction fields: roughly speaking,

$$\frac{1}{T\cdot\lambda_b^*(\eta)}\sum_{i=1}^{T\cdot\lambda_b^*(\eta)} \mathbb{I}\Big\{X_i^{\eta|b}(x) \in \bigcup_{j:\ m_j\in \text{ wide minima}} (m_j-\epsilon, m_j+\epsilon)\Big\} \xrightarrow{p} 1 \qquad \text{as } \eta\downarrow 0$$

where $\lambda_b^*(\eta)$ is a scaling function of η that is regularly varying with index $\mathcal{J}_b^*(V)\cdot(\alpha-1)+1$. Here $\mathcal{J}_b^*(V)$ is the maximal relative width of U's attraction fields. This uncovers an intriguing phenomenon: combined with truncation, the heavy-tailed processes will avoid any local minimum of U that is not the widest. The precise definitions of the widest attraction fields and the associated local minima are given in Section 3, but here we note that the width is measured by the number of jumps (with sizes bounded by b) required to exit the attraction field.

Figure 1.1 (Left, Middle) clearly illustrates these points with the histograms of the sample trajectories of SGDs. Note first that SGDs with light-tailed gradient noise—(c) and (d) of Figure 1.1 (Left, Middle)—never manages to escape a (sharp) minimum regardless of gradient clipping. In contrast, SGDs with heavy-tailed gradient noise—(a) and (b) of Figure 1.1 (Left, Middle)—easily escapes from local minima. Moreover, there is a clear difference between SGDs with gradient clipping and without gradient clipping. In (a) of Figure 1.1 (Left), SGD without gradient clipping spends a significant amount of time at each of all four local minima ($\{m_1, m_2, m_3, m_4\}$), although it spends more time around the wide ones ($\{m_2, m_4\}$) than the sharp ones ($\{m_1, m_3\}$). On the other hand, in (b) of Figure 1.1 (Left), SGD with gradient clipping not only escapes from local minima but also avoids sharp minima ($\{m_1, m_3\}$) almost completely. This means that after we run SGD for long enough (more precisely, the required run length $\lambda_b^*(\eta)$ that is roughly of polynomial order; see Section 3.2 for details), it is almost guaranteed that it won't be at a sharp minimum, effectively eliminating sharp minima from its training trajectories.

The theoretical developments in our work hinge on the technical framework in [34] that connects the large deviation and metastability analysis for heavy-tailed dynamical systems. Specifically,

- [34] establishes a new formulation of heavy-tailed large deviations that is locally uniform with respect to the initial values. These results characterize the *catastrophe principle* that reveals a discrete hierarchy governing the causes and probabilities of a wide variety of rare events associated with heavy-tailed stochastic difference/differential equations. Moreover, this locally uniform formulation of large deviations proves to be the right tool for the analysis of local stability in [34] and the characterization of global dynamics in our work.

- In terms of local stability, [34] obtains sharp asymptotics of the joint law of the (scaled) exit-times and exit-locations for heavy-tailed dynamical systems. In particular, the results reveal how the

local stability of $X_j^{\eta|b}$ within an attraction field varies with the truncation threshold b in (1.1). Building on such exit-time and exit-location analyses, we establish a scaling limit of the heavy-tailed dynamical systems over a multi-well potential at the process level. In particular, a key step in our work is the development of a technical framework that elevates the characterization of local behaviors to the global dynamics of $X_j^{\eta|b}$ over a multi-well potential U.

Section 2 provides a review of the results in [34] that are most relevant to our work, and Section 3 presents our theoretical analysis of the global dynamics of heavy-tailed SGDs.

Control of Global Dynamics of SGDs using Truncated Heavy Tails: We also propose a novel computational strategy that takes advantage of our newly discovered global dynamics of the heavy-tailed SGD. While the evidence of heavy tails were reported in many deep learning tasks [31, 30, 4, 7, 9, 24, 19, 32, 36], there seem to be plenty of deep learning contexts where the stochastic gradient noises are light-tailed [25] as well. Guided by our new theory, we propose an algorithm that injects heavy-tails to SGD by inflating the tail distribution of the gradient noise and facilitating the discovery of a local minimum that generalizes better. Our experiments with image classification tasks, reported in Tables 5.1 and 5.2, illustrate that the tail-inflation strategy we propose here can indeed improve the generalization performance of the SGD as predicted by our theory.

Some of the results of this paper have been presented in a preliminary form at a conference [33]. However, the current paper provides significant extension and generalization of results in [33]. For instance, we make much weaker assumptions, allowing for non-constant diffusion coefficient $\sigma(\cdot)$ and eliminating the need for regularity conditions such as $U \in \mathcal{C}^2(\mathbb{R})$ or the confinement of $X_j^{\eta|b}$ within a compact set. Besides, the global dynamics (i.e., sample path convergence results in Section 3.2) are characterized not only in terms of finite-dimensional distributions but also w.r.t. the L_p topology of the càdlàg space. Besides, compared to the brute force approach in [33], the current paper develops a systematic framework for sample path convergence w.r.t. the L_p topology of the càdlàg space and for elevating the local stability results to a characterization of the global dynamics.

The rest of the paper is organized as follows. Section 2 reviews results in [34] that are central to the developments in this work. Section 3 provides theoretical characterization of the global dynamics of the SGDs driven by heavy-tailed noises. Section 4 presents numerical experiments that confirm our theory. Section 5 proposes a new algorithm that artificially injects heavy tailed gradient noise in actual deep learning tasks and demonstrate the improved performance. In the Appendix, we collect the technical proofs for results in Section 3 and the details of the experiments presented in Sections 4 and 5.

2 Preliminaries

In this section, we introduce notations and review results that will be frequently used throughout this paper. Section 2.1 presents the sample-path large deviations for heavy-tailed stochastic difference equations $X_j^{\eta|b}(x)$'s defined in (2.1), and Section 2.2 discusses the local stability of $X_j^{\eta|b}(x)$. By applying these mathematical machineries in Section 3, we are able to obtain a tight characterization of the global dynamics of heavy-tailed SGDs over a multimodal potential (i.e., specializing to the case where $a(\cdot) = -U'(\cdot)$ for some multimodal function U).

First, we set frequently used notations. Let $\mathbb{Z}$ be the set of integers, $\mathbb{N} = \{1, 2, \cdots\}$ be the set of positive integers, and $\mathbb{Z}_+ = \{0, 1, 2, \cdots\}$ be the set of non-negative integers. Let $[n] = \{1, 2, \cdots, n\}$ for any positive integer n. Consider a metric space $(\mathbb{S}, \boldsymbol{d})$ with $\mathscr{S}_{\mathbb{S}}$ being the corresponding Borel σ-algebra. For any $E \subseteq \mathbb{S}$, let E° and E^- be the interior and closure of E, respectively. For any $\epsilon > 0$, let $E^r \triangleq \{y \in \mathbb{S} : \ \boldsymbol{d}(E, y) \le \epsilon\}$ be the ϵ-enlargement of E. Here, for any set $A \subseteq \mathbb{S}$ and any $x \in \mathbb{S}$, we define $\boldsymbol{d}(A, x) \triangleq \inf\{\boldsymbol{d}(y, x) : \ y \in A\}$. Let $E_\epsilon \triangleq ((E^c)^\epsilon)^c$ be the ϵ-shrinkage of E. It is

worth noticing that, for any set E, the enlargement E^r of E is closed, while the shrinkage E_r of E is open. We say that set $A \subseteq \mathbb{S}$ is bounded away from $B \subseteq \mathbb{S}$ under $\boldsymbol{d}$ if $\inf_{x\in A, y\in B} \boldsymbol{d}(x,y) > 0$.

Throughout this paper, we characterize heavy-tails using the notion of regular variation. For a measurable function $\phi : (0,\infty) \to (0,\infty)$, we say that ϕ is regularly varying as $x \to \infty$ with index β (denoted as $\phi(x) \in \mathcal{RV}_\beta(x)$ as $x \to \infty$) if $\lim_{x\to\infty} \phi(tx)/\phi(x) = t^\beta$ for all $t > 0$. For details of the properties of regularly varying functions, see, for example, Chapter 2 of [28]. We say that a measurable function $\phi(\eta)$ is regularly varying as $\eta \downarrow 0$ with index β if $\lim_{\eta\downarrow 0} \phi(t\eta)/\phi(\eta) = t^\beta$ for any $t > 0$. We denote this as $\phi(\eta) \in \mathcal{RV}_\beta(\eta)$ as $\eta \downarrow 0$.

2.1 Sample Path Large Deviations for Heavy-Tailed Dynamical Systems

In this section, we review the sample path large deviations developed in [34] for heavy-tailed dynamical systems. Specifically, we consider the following form of stochastic difference equations driven by heavy-tailed perturbations and potentially subject to truncation. Let

$$\varphi_c(w) \triangleq (w \wedge c) \vee (-c) \quad \forall w \in \mathbb{R},\ c > 0$$

be the projection operator onto $[-c, c]$. Given any $\eta > 0$, $b > 0$, and $x \in \mathbb{R}$, define $(X_j^{\eta|b}(x))_{j\geq 0}$ through the recursion

$$X_0^{\eta|b}(x) = x, \qquad X_j^{\eta|b}(x) = X_{j-1}^{\eta|b}(x) + \varphi_b\Big(\eta a\big(X_{j-1}^{\eta|b}(x)\big) + \eta\sigma\big(X_{j-1}^{\eta|b}(x)\big)Z_j\Big) \ \ \forall j \geq 1, \tag{2.1}$$

Here, Z_j's are iid copies of some random variable Z, and the coefficients $a(\cdot)$ and $\sigma(\cdot)$ are some functions. In case that $b = \infty$, as a convention we set $\varphi_\infty(w) = w$ as the identity mapping, and write $X_j^\eta(x) = X_j^{\eta|\infty}(x)$. In other words, $b = \infty$ corresponds to the untruncated case where the recursion degenerates to

$$X_0^\eta(x) = x; \qquad X_j^\eta(x) = X_{j-1}^\eta(x) + \eta a\big(X_{j-1}^\eta(x)\big) + \eta\sigma\big(X_{j-1}^\eta(x)\big)Z_j \ \ \forall j \geq 1. \tag{2.2}$$

Let

$$H^{(+)}(x) \triangleq \mathbf{P}(Z > x), \quad H^{(-)}(x) \triangleq \mathbf{P}(Z < -x), \quad H(x) \triangleq H^{(+)}(x) + H^{(-)}(x) = \mathbf{P}(|Z| > x). \tag{2.3}$$

We focus on the case where Z is heavy-tailed.

Assumption 1 (Regularly Varying Noises)**.** $\mathbf{E}Z = 0$. *Besides, there exist* $\alpha > 1$ *and* $p^{(+)}, p^{(-)} \in (0,1)$ *with* $p^{(+)} + p^{(-)} = 1$ *such that*

$$H(x) \in \mathcal{RV}_{-\alpha}(x) \ \ \text{as } x \to \infty; \quad \lim_{x\to\infty} \frac{H^{(+)}(x)}{H(x)} = p^{(+)}; \quad \lim_{x\to\infty} \frac{H^{(-)}(x)}{H(x)} = p^{(-)} = 1 - p^{(+)}.$$

Next, we introduce a few assumptions on $a : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$. It is worth noticing that, as we will see in Result 1 below, Assumption 4 can be safely dropped when $b \in (0,\infty)$ in (2.1), i.e., when truncation is in effect.

Assumption 2 (Lipschitz Continuity)**.** *There exists some* $D \in (0,\infty)$ *such that*

$$|\sigma(x) - \sigma(y)| \vee |a(x) - a(y)| \leq D|x - y| \quad \forall x, y \in \mathbb{R}.$$

Assumption 3 (Nondegeneracy)**.** $\sigma(x) > 0 \ \ \forall x \in \mathbb{R}$.

Assumption 4 (Boundedness)**.** *There exists some* $C \in (0,\infty)$ *such that*

$$|a(x)| \vee |\sigma(x)| \leq C \qquad \forall x \in \mathbb{R}.$$

Let $(\mathbb{D}[0,T], \boldsymbol{d}_{J_1}^{[0,T]})$ be a metric space, where $\mathbb{D}[0,T]$ is the space of all càdlàg functions on $[0,T]$ and $\boldsymbol{d}_{J_1}^{[0,T]}$ is the Skorodkhod J_1 metric

$$\boldsymbol{d}_{J_1}^{[0,T]}(x,y) \triangleq \inf_{\lambda\in\Lambda_T} \sup_{t\in[0,T]} |\lambda(t)-t| \vee |x(\lambda(t))-y(t)|.$$

Here, Λ_T is the set of homeomorphism on $[0,T]$. Given any $A \subseteq \mathbb{R}$, let $A^{k\uparrow} \triangleq \{(t_1,\cdots,t_k) \in A^k : t_1 < t_2 < \cdots < t_k\}$ be the set of ordered sequence of real numbers with length k on A. For any b, $T \in (0,\infty)$ and $k \in \mathbb{N}$, define the mapping $h_{[0,T]}^{(k)|b} : \mathbb{R}\times\mathbb{R}^k\times(0,T]^{k\uparrow} \to \mathbb{D}[0,T]$ as follows. Given $x_0 \in \mathbb{R}$, $\boldsymbol{w} = (w_1,\cdots,w_k) \in \mathbb{R}^k$, and $\boldsymbol{t} = (t_1,\cdots,t_k) \in (0,T]^{k\uparrow}$, let $\xi = h_{[0,T]}^{(k)|b}(x_0,\boldsymbol{w},\boldsymbol{t})$ solves

$$\xi_0 = x_0; \tag{2.4}$$

$$\frac{d\xi_s}{ds} = a(\xi_s) \quad \forall s \in [0,T],\ s \neq t_1,t_2,\cdots,t_k; \tag{2.5}$$

$$\xi_s = \xi_{s-} + \varphi_b\big(\sigma(\xi_{s-})\cdot w_j\big) \qquad \text{if } s = t_j \text{ for some } j \in [k] \tag{2.6}$$

In other words, $h_{[0,T]}^{(k)|b}(x_0,\boldsymbol{w},\boldsymbol{t})$ returns an ODE path perturbed by k jumps, where the size of each jump is modulated by $\sigma(\cdot)$ and truncated under b. For $k=0$, we adopt the convention that $\xi = h_{[0,T]}^{(0)|b}(x_0)$ is the solution to the ODE $d\xi_s/ds = a(\xi_s)$ $\forall s \in [0,T]$ under the initial condition $\xi_0 = x_0$. For any $T \in (0,\infty)$, $b \in (0,\infty]$, $A \subseteq \mathbb{R}$ and $k \in \mathbb{Z}_+$, let

$$\mathbb{D}_A^{(k)|b}[0,T] \triangleq h_{[0,T]}^{(k)|b}\big(A\times\mathbb{R}^k\times(0,T]^{k\uparrow}\big) \tag{2.7}$$

be the set of all ODE paths with k jumps, where the size of each jump is modulated by the diffusion coefficient $\sigma(\cdot)$ and then truncated under threshold b. We adopt the convention that $\mathbb{D}_A^{(-1)|b}[0,T] \triangleq \emptyset$. Given any $x \in \mathbb{R}$, $k \in \mathbb{Z}_+$, and $b, T \in (0,\infty)$, let

$$\mathbf{C}_{[0,T]}^{(k)|b}(\,\cdot\,;x) \triangleq \int \mathbf{I}\Big\{h_{[0,T]}^{(k)|b}(x,\boldsymbol{w},\boldsymbol{t}) \in\ \cdot\ \Big\}\nu_\alpha^k(d\boldsymbol{w})\times\mathcal{L}_T^{k\uparrow}(d\boldsymbol{t}). \tag{2.8}$$

Here, $\nu_\beta^k(\cdot)$ is the k-fold product measure of the (Borel) measure

$$\nu_\beta[x,\infty) = p^{(+)}x^{-\beta}, \quad \nu_\beta(-\infty,-x] = p^{(-)}x^{-\beta}, \quad \forall x>0,\ \beta>0, \tag{2.9}$$

$\mathcal{L}_t$ is the Lebesgue measure restricted on $(0,t)$, and $\mathcal{L}_t^{k\uparrow}$ is the Lebesgue measure restricted on $(0,t)^{k\uparrow}$. Let $\boldsymbol{X}_{[0,T]}^{\eta|b}(x) \triangleq \{X_{\lfloor t/\eta\rfloor}^{\eta|b}(x) :\ t\in[0,T]\}$ be the time-scaled version of $X_j^{\eta|b}(x)$ embedded in $\mathbb{D}[0,T]$. Similar notations are adopted for $h_{[0,T]}^{(k)} = h_{[0,T]}^{(k)|\infty}$, $\mathbb{D}_A^{(k)}[0,T] = \mathbb{D}_A^{(k)|\infty}[0,T]$, $\mathbf{C}_{[0,T]}^{(k)} = \mathbf{C}_{[0,T]}^{(k)|\infty}$, and $\boldsymbol{X}_{[0,T]}^{\eta}(x) = \boldsymbol{X}_{[0,T]}^{\eta|\infty}(x)$.

Let $\mathscr{S}_{\mathbb{D}[0,T]}$ be the Borel σ-algebra of $(\mathbb{D}[0,T], \boldsymbol{d}_{J_1}^{[0,T]})$. Let

$$\lambda(\eta) \triangleq \eta^{-1}H(\eta^{-1})$$

and $\lambda^k(\eta) = (\lambda(\eta))^k$. By Assumption 1, $\lambda(\eta) \in \mathcal{RV}_{\alpha-1}(\eta)$ as $\eta\downarrow 0$. As a summary of Theorems 2.3 and 2.4 in [34], Result 1 provides sharp asymptotics for rare events in $X_j^{\eta|b}(x)$ driven by heavy-tailed perturbations.

Result 1 (Sample Path Large Deviations)**.** *Let Assumptions 1, 2, and 3 hold. Let $k \in \mathbb{Z}_+$ and $T, b \in (0,\infty)$. For any $B \in \mathscr{S}_{\mathbb{D}[0,T]}$ that is bounded away from $\mathbb{D}_A^{(k-1)|b}[0,T]$ under $\boldsymbol{d}_{J_1}^{[0,T]}$,*

$$\begin{aligned}\inf_{x\in A}\mathbf{C}_{[0,T]}^{(k)|b}(B^\circ;x) &\le \liminf_{\eta\downarrow0}\frac{\inf_{x\in A}\mathbf{P}\big(\boldsymbol{X}_{[0,T]}^{\eta|b}(x)\in B\big)}{\lambda^k(\eta)}\\ &\le \limsup_{\eta\downarrow0}\frac{\sup_{x\in A}\mathbf{P}\big(\boldsymbol{X}_{[0,T]}^{\eta|b}(x)\in B\big)}{\lambda^k(\eta)} \le \sup_{x\in A}\mathbf{C}_{[0,T]}^{(k)|b}(B^-;x) < \infty.\end{aligned} \tag{2.10}$$

Furthermore, if Assumption 4 holds, then claim (2.10) is also valid for $b = \infty$.

To conclude this section, we add a few remarks regarding Result 1. This result is a manifestation of the catastrophe principle that governs the rare events in heavy-tailed systems: the catastrophic failures (i.e., extreme deviations from typical behaviors) in a small number of components lead to system-wide rare events. Specifically, the index k that leads to non-degenerate bounds in (2.10) corresponds to the minimum number of jumps that needs to be added to the ODE $\boldsymbol{y}_t(x)$ for it to enter the set B given $x \in A$, where

$$\boldsymbol{y}_0(x) = x, \qquad \frac{d\boldsymbol{y}_t(x)}{dt} = a\big(\boldsymbol{y}_t(x)\big) \ \ \forall t \geq 0. \tag{2.11}$$

Such an index k dictates the polynomial rate of decay of the probability of rare events and plays a role similar to the infimum of rate function of the classical large deviation principles. Furthermore, it dictates the most likely scenario of the rare events (i.e., they are almost always caused by exactly k large jumps in the system); see Corollary 2.5 of [34].

2.2 First Exit Analysis for Heavy-Tailed Dynamical Systems

Next, we review the results about the local stability of $X_j^{\eta}(x)$ and $X_j^{\eta|b}(x)$ in Section 2.3 of [34]. Specifically, let $I = (s_{\text{left}}, s_{\text{right}})$ be an open interval where $s_{\text{left}} < 0 < s_{\text{right}}$. Define

$$\tau^{\eta}(x) \triangleq \min\big\{j \geq 0 : \ X_j^{\eta}(x) \notin I\big\}, \qquad \tau^{\eta|b}(x) \triangleq \min\big\{j \geq 0 : \ X_j^{\eta|b}(x) \notin I\big\} \tag{2.12}$$

as the first exit times of $X_j^{\eta}(x)$ and $X_j^{\eta|b}(x)$ from $I = (s_{\text{left}}, s_{\text{right}})$, respectively. As will be demonstrated in Section 3, the global dynamics of heavy-tailed SGDs are understood by characterizing when and how we exit from different regions and traverse the loss landscape, which hinges on the first exit times and exit locations under different choices of I.

Specifically, in Section 2.2 we impose Assumption 5 on $a(\cdot)$. In case that $a(\cdot) = -U'(\cdot)$ for some $U \in \mathcal{C}^1(\mathbb{R})$, Assumption 5 dictates that, over the domain I, the potential function $U(\cdot)$ has a unique local minimum at $x = 0$. Moreover, since $U'(x)x = -a(x)x > 0$ for all $x \in I \setminus \{0\}$, the domain I is a subset of the attraction field of the origin, as $\lim_{t\to\infty} \boldsymbol{y}_t(x) = 0$ holds for all $x \in I$.

Assumption 5. *$a(0) = 0$. Besides, it holds for all $x \in I \setminus \{0\}$ that $a(x)x < 0$.*

To present the result, we introduce a few definitions. For any $k \in \mathbb{N}$ and $b \in (0, \infty)$, let the mapping $\check{g}^{(k)|b} : \mathbb{R} \times \mathbb{R}^k \times (0,\infty)^{k\uparrow} \to \mathbb{R}$ be defined as

$$\check{g}^{(k)|b}(x, \boldsymbol{w}, \boldsymbol{t}) \triangleq h^{(k)|b}_{[0,t_k+1]}(x, \boldsymbol{w}, \boldsymbol{t})(t_k) \tag{2.13}$$

where $\boldsymbol{t} = (t_1, \ldots, t_k) \in (0,\infty)^{k\uparrow}$, $\boldsymbol{w} = (w_1, \ldots, w_k) \in \mathbb{R}^k$, and $h^{(k)|b}_{[0,T]} : \mathbb{R} \times \mathbb{R}^k \times (0,T]^{k\uparrow} \to \mathbb{D}[0,T]$ is as defined in (2.4)–(2.6). As a convention, we set $\check{g}^{(0)|b}(x) \triangleq x$. Next, define Borel measures (for each $k \geq 1$ and $b \in (0,\infty)$)

$$\check{\mathbf{C}}^{(k)|b}(\ \cdot\ ; x) \triangleq \int \mathbf{I}\Big\{\check{g}^{(k-1)|b}\big(x + \varphi_b\big(\sigma(x)\cdot w_0\big), \boldsymbol{w}, \boldsymbol{t}\big) \in \ \cdot\ \Big\}\nu_\alpha(dw_0) \times \nu_\alpha^{k-1}(d\boldsymbol{w}) \times \mathcal{L}_\infty^{k-1\uparrow}(d\boldsymbol{t}) \tag{2.14}$$

with $\mathcal{L}_\infty^{k\uparrow}$ being the Lebesgue measure restricted on $\{(t_1, \cdots, t_k) \in (0,\infty)^k : \ 0 < t_1 < t_2 < \cdots < t_k\}$. Also, define

$$\check{\mathbf{C}}(\ \cdot\ ; x) \triangleq \int \mathbf{I}\Big\{x + \sigma(x)\cdot w \in \ \cdot\ \Big\}\nu_\alpha(dw). \tag{2.15}$$

In case that $x = 0$, we write $\check{\mathbf{C}}^{(k)|b}(\cdot) \triangleq \check{\mathbf{C}}^{(k)|b}(\ \cdot\ ; 0)$. and $\check{\mathbf{C}}(\cdot) \triangleq \check{\mathbf{C}}(\ \cdot\ ; 0)$. Also, let

$$l \triangleq \inf_{x \in I^c} |x| = |s_{\text{left}}| \wedge s_{\text{right}}, \qquad \mathcal{J}_b^* \triangleq \lceil l/b \rceil. \tag{2.16}$$

Intuitively speaking, l is the distance between the origin and I^c, and $\mathcal{J}_b^*$ is the smallest number of jumps required to exit from I if the size of each jump is bounded by b.

Recall that $H(\cdot) = \mathbf{P}(|Z| > \cdot)$ and $\lambda(\eta) = \eta^{-1}H(\eta^{-1})$. Result 2 provides sharp asymptotics for the joint distribution of the first exit times and exit locations of $X_j^{\eta|b}(x)$ and $X_j^{\eta}(x)$.

Result 2 (First Exit Times and Locations)**.** *Let Assumptions 1, 2, 3, and 5 hold.*

(a) Let $b > 0$ be such that $s_{left}/b \notin \mathbb{Z}$ and $s_{right}/b \notin \mathbb{Z}$. For any $\epsilon > 0$, $t \geq 0$, and measurable set $B \subseteq I^c$,

$$\limsup_{\eta\downarrow 0} \sup_{x\in I_\epsilon} \mathbf{P}\Big(C_b^*\eta \cdot \lambda^{\mathcal{J}_b^*}(\eta)\tau^{\eta|b}(x) > t;\ X^{\eta|b}_{\tau^{\eta|b}(x)}(x) \in B\Big) \leq \frac{\check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(B^-)}{C_b^*} \cdot \exp(-t),$$

$$\liminf_{\eta\downarrow 0} \inf_{x\in I_\epsilon} \mathbf{P}\Big(C_b^*\eta \cdot \lambda^{\mathcal{J}_b^*}(\eta)\tau^{\eta|b}(x) > t;\ X^{\eta|b}_{\tau^{\eta|b}(x)}(x) \in B\Big) \geq \frac{\check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(B^\circ)}{C_b^*} \cdot \exp(-t)$$

where $C_b^* \triangleq \check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(I^c) \in (0,\infty)$.

(b) For any $t \geq 0$ and measurable set $B \subseteq I^c$,

$$\limsup_{\eta\downarrow 0} \sup_{x\in I_\epsilon} \mathbf{P}\Big(C^*\eta \cdot \lambda(\eta)\tau^{\eta}(x) > t;\ X^{\eta}_{\tau^{\eta}(x)}(x) \in B\Big) \leq \frac{\check{\mathbf{C}}(B^-)}{C^*} \cdot \exp(-t),$$

$$\liminf_{\eta\downarrow 0} \inf_{x\in I_\epsilon} \mathbf{P}\Big(C^*\eta \cdot \lambda(\eta)\tau^{\eta}(x) > t;\ X^{\eta}_{\tau^{\eta}(x)}(x) \in B\Big) \geq \frac{\check{\mathbf{C}}(B^\circ)}{C^*} \cdot \exp(-t)$$

where $C^* \triangleq \check{\mathbf{C}}(I^c) \in (0,\infty)$.

To conclude, we stress that Result 2 is another intriguing manifestation of the catastrophe principle in heavy-tailed systems and reveals a discrete hierarchy in the first exit analysis of $X_j^{\eta|b}(x)$. In particular, $\mathcal{J}_b^*$ can be interpreted as the "discretized width" of domain I relative to the truncation threshold b, and Result 2 shows that $\tau^{\eta|b}(x)$ is roughly of order $\eta^{-1}\lambda^{-\mathcal{J}_b^*}(\eta)$, which is regularly varying in $1/\eta$ with index $1+\mathcal{J}_b^*(\alpha-1)$. In other words, the order of the exit times are dictated by the minimum number of jumps required for exit; the wider the domain I is (in terms of the relative width $\mathcal{J}_b^*$), the much longer it takes to exit from I, asymptotically. Therefore, in light of Result 2, it is natural to expect the following: over some multimodal potential function U, the truncated heavy-tailed SGDs will spend almost all the time around the widest local minima, thus avoiding all the narrow minima. This will be established rigorously in Section 3. In fact, thanks to the sharp characterization of both exit times and exit locations in Result 2, we are able to provide asymptotics for the entire sample path of heavy-tailed SGDs.

3 Global Dynamics of Heavy-Tailed SGDs

In this section, we consider the case where $a(\cdot) = -U'(\cdot)$ for some general multimodal potential function $U : \mathbb{R} \to \mathbb{R}$ and characterize the global behavior of $X_j^{\eta}(x)$ and $X_j^{\eta|b}(x)$. We show that, after proper scaling, their sample paths converge to those of Markov jump processes whose state spaces consist of local minima of U. Curiously, the state space of the limit process associated with $X_j^{\eta|b}(x)$ consists of only the widest local minima.

3.1 Problem Setting

We consider a multimodal potential function with local minima $\{m_1, m_2, \ldots, m_{n_{min}}\}$. More precisely, we make the following assumption throughout this section.

Assumption 6. *Let $U : \mathbb{R} \to \mathbb{R}$ be a function in $\mathcal{C}^1(\mathbb{R})$. Besides, there exist a positive integer $n_{min} \geq 2$ and an ordered sequence of real numbers $-\infty < m_1 < s_1 < m_2 < s_2 < \cdots < s_{n_{min}-1} < m_{n_{min}} < \infty$ such that (under the convention $s_0 = -\infty$ and $s_{n_{min}} = \infty$)*

(i) $U'(x) = 0$ iff $x \in \{m_1, s_1, \cdots, s_{n_{min}-1}, m_{n_{min}}\}$;

(ii) $U'(x) < 0$ for all $x \in \bigcup_{j\in[n_{min}]}(s_{j-1}, m_j)$;

(iii) $U'(x) > 0$ for all $x \in \bigcup_{j\in[n_{min}]}(m_j, s_j)$.

See Figure 3.1 (Left) for an illustration of such function U with $n_{\min} = 3$. Note that the local maxima $s_1, \cdots, s_{n_{\min}-1}$ divide $\mathbb{R}$ into different regions $I_i \triangleq (s_{i-1}, s_i)$ for $i = 0, \ldots, n_{min}$. Such regions can be viewed as the *attraction fields* of the local minima m_i's. That is, the ODE $\boldsymbol{y}_t(x)$ defined in (2.11) (with $a = -U'$) admits the limit $\lim_{t\to\infty} \boldsymbol{y}_t(x) = m_i$ if $x \in I_i$. Note that we impose the condition $n_{\min} \geq 2$ simply to avoid the trivial case of $n_{\min} = 1$, in which case there exists only one attraction field.

In order to present the main results, we introduce some definitions to facilitate the characterization of the geometry of U. First, for each attraction field I_i, let

$$l_i \triangleq \inf_{x\in I_i^c} |x - m_i| = |m_i - s_{i-1}| \wedge |s_i - m_i| \tag{3.1}$$

be the effective "width" of I_i, i.e., the minimum distance between m_i and the outside of I_i. Next, for any $i \in [n_{\min}]$ and $j \in [n_{\min}]$ with $j \neq i$, let

$$\mathcal{J}_b^*(i) \triangleq \lceil l_i/b \rceil, \qquad l_{i,j} \triangleq \inf_{x\in I_j} |x - m_i| = \begin{cases} s_{j-1} - m_i & \text{if } \ j > i \\ m_i - s_j & \text{if } \ j < i \end{cases}, \qquad \mathcal{J}_b^*(i,j) \triangleq \lceil l_{i,j}/b \rceil. \tag{3.2}$$

Recall that b is the truncation threshold for $X_j^{\eta|b}(x)$. Here, $\mathcal{J}_b^*(i)$ can be interpreted as the *discretized width* of I_i w.r.t. the *resolution* b, in the sense that it is the minimum number of jumps (with sizes bounded by b) required to escape from I_i when starting from m_i. Furthermore, $\mathcal{J}_b^*(i,j)$ is the minimal number of steps required to travel from m_i to I_j under the truncation threshold b. By definition, we must have $\mathcal{J}_b^*(i,j) \geq \mathcal{J}_b^*(i)$. With $\mathcal{J}_b^*(i)$ and $\mathcal{J}_b^*(i,j)$, we define the *typical transition graph* as follows.

Definition 3.1 (Typical Transition Graph)**.** *Given a function U satisfying Assumption 6 and some $b > 0$, the typical transition graph associated with threshold b is a directed graph $\mathcal{G}_b = (V, E_b)$ such that*

- *$V = \{m_1, \cdots, m_{n_{min}}\}$;*
- *An edge $(m_i \to m_j)$ is in E_b iff $\mathcal{J}_b^*(i,j) = \mathcal{J}_b^*(i)$.*

The graph $\mathcal{G}_b$ can be decomposed into different communication classes that are mutually exclusive. For $m_i, m_j \in V$ with $i \neq j$, we say that m_i and m_j communicate if and only if there exists a path $(m_i \to m_{k_1} \to \cdots \to m_{k_n} \to m_j)$ as well as a path $(m_j \to m_{k'_1} \to \cdots \to m_{k'_{n'}} \to m_i)$ on $\mathcal{G}_b$. In this section, we focus on the case where $\mathcal{G}_b$ is irreducible, i.e., all nodes communicate with each other on graph $\mathcal{G}_b$. See Figure 3.1 (Middle) and (Right) for the illustration of irreducible and reducible cases, respectively. Specifically, we impose the following assumption on the truncation threshold b. We note that the second condition is a mild one, as it holds for almost every $b > 0$ except for countably many.

Assumption 7. *$b \in (0, \infty)$ is such that*

- *$\mathcal{G}_b$ is irreducible,*
- *$|s_j - m_i|/b \notin \mathbb{Z}$ for all $i \in [n_{min}]$ and $j \in [n_{min} - 1]$.*

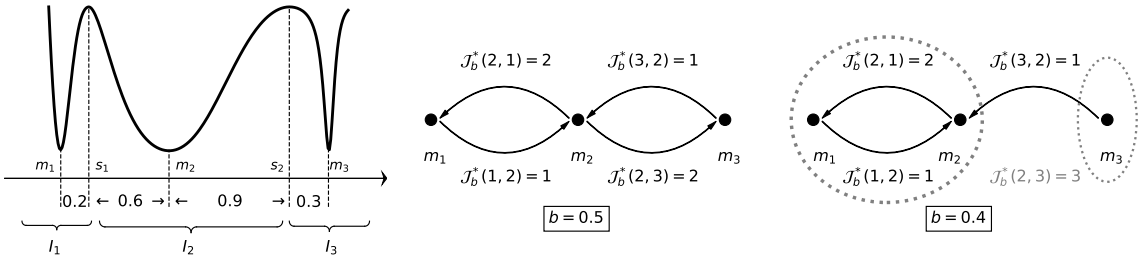

Figure 3.1: Typical transition graphs $\mathcal{G}_b$ associated with different gradient clipping thresholds b. **(Left)** The potential function U illustrated here has 3 attraction fields. For the second one $I_2 = (s_1, s_2)$, we have $s_2 - m_2 = 0.9, m_2 - s_1 = 0.6$. **(Middle)** The typical transition graph associated with $b = 0.5$. The entire graph $\mathcal{G}_b$ is irreducible since all nodes communicate with each other. **(Right)** The typical transition graph associated with $b = 0.4$. When $b = 0.4$, since $0.6 < 2b$ and $0.9 > 2b$, we have $\mathcal{J}_b^*(2,1) = 2$ and $\mathcal{J}_b^*(2,3) = 3$, and hence $\mathcal{J}_b^*(2) = 2 = \mathcal{J}_b^*(2,1) < \mathcal{J}_b^*(2,3)$. Therefore, the graph $\mathcal{G}_b$ does not contain the edge $m_2 \to m_3$ and there are two communication classes: $G_1 = \{m_1, m_2\}, G_2 = \{m_3\}$.

3.2 Sample Path Convergence

We are now ready to present the main result of this section. Theorem 3.2 establishes that, under a proper time scaling, the sample path of $X_j^{\eta|b}(x)$ converges to that of a Markov jump process, which only visits the widest local minima of U. Here, the width of each attraction field I_i is characterized by $\mathcal{J}_b^*(i)$ defined in (3.2). We use

$$\mathcal{J}_b^*(V) \triangleq \max_{i:\ m_i \in V} \mathcal{J}_b^*(i) \tag{3.3}$$

to denote the largest width—when discretized w.r.t. the resolution b—among all attraction fields. Next, define

$$V_b^* \triangleq \{m_i :\ \mathcal{J}_b^*(i) = \mathcal{J}_b^*(V)\} \tag{3.4}$$

as the set containing all the widest local minima.

In Theorem 3.2, the scaling limit of the sample path of $X_j^{\eta|b}(x)$ will be characterized in terms of the following two modes of convergence. First, we say that $\{S_t^\eta : t > 0\}$ converges to $\{S_t^* : t > 0\}$ *in finite-dimensional distributions* (f.d.d.) if we have $\left(S_{t_1}^\eta, \cdots, S_{t_k}^\eta\right) \Rightarrow \left(S_{t_1}^*, \cdots, S_{t_k}^*\right)$ as $\eta \downarrow 0$ for any $k \geq 1$ and $0 < t_1 < t_2 < \cdots < t_k < \infty$. We also denote this as $\{S_t^\eta : t > 0\} \overset{f.d.d.}{\to} \{S_t^* : t > 0\}$.

Remark 1. *We consider the convergence in f.d.d. only on $(0, \infty)$, thus excluding $t = 0$. This is because in Theorem 3.2, under the proper time scaling, the value of $X_{\lfloor t \rfloor}^{\eta|b}(x)$ for some t close to 0 will quickly converges to that of $Y_0^{*|b}$ (i.e., the initial value of the limit Markov jump process), but not exactly at $t = 0$. The same applies to Theorem 3.3.*

Next, we recall the convergence w.r.t. the L_p topology in $\mathbb{D}[0, \infty)$. For any $p \in [1, \infty)$ and $T \in (0, \infty)$, let

$$\boldsymbol{d}_{L_p}^{[0,T]}(x, y) \triangleq \left(\int_0^T |x_t - y_t|^p dt \right)^{1/p} \qquad \forall x, y \in \mathbb{D}[0, T] \tag{3.5}$$

be the L_p metric on $\mathbb{D}[0, T]$. For any $T > 0$, define the projection $\pi_T : \mathbb{D}[0, \infty) \to \mathbb{D}[0, T]$ such that

$$\pi_T(\xi)_t = \xi_t \qquad \forall t \in [0, T]. \tag{3.6}$$

Now, we define

$$\boldsymbol{d}_{L_p}^{[0,\infty)}(x,y) \triangleq \sum_{k\geq 1} \frac{1 \wedge \boldsymbol{d}_{L_p}^{[0,k]}\big(\pi_k(x),\pi_k(y)\big)}{2^k} \qquad \forall x,y \in \mathbb{D}[0,\infty) \tag{3.7}$$

and note that $\boldsymbol{d}_{L_p}^{[0,\infty)}$ is a metric on $\mathbb{D}[0,\infty)$. Hereafter in this paper, the continuity of a functional $f : \mathbb{D}[0,\infty) \to \mathbb{R}$ is understood w.r.t. the topology induced by $\boldsymbol{d}_{L_p}^{[0,\infty)}$. We say that the sequence of càdlàg processes $\{S_t^\eta : \ t \geq 0\}$ converges in distribution to $\{S_t^* : \ t \geq 0\}$ *w.r.t. the L_p topology in* $\mathbb{D}[0,\infty)$ as $\eta \downarrow 0$ if $\lim_{\eta\downarrow 0} \mathbf{E} f(S_\cdot^\eta) = \mathbf{E} f(S_\cdot^*)$ for all $f : \mathbb{D}[0,\infty) \to \mathbb{R}$ that is bounded and continuous. We denote this with $S_\cdot^\eta \Rightarrow S_\cdot^*$ in $(\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)})$ or $\{S_t^\eta : t \geq 0\} \Rightarrow \{S_t^* : t \geq 0\}$ in $(\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)})$.

Recall that $H(\cdot) = \mathbf{P}(|Z_1| > \cdot)$ and $\lambda(\eta) = \eta^{-1} H(\eta^{-1}) \in \mathcal{RV}_{\alpha-1}(\eta)$. Define a scaling function

$$\lambda_b^*(\eta) \triangleq \eta \cdot \big(\lambda(\eta)\big)^{\mathcal{J}_b^*(V)} \in \mathcal{RV}_{\mathcal{J}_b^*(V)\cdot(\alpha-1)+1}(\eta) \qquad \text{as } \eta \downarrow 0. \tag{3.8}$$

We are now ready to state the main result.

Theorem 3.2. *Let Assumptions 1, 2, 3, 6, and 7 hold. Let $p \in [1,\infty)$, $i \in [n_{min}]$, and $x \in I_i$. As $\eta \downarrow 0$,*

$$\{X^{\eta|b}_{\lfloor \cdot/\lambda_b^*(\eta)\rfloor}(x) : \ t > 0\} \overset{f.d.d.}{\to} \{Y_t^{*|b} : \ t > 0\} \quad \textit{and} \quad X^{\eta|b}_{\lfloor \cdot/\lambda_b^*(\eta)\rfloor}(x) \Rightarrow Y_\cdot^{*|b} \textit{ in } (\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)}),$$

where $Y_t^{*|b}$ is a continuous-time Markov chain with a finite state space V_b^*, initial distribution (see (3.11) for the definition of θ_b),

$$\mathbf{P}(Y_0^{*|b} = m_j) = \theta_b(m_j|m_i) \qquad \forall m_j \in V_b^*, \tag{3.9}$$

and infinitesimal generator (see (3.10) for the definition of q_b)

$$\begin{aligned} Q^{*|b}(i,j) &= \sum_{j' \in [n_{min}]:\ j' \neq i} q_b(i,j')\theta_b(m_j|m_{j'}) \qquad \forall m_i,\ m_j \in V_b^* \textit{ with } m_i \neq m_j, \\ Q^{*|b}(i,i) &= -\sum_{m_j \in V_b^*:\ j \neq i} Q^{*|b}(i,j) \qquad \forall m_i \in V_b^*. \end{aligned}$$

We provide the proof of Theorem 3.2 in Section B. Here, we specify the law of limiting Markov jump process $Y^{*|b}$. Recall the definition of $\check{\mathbf{C}}^{(k)|b}(\ \cdot\ ; x)$ in (2.14). Let

$$q_b(i,j) \triangleq \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_j; m_i), \qquad q_b(i) \triangleq \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_i^c; m_i). \tag{3.10}$$

Note that $\sum_{j\in[n_{\min}]:\ j\neq i} q_b(i,j) = q_b(i) \in (0,\infty)$ for any $i \in [n_{\min}]$; see (C.3). This allows us to define a discrete-time Markov chain $(S_n)_{n\geq 0}$ over state space V, with any state $v \in V_b^*$ being an absorbing state, such that the one-step transition kernel $\mathbf{P}(S_{n+1} = m_j | S_n = m_i) = q_b(i,j)/q_b(i)$ holds for any $m_i \in V \setminus V_b^*$ and any $m_j \in V$. Next, define

$$\theta_b(m_j|m_i) \triangleq \mathbf{P}(S_n = m_j \text{ for some } n \geq 0 \mid S_0 = m_i) \tag{3.11}$$

for any $m_i \in V$ and any $m_j \in V_b^*$ as the probability of being absorbed at m_j when starting from m_i. For any $m_i \in V_b^*$, by definition of $\theta_b(\cdot|m_i)$, we see that $\theta_b(m_i|m_i) = 1$. In case that $m_i \in V \setminus V_b^*$, the evaluation of $\theta_b(m_j|m_i)$ is straightforward using the fundamental matrix of the Markov chain; see, for instance, Chapter 3.3 of [14]. Lastly, given the generator of $Y^{*|b}$, we have

$$\mathbf{P}(Y_{t+h}^{*|b} = m_j \mid Y_t^{*|b} = m_i) = h \cdot \sum_{j'\in[n_{\min}]:\ j'\neq i} q_b(i,j')\theta_b(m_j|m_{j'}) + \boldsymbol{o}(h) \qquad \text{as } h \downarrow 0$$

for any $m_i,\ m_j \in V_b^*$ with $m_i \neq m_j$.

Moving onto the untruncated process $X_j^\eta(x)$, Theorem 3.3 establishes a sample-path level convergence of $X_j^\eta(x)$ by sending $b \to \infty$ in Theorem 3.2. In particular, given any $T > 0$, there is a high chance that $X_j^\eta(x)$ coincides with the truncated dynamics $X_j^{\eta|b}(x)$ for all $j \leq T$ if the truncation threshold b is large. Therefore, as the truncation threshold b of $X_j^{\eta|b}(x)$ tends to ∞ in Theorem 3.2, we recover the results for $X_j^\eta(x)$. More precisely, recall the definition of measure $\check{\mathbf{C}}(\ \cdot\ ; x)$ in (2.15). For $i, j \in [n_{\min}]$ with $i \neq j$, let

$$q(i,j) \triangleq \check{\mathbf{C}}(I_j; m_i), \qquad q(i) \triangleq \sum_{j\in[n_{\min}]:\ j\neq i} q(i,j). \tag{3.12}$$

Recall that $H(\cdot) = \mathbf{P}(|Z| > \cdot)$. Theorem 3.3 shows that, under time scaling $1/H(\eta^{-1})$, the process X_j^η converges in distribution to a Markov jump process at the sample-path level. The proof is given in Section B.

Theorem 3.3. *Let Assumptions 1, 2, 3, 4, and 6 hold. Let $p \in [1,\infty)$, $i \in [n_{min}]$, and $x \in I_i$. As $\eta \downarrow 0$,*

$$\{X^\eta_{\lfloor t/H(\eta^{-1})\rfloor}(x) :\ t > 0\} \overset{f.d.d.}{\to} \{Y_t^* :\ t > 0\} \quad \text{and} \quad X^\eta_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \Rightarrow Y_\cdot^* \text{ in } (\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)})$$

where Y_t^* is a continuous-time Markov chain with a finite state space V, initial value $Y_0^* = m_i$, and infinitesimal generator

$$Q^*(i,j) = q(i,j) \qquad \forall m_i,\ m_j \in V \text{ with } m_i \neq m_j,$$
$$Q^*(i,i) = -\sum_{j\in[n_{min}]:\ j\neq i} Q^*(i,j) = -q(i) \qquad \forall m_i \in V.$$

Finally, we state a direct corollary of Theorem 3.2 that highlights the elimination of sharp minima under truncated heavy-tailed dynamics. Theorem 3.2 reveals that, under small η, the sample path of the truncated dynamics $X_j^{\eta|b}(x)$ closely resembles that of a Markov jump process that completely avoids all the narrower attraction fields of the potential U. Corollary 3.4 then further demonstrates that the fraction of time $X_j^{\eta|b}(x)$ spends around sharp minima converges in probability to 0 as $\eta \downarrow 0$. This result follows directly from Theorem 3.2 and the continuous mapping argument. In particular, given any $\epsilon,\ T > 0$ and mapping

$$f(\xi) = \frac{1}{T}\int_0^T \mathbf{I}\Big\{\xi_t \in \bigcup_{j:\ m_j\in V_b^*} (m_j - \epsilon, m_j + \epsilon)\Big\} dt,$$

one can see that $f : \mathbb{D}[0,\infty) \to \mathbb{R}$ is continuous at any ξ that, over $[0,T]$, only takes values in V_b^* and only makes finitely many jumps.

Corollary 3.4. *Let Assumptions 1, 2, 3, 6, and 7 hold. For any $i \in [n_{min}]$, $x \in I_i$, $T > 0$, and $\epsilon > 0$,*

$$\frac{1}{\lfloor T/\lambda_b^*(\eta)\rfloor}\sum_{t=1}^{\lfloor T/\lambda_b^*(\eta)\rfloor} \mathbf{I}\Big\{X_t^{\eta|b}(x) \in \bigcup_{j:\ m_j\in V_b^*} (m_j - \epsilon, m_j + \epsilon)\Big\} \xrightarrow{p} 1 \qquad \text{as } \eta \downarrow 0$$

where $\xrightarrow{p}$ stands for convergence in probability.

4 Simulation Experiments

We empirically demonstrate that (a) as indicated by Theorem 2, the minimum jump number defined in (3.2) accurately characterizes the first exit times of the SGDs with clipped heavy-tailed gradient

noises; (b) sharp minima can be effectively eliminated from such SGD; and (c) these properties are exclusive to heavy-tails. Under light-tailed noises, SGDs are trapped in sharp minima for a long time. The test function $f \in \mathcal{C}^2(\mathbb{R})$ is the same as in Fig. 1.1 (Left,e). m_1 and m_3 are sharp minima in narrow attraction fields, while m_2 and m_4 are flatter and located in larger attraction fields. Heavy-tailed noises have tail index $\alpha = 1.2$, and *light-tailed* noises are $\mathcal{N}(0, 1)$. See Section F for details.

First, we compare the first exit time of heavy-tailed SGD (when initialized at -0.7) from $\Omega_2 = (-1.3, 0.2)$ under 3 different clipping mechanism: (1) $b = 0.28$, where the minimum jump number required to escape is $l^* = 3$; (2) $b = 0.5$, where $l^* = 2$; (3) no gradient clipping, where $l^* = 1$ obviously. According to Theorem 2, the first exit times for the aforementioned 3 clipping mechanism are of order $(1/\eta)^{1.6}, (1/\eta)^{1.4}$ and $(1/\eta)^{1.2}$ respectively. These theoretical predictions are accurate as demonstrated in Figure 1.1 (Right). Next, we investigate the global dynamics of heavy-tailed SGD. We compared the clipped case (with $b = 0.5$) against the case without clipping. Figure 1.1 (Left, a, b) show the histograms of the empirical distributions of SGD, and Figure 1.1(Middle, a,b) plots the SGD trajectories. Without gradient clipping, X_n still visits the two sharp minima m_1, m_3. Under gradient clipping, the time spent at m_1, m_3 is almost completely eliminated and is negligible compared to the time X_n spent at m_2, m_4 in larger attraction fields. This matches the predictions of Theorem 3.2 and Corollary 3.4, i.e., the elimination of sharp minima with truncated heavy-tailed noises. We stress that the said properties are exclusive to heavy-tailed SGD. As shown in Figure 1.1(Left,c,d) and Figure 1.1(Middle, c,d), light-tailed SGD are easily trapped at sharp minima for extremely long time.

Figure 4.1 illustrates the same phenomena in $\mathbb{R}^2$, where f has several saddle points and infinitely many local minima—the local minima of Ω_2 form a line segment, which is an uncountably infinite set. Under clipping threshold b, attraction fields Ω_1 and Ω_2 are the *larger* ones since the escape from them requires at least two jumps. This suggests that the theoretical results from Section 3 also hold for the more general multidimensional settings.

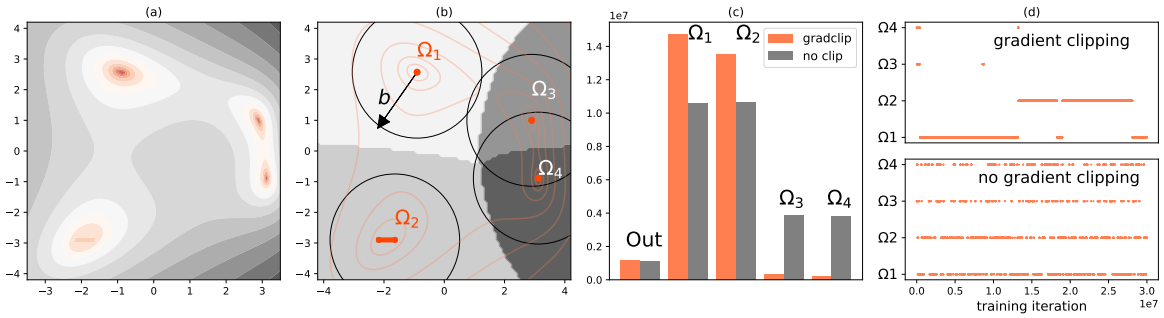

Figure 4.1: Experiment result of heavy-tailed SGD when optimizing the modified Himmelblau function. **(a)** Contour plot of the test function f. **(b)** Different shades of gray are used to indicate the area of the four different attraction fields $\Omega_1, \Omega_2, \Omega_3, \Omega_4$ of f. We say that a point belongs to an attraction field Ω_i if, when initializing at this point, the gradient descent iterates converge to the local minima in Ω_i, which are indicated by the colored dots. The circles are added to imply whether the SGD iterates can escape from each Ω_i with one large jump or not under clipping threshold b. **(c)** The time heavy-tailed SGD spent at different region. An iterate X_k is considered "visiting" Ω_i if its distance to the local minimizer of Ω_i is less than 0.5; otherwise we label X_k as "out". **(d)** The transition trajectories of heavy-tailed SGD. The dots represent the last "visited" attraction field at each iteration.

Table 5.1: Test accuracy and expected sharpness of different methods across different tasks. The reported numbers are the averages over 5 replications. For 95% CI, see Section F.

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
FashionMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
FashionMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

5 Deep Learning Experiments: An Ablation Study

In this section, we verify our theoretical results and demonstrate the effectiveness of clipped heavy-tailed noise in training deep neural networks. Let θ be the current model weight during training, $g_{SB}(\theta)$ be the typical small-batch gradient, and $g_{GD}(\theta)$ be the true (deterministic) gradient evaluated on the entire training dataset. Then by evaluating $g_{SB}(\theta) - g_{GD}(\theta)$ we obtain a sample of the gradient noise. Due to the prohibitive cost of evaluating $g_{GD}(\theta)$, we instead use $g_{SB}(\theta) - g_{LB}(\theta)$ as its approximation where g_{LB} denotes the gradient evaluated on a larger batch. This is justified by the unbiasedness in $\mathbf{E}_{LB}[g_{LB}(\theta)] = g_{GD}(\theta)$. For some heavy-tailed random variable Z, by multiplying Z with SGD noise, we obtain the following perturbed gradient:

$$g_{heavy}(\theta) \;=\; g_{SB}(\theta) + Z\big(g_{SB*}(\theta) - g_{LB}(\theta)\big) \tag{5.1}$$

where SB and $SB*$ are two mini batches that may or may not be identical. We use the following update recursion under gradient clipping threshold b: $X_{k+1}^{\eta} = X_k^{\eta} - \varphi_b(\eta g_{heavy}(X_k^{\eta}))$ where φ_b is the truncation operator. We consider two different implementations: in *our method 1* (labeled as "our 1" in Table 5.1), SB and $SB*$ are chosen independently, while in *our method 2* (labeled as "our 2" in Table 5.1), we use the same batch for SB and $SB*$. In summary, by simply multiplying gradient noise with heavy-tailed random variables, we inject heavy-tailed noise into the optimization procedure.

We conduct an ablation study and benchmark the proposed clipped heavy-tailed methods against the following optimization methods. *LB*: large-batch SGD with $X_{k+1}^{\eta} = X_k^{\eta} - \eta g_{LB}(X_k^{\eta})$; *SB*: small-batch SGD with $X_{k+1}^{\eta} = X_k^{\eta} - \eta g_{SB}(X_k^{\eta})$; *SB + Clip*: the update recursion is $X_{k+1}^{\eta} = X_k^{\eta} - \varphi_b(\eta g_{SB}(X_k^{\eta}))$; *SB + Noise*: Our method 2 WITHOUT the gradient clipping mechanism, leading to the update recursion $X_{k+1}^{\eta} = X_k^{\eta} - \eta g_{heavy}(X_k^{\eta})$.

The experiment setting and choice of hyperparameters are adapted from [37]. We consider three different tasks: (1) LeNet [17] on corrupted FashionMNIST [35], (2) VGG11 [29] on SVHN [21], (3) VGG11 on CIFAR10 [16] (see Section F for details). Here we highlight a few points: First, within the same task, for all the 6 candidate methods will use the same η, batch size, training iteration, and (when needed) the same clipping threshold b and heavy-tailed RV Z for a fair comparison; the training duration is long enough so that LB and SB have attained 100% training accuracy and close-to-0 training loss long before the end of training (the exception here is "*SB + Noise*" method; see Section F for the details); Second, to facilitate convergence to local minima for *our methods 1 and 2*, we remove heavy-tailed noise for last final 5,000 iterations and run LB instead[1].

Table 5.1 shows that in all 3 tasks both *our method 1* and *our method 2* attain better test accuracy than the other candidate methods. Meanwhile, both methods exhibit similar test performance, implying that the implementation of the heavy-tailed method may not be a the deciding factor. We also

[1]The proposed method can be interpreted as a simplified version of GD + annealed heavy-tailed perturbation, where a detailed annealing is substituted by a two-phase training schedule. In the first *exploration* phase the clipped heavy-tailed noises drive the iterates to explore the loss landscape and identify "wide" attraction fields. In the second *exploitation* phase, removing the artificial perturbation accelerates convergence to local minima.

Table 5.2: Our method's gain on test accuracy persists even when applied with techniques such as data augmentation and scheduled learning rates. For 95% CI, see Section F.

CIFAR10-VGG11	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Average
SB+Clip	89.40%	89.41%	89.89%	89.52%	89.47%	89.54%
Our 1	90.76%	90.57%	90.49%	90.85%	90.79%	**90.67%**
Our 2	90.67%	90.23%	90.52%	90.13%	90.70%	90.45%
CIFAR100-VGG16	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Average
SB+Clip	55.76%	56.8%	56.38%	56.35%	56.32%	56.32%
Our 1	67.43%	65.12%	65.14%	65.96%	63.57%	**65.44%**
Our 2	67.19%	61.17%	60.97%	64.75%	60.90%	62.99%

report the *expected sharpness* metric $\mathbf{E}_{\nu\sim\mathcal{N}(\mathbf{0},\delta^2\mathbf{I})}|L(\theta^*+\nu)-L(\theta^*)|$ used in [37, 22] where $\mathcal{N}(\mathbf{0},\delta^2\mathbf{I})$ is a Gaussian distribution, θ^* is the trained model weight and L is training loss. In our experiment, we use $\delta = 0.01$ and the expectation is evaluated by averaging over 100 samples. We conduct 5 replications for each experiment scenario and report the averaged performance in Table 5.1. Smaller sharpness of our methods 1 and 2 confirms that they encourage minimizers with a "flatter" geometry, thus attaining better test performances.

The ablation study in Table 5.1 shows that both heavy-tailed noise and gradient clipping are necessary to find a flat minima and hence achieve better generalization, which is predicted by our analyses. *SB* and *SB + Clip* achieve similar inferior performances, confirming that clipping does not help when noise is light-tailed. *SB + Noise* injects heavy-tailed noise without gradient clipping, which achieves an inferior performance. This poor performance—even after extensive parameter tuning and engineering (see Section F for more details)—demonstrates the difficulty on the optimization front, especially when heavy-tailed noise is present yet little effort is put into controlling the highly volatile gradient noises. This is aligned with the observations in [36, 5] where adaptive gradient clipping methods are proposed to improve convergence of SGD in the presence of heavy-tailed noises. This confirms that gradient clipping is crucial for heavy-tailed SGD.

Lastly, Table 5.2 shows that even in the more sophisticated settings with training techniques such as data augmentations and scheduled learning rates, truncated heavy-tailed SGD still manages to consistently find solutions with better test performance. For experiment details, see Section F. In Table F.5 we also report the sharpness of the obtained solutions.

References

[1] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2nd ed edition, 1999.

[2] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[3] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020.

[4] S. Garg, J. Zhanson, E. Parisotto, A. Prasad, J. Z. Kolter, Z. C. Lipton, S. Balakrishnan, R. Salakhutdinov, and P. K. Ravikumar. On proximal policy optimization's heavy-tailed gradients, 2021.

[5] E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053. Curran Associates, Inc., 2020.

[6] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[7] M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in sgd. *arXiv preprint arXiv:2006.04740*, 2020.

[8] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[9] L. Hodgkinson and M. W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. *arXiv preprint arXiv:2006.06293*, 2020.

[10] P. Imkeller and I. Pavlyukevich. Metastable behaviour of small noise lévy-driven diffusions. *ESAIM: PS*, 12:412–437, 2008.

[11] P. Imkeller, I. Pavlyukevich, and M. Stauch. First exit times of non-linear dynamical systems in $\mathbb{R}^d$ perturbed by multifractal Lévy noise. *Journal of Statistical Physics*, 141(1):94–119, 2010.

[12] P. Imkeller, I. Pavlyukevich, and T. Wetzel. The hierarchy of exit times of Lévy-driven Langevin equations. *The European Physical Journal Special Topics*, 191(1):211–222, 2010.

[13] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[14] J. Kemeny and J. Snell. Finite markov chains. with a new appendix. In *Generalization of a fundamental matrix*. Springer, 1983.

[15] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[16] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[17] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[18] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6391–6401, 2018.

[19] M. Mahoney and C. Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

[20] S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*, 2018.

[21] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[22] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro. Exploring generalization in deep learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[23] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.

[24] T. H. Nguyen, U. Simsekli, and G. Richard. Non-asymptotic analysis of fractional langevin monte carlo for non-convex optimization. In *International Conference on Machine Learning*, pages 4810–4819. PMLR, 2019.

[25] A. Panigrahi, R. Somani, N. Goyal, and P. Netrapalli. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

[26] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR, 2013.

[27] I. Pavlyukevich. Cooling down lévy flights. *Journal of Physics A: Mathematical and Theoretical*, 40(41):12299, 2007.

[28] S. I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] U. Şimşekli, M. Gürbüzbalaban, T. H. Nguyen, G. Richard, and L. Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

[31] U. Şimşekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[32] V. Srinivasan, A. Prasad, S. Balakrishnan, and P. K. Ravikumar. Efficient estimators for heavy-tailed machine learning, 2021.

[33] X. Wang, S. Oh, and C.-H. Rhee. Eliminating sharp minima from SGD with truncated heavy-tailed noise. In *International Conference on Learning Representations*, 2022.

[34] X. Wang and C.-H. Rhee. Large deviations and metastability analysis for heavy-tailed dynamical systems, 2023.

[35] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[36] J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[37] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7654–7663. PMLR, 09–15 Jun 2019.

The appendices are structured as follows. Section A develops a theoretical framework for establishing the sample path convergence of jump processes. Applying this framework, in Sections B–D we provide the proof of Theorems 3.2 and 3.3. The proof utilizes several results established in [34], which are collected in Section E. Section F provides the details of the simulation experiments in Section 4 and the deep learning experiments in Section 5.

A Technical Lemmas for Theorem 3.2

Let $Y_\cdot^\eta$ and $Y_\cdot^*$ be random elements in $\mathbb{D}[0,\infty)$, i.e., $\mathbb{R}$-valued càdlàg processes. We start by discussing a few properties of the weak convergence in $(\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)})$. In particular, a similar mode of convergence in $(\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$ can be defined analogously for any $T \in (0,\infty)$. Recall the projection mapping π_T defined in (3.6). We say that $Y_\cdot^\eta \Rightarrow Y_\cdot^*$ in $(\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$ if

$$\lim_{\eta\downarrow 0} \mathbf{E} f\big(\pi_T(S_\cdot^\eta)\big) = \mathbf{E} f\big(\pi_T(S_\cdot^*)\big) \qquad \forall f : \mathbb{D}[0,T] \to \mathbb{R} \text{ continuous and bounded;}$$

see (3.5) for the definition of $\boldsymbol{d}_{L_p}^{[0,T]}$. More precisely, the L_p norm $\boldsymbol{d}_{L_p}^{[0,T]}$ induces a metric over a quotient space $\mathbb{D}[0,T]/\mathcal{N}$. In particular, since we are dealing with the càdlàg space $\mathbb{D}[0,T]$, we set $\mathcal{N} = \{\xi \in \mathbb{D}[0,T] : \ \xi_t \equiv 0 \ \forall t \in [0,T)\}$, which is the set containing all paths in $\mathbb{D}[0,T]$ that is constant zero except for the endpoint.

First, Lemma A.1 shows that the convergence in $(\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)})$ follows from the convergence in $(\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$.

Lemma A.1. *Let $p \in [1,\infty)$. If $Y_\cdot^\eta \Rightarrow Y_\cdot^*$ in $(\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$ as $\eta \downarrow 0$ for any positive integer T, then $Y_\cdot^\eta \Rightarrow Y_\cdot^*$ in $(\mathbb{D}[0,\infty), \boldsymbol{d}_{L_p}^{[0,\infty)})$ as $\eta \downarrow 0$.*

Proof. By Portmanteau Theorem, it suffices to show that $\lim_{\eta\downarrow 0} \mathbf{E} f(Y_\cdot^\eta) = \mathbf{E} f(Y_\cdot^*)$ holds for any $f : \mathbb{D}[0,\infty) \to \mathbb{R}$ that is bounded and uniformly continuous. To proceed, we arbitrarily pick one such f and some $\epsilon > 0$. By virtue of the uniform continuity of f, there exists some $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $\boldsymbol{d}_{L_p}^{[0,\infty)}(x,y) < \delta$. By definition of $\boldsymbol{d}_{L_p}^{[0,\infty)}$ in (3.7), we must have $\boldsymbol{d}_{L_p}^{[0,\infty)}(x,y) < 1/2^{\lfloor T \rfloor - 1}$ if $x_t = y_t$ for all $t \in [0,T)$. Now, we fix some positive integer T large enough such that $1/2^{T-1} < \delta$. Define $\widetilde{\pi}_T : \mathbb{D}[0,\infty) \to \mathbb{D}[0,\infty)$ by

$$\widetilde{\pi}_T(\xi)_t \triangleq \begin{cases} \xi_t & \text{if } t \in [0,T) \\ 0 & \text{if } t \geq T \end{cases}$$

and set $\widetilde{f}_T(\xi) \triangleq f\big(\widetilde{\pi}_T(\xi)\big)$. We now have $\boldsymbol{d}_{L_p}^{[0,\infty)}\big(\xi, \widetilde{\pi}_T(\xi)\big) < \delta$ and hence $|f(\xi) - \widetilde{f}_T(\xi)| < \epsilon$ for any $\xi \in \mathbb{D}[0,\infty)$. As a result,

$$\limsup_{\eta\downarrow 0} |\mathbf{E} f(Y_\cdot^\eta) - \mathbf{E}\widetilde{f}_T(Y_\cdot^\eta)| < \epsilon, \qquad |\mathbf{E} f(Y_\cdot^*) - \mathbf{E}\widetilde{f}_T(Y_\cdot^*)| < \epsilon. \tag{A.1}$$

Furthermore, let $\pi_T^\dagger : \mathbb{D}[0,T] \to \mathbb{D}[0,\infty)$ be defined as

$$\pi^\dagger(\xi')_t \triangleq \begin{cases} \xi'_t & \text{if } t \in [0,T) \\ 0 & \text{if } t \geq T \end{cases},$$

which, at an intuitive level, is interpreted as a "pseudo inverse" of the projection mapping π_T defined in (3.6). Also, define functional $f_T : \mathbb{D}[0,T] \to \mathbb{R}$ by $f_T(\cdot) \triangleq f\big(\pi_T^\dagger(\cdot)\big)$. It is easy to see that (i) f_T is continuous due to the continuity of f and $\pi_T^\dagger$, and (ii) for any $\xi \in \mathbb{D}[0,\infty)$, we have $\widetilde{f}_T(\xi) = f_T\big(\pi_T(\xi)\big)$. Due to the assumption $Y_\cdot^\eta \Rightarrow Y_\cdot^*$ in $(\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$, we now yield

$$\lim_{\eta\downarrow 0} |\mathbf{E}\widetilde{f}_T(Y_\cdot^\eta) - \mathbf{E}\widetilde{f}_T(Y_\cdot^*)| = 0. \tag{A.2}$$

Combining (A.1) and (A.2), we get $\limsup_{\eta\downarrow 0}|\mathbf{E}f(Y_\cdot^\eta)-\mathbf{E}f(Y_\cdot^*)|<2\epsilon$. Driving $\epsilon\to 0$, we conclude the proof. □

Lemma A.2 then provides a Prohorov-type argument where weak convergence in $(\mathbb{D}[0,T],\boldsymbol{d}_{L_p}^{[0,T]})$ can be established using the convergence in f.d.d. and a tightness condition. The proof is a straightforward adaptation of its J_1 counterparts. For the sake of clarity, the next proof will, w.l.o.g., focus on the case where $T=1$, but we stress that the arguments can be easily extended to $\mathbb{D}[0,T]$ with arbitrary $T\in(0,\infty)$. Recall that we write $\mathbb{D}=\mathbb{D}[0,1]$.

Lemma A.2. *Let $T\in(0,\infty)$, $p\in[1,\infty)$, and $\mathcal{T}$ be a dense subset of $(0,T)$. Suppose that the laws of $Y_\cdot^{\eta_n}$ are tight in $(\mathbb{D}[0,T],\boldsymbol{d}_{L_p}^{[0,T]})$ for any sequence $\eta_n>0$ with $\lim_n\eta_n=0$, and*

$$(Y_{t_1}^\eta,\cdots,Y_{t_k}^\eta)\Rightarrow(Y_{t_1}^*,\cdots,Y_{t_k}^*)\ \text{as}\ \eta\downarrow 0\qquad\forall k=1,2,\cdots,\ \forall(t_1,\cdots,t_k)\in\mathcal{T}^{k\uparrow}. \tag{A.3}$$

Then $Y_\cdot^\eta\Rightarrow Y_\cdot^*$ in $(\mathbb{D}[0,T],\boldsymbol{d}_{L_p}^{[0,T]})$ as $\eta\downarrow 0$.

Proof. The arguments below are adapted from the standard proofs in [1] for J_1 topology. For any $0\le t_1<t_2<\cdots<t_k\le 1$, let $\pi_{(t_1,\cdots,t_k)}:\mathbb{D}\to\mathbb{R}^k$ be the projection mapping, i.e., $\pi_{(t_1,\cdots,t_k)}(\xi)=(\xi_{t_1},\xi_{t_2},\cdots,\xi_{t_k})$. Let $\mathcal{R}^k$ be the Borel σ-algebra for $\mathbb{R}^k$. Let $p[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]$ be the collection of all sets of form $\pi_{(t_1,\cdots,t_k)}^{-1}H$, where $k\ge 1$, $H\in\mathcal{R}^k$, and $t_1<\cdots<t_k$ with $t_i\in\mathcal{T}$ for each $i\in[k]$. It suffices to show that (recall that $\boldsymbol{d}_{L_p}=\boldsymbol{d}_{L_p}^{[0,1]}$ and let $\mathcal{D}_p$ be the Borel σ-algebra of $(\mathbb{D},\boldsymbol{d}_{L_p})$)

$$p[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]\ \text{is a separating class for}\ (\mathbb{D},\boldsymbol{d}_{L_p}). \tag{A.4}$$

In other words, any two Borel probability measures μ and ν over $(\mathbb{D},\boldsymbol{d}_{L_p})$ would coincide (i.e., $\mu(A)=\nu(A)\ \forall A\in\mathcal{D}_p$) if $\mu(A)=\nu(A)\ \forall A\in p[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]$. To see why claim (A.4) is a sufficient condition, note that the tightness condition in Lemma A.2 implies that the sequence $Y_\cdot^{\eta_n}$ has a converging subsequence, while the claim (A.4) and assumption (A.3) dictate that the limiting distribution must be that of $Y_\cdot^*$.

The remainder of this proof is devoted to establishing claim (A.4). First, we show that the projection mapping of form $\pi_{(t_1,\cdots,t_k)}:\mathbb{D}\to\mathbb{R}^k$ is $\mathcal{D}_p/\mathcal{R}^k$ measurable when $0\le t_1<\cdots<t_k<1$, which immediately confirms that $p[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]\subseteq\mathcal{D}_p$. To do so, it suffices to prove that $\pi_{(t)}$ is measurable for any given $t\in[0,1)$. Define $h_\epsilon(x):\mathbb{D}\to\mathbb{R}$ by $h_\epsilon(x)=\epsilon^{-1}\int_t^{t+\epsilon}x_s ds$. W.l.o.g. we only consider ϵ small enough such that $t+\epsilon\le 1$. For any $x,\ y\in\mathbb{D}$ and $\Delta\in(0,1)$,

$$\begin{aligned}|h_\epsilon(x)-h_\epsilon(y)|&\le\epsilon^{-1}\int_t^{t+\epsilon}|x_s-y_s|ds\\&=\epsilon^{-1}\int_t^{t+\epsilon}|x_s-y_s|\mathbf{I}\{|x_s-y_s|>\Delta\}ds+\epsilon^{-1}\int_t^{t+\epsilon}|x_s-y_s|\mathbf{I}\{|x_s-y_s|\le\Delta\}ds\\&\le\epsilon^{-1}\int_t^{t+\epsilon}\frac{|x_s-y_s|^p}{|\Delta|^p}ds+\Delta.\end{aligned}$$

Therefore, for any sequence $y^{(n)}\in\mathbb{D}$ such that $\boldsymbol{d}_{L_p}(y^{(n)},x)\to 0$, we have $\limsup_{n\to\infty}|h_\epsilon(x)-h_\epsilon(y^{(n)})|\le\Delta$. Driving $\Delta\downarrow 0$, we see that $h_\epsilon(\cdot)$ is a continuous mapping. On the other hand, the right continuity of all paths in $\mathbb{D}$ implies that $h_\epsilon(x)\to\pi_{(t)}(x)$ as $\epsilon\to 0$ for all $x\in\mathbb{D}$. As a result, the limiting mapping $\pi_{(t)}$ must be $\mathcal{D}_p/\mathcal{R}$ measurable.

Let $\sigma[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]$ be the σ-algebra generated by $p[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]$. Note that we have verified $p[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]\subseteq\mathcal{D}_p$, which implies $\sigma[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]\subseteq\mathcal{D}_p$ since $\mathcal{D}_p$ is also a σ-algebra. Furthermore, suppose that we can show

$$\sigma[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]\supseteq\mathcal{D}_p\qquad(\text{and hence}\ \sigma[\pi_{\boldsymbol{t}}:\boldsymbol{t}\in\mathcal{T}]=\mathcal{D}_p), \tag{A.5}$$

then we can confirm claim (A.4) using $\pi-\lambda$ Theorem. Indeed, for any Borel probability measures μ and ν over $(\mathbb{D}, \boldsymbol{d}_{L_p})$, note that $\mathcal{L} \triangleq \{A \in \mathcal{D}_p : \ \mu(A) = \nu(A)\}$ is a λ-system. Whenever $p[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}] \subseteq \mathcal{L}$, by applying $\pi - \lambda$ Theorem we then get $\sigma[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}] = \mathcal{D}_p \subseteq \mathcal{L}$. This concludes that $p[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}]$ is a separating class.

Now, it only remains to prove claim (A.5). Since $\mathcal{T}$ is a dense subset of $(0, T)$, for each $m \geq 1$ we can pick some positive integer k and some $0 < s_1 < \cdots < s_k < 1$, with $s_i \in \mathcal{T}$, such that $\max_{i\in[k+1]} |s_{i+1} - s_i| < m^{-1}$, under the convention that $s_0 = 0$ and $s_{k+1} = 1$. Now, construct a mapping $V_m : \mathbb{R}^k \to \mathbb{D}$ as follows: for each $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k) \in \mathbb{R}^k$, define $\xi = V_m(\boldsymbol{\alpha})$ by setting $\xi_t = \alpha_i$ if $t \in [s_i, s_{i+1})$ for each $i \in [k + 1]$ (with the convention that $\alpha_0 = 0$) and $\xi_1 = \alpha_k$. It is easy to see that V_m is continuous, and hence $\mathcal{R}^k/\mathcal{D}_p$ measurable. Besides, we have shown that $\pi_{(t_1,\cdots,t_k)}$ is $\sigma[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}]/\mathcal{R}^k$ measurable. As a result, the composition $V_m^* \triangleq V_m \pi_{s_1,\cdots,s_k} : \mathbb{D} \to \mathbb{D}$ is $\sigma[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}]/\mathcal{D}_p$ measurable.

To proceed, fix some $\epsilon > 0$. For any $x \in \mathbb{D}$, define $x' \in \mathbb{D}$ such that $x'_t = x_t$ for all $t \in [\epsilon, 1 - \epsilon)$ and $x'_t = 0$ otherwise. The boundedness of any path in $\mathbb{D}$ implies the existence of some $M_x \in (0, \infty)$ such that $\sup_t |x_t| \leq M_x$. Next, note that

$$\boldsymbol{d}_{L_p}(V_m^* x, x) \leq \underbrace{\boldsymbol{d}_{L_p}(V_m^* x', x')}_{\text{(I)}} + \underbrace{\boldsymbol{d}_{L_p}(V_m^* x', V_m^* x)}_{\text{(II)}} + \underbrace{\boldsymbol{d}_{L_p}(x', x)}_{\text{(III)}}.$$

First, it was shown in Theorem 12.5 of [1] that $\lim_{m\to\infty} \boldsymbol{d}_{J_1}(V_m^* x', x') = 0$. This immediately implies that $\lim_{m\to\infty} \boldsymbol{d}_{L_p}(V_m^* x', x') = 0$. Next, by definition of x', we have $\limsup_{m\to\infty} \big[\text{(II)}\big]^p \leq (2M_x)^p \cdot 2\epsilon$ and $\limsup_{m\to\infty} \big[\text{(III)}\big]^p \leq (2M_x)^p \cdot 2\epsilon$. Driving $\epsilon \downarrow 0$, we obtain that $\lim_{m\to\infty} \boldsymbol{d}_{L_p}(V_m^* x, x) = 0$ for all $x \in \mathbb{D}$. This implies that the identity mapping $\mathbf{I}(\xi) = \xi$ is also $\sigma[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}]/\mathcal{D}_p$ measurable, which leads to $\mathcal{D}_p \subseteq \sigma[\pi_{\boldsymbol{t}} : \boldsymbol{t} \in \mathcal{T}]$ and concludes the proof. □

Moreover, consider a family of $\mathbb{R}$-valued càdlàg processes $\hat{Y}_t^{\eta,\epsilon}$, supported on the same underlying probability space with process Y_t^η, that satisfies the following condition.

Condition 1. *For any $T \in (0, \infty)$ and $p \in [1, \infty)$, the following claims hold for all $\epsilon > 0$ small enough:*

(i) $\{\hat{Y}_t^{\eta,\epsilon} : \ t > 0\} \overset{f.d.d.}{\to} \{Y_t^* : \ t > 0\}$ *and* $\hat{Y}_\cdot^{\eta,\epsilon} \Rightarrow Y_\cdot^*$ *in* $(\mathbb{D}[0, T], \boldsymbol{d}_{L_p}^{[0,T]})$ *as* $\eta \downarrow 0$*;*

(ii) $\lim_{\eta\to 0} \mathbf{P}(|\hat{Y}_T^{\eta,\epsilon} - Y_T^\eta| \geq \epsilon) = 0$ *and* $\lim_{\eta\downarrow 0} \mathbf{P}\Big(\boldsymbol{d}_{L_p}^{[0,T]}\big(\hat{Y}_\cdot^{\eta,\epsilon}, Y_\cdot^\eta\big) \geq 2\epsilon\Big) = 0.$

As the first component of our framework, Lemma A.3 shows that, under Condition 1, both Y_t^η and $\hat{Y}_t^{\eta,\epsilon}$ admit the same limit Y_t^*.

Lemma A.3. *If Condition 1 holds, then* $\{Y_t^\eta : \ t > 0\} \overset{f.d.d.}{\to} \{Y_t^* : \ t > 0\}$ *and, for any $T > 0$,* $Y_\cdot^\eta \Rightarrow Y_\cdot^*$ *in* $(\mathbb{D}[0, T], \boldsymbol{d}_{L_p}^{[0,T]})$ *as* $\eta \downarrow 0$.

Proof. We start with the L_p convergence. By Portmanteau Theorem, it suffices to show that $\liminf_{\eta\downarrow 0} \mathbf{P}(Y_\cdot^\eta \in G) \geq \mathbf{P}(Y_\cdot^* \in G)$ for any open set G in the L_p topology of $\mathbb{D}[0, T]$. Next, (recall that G_ϵ is the ϵ-shrinkage of G, and G_ϵ is also an open set)

$$\begin{aligned}\mathbf{P}(Y_\cdot^\eta \in G) &\geq \mathbf{P}(Y_\cdot^\eta \in G, \ \boldsymbol{d}_{L_p}^{[0,T]}(\hat{Y}_\cdot^{\eta,\epsilon}, Y_\cdot^\eta) < 2\epsilon) \geq \mathbf{P}(\hat{Y}_\cdot^{\eta,\epsilon} \in G_{2\epsilon}, \ \boldsymbol{d}_{L_p}^{[0,T]}(\hat{Y}_\cdot^{\eta,\epsilon}, Y_\cdot^\eta) < 2\epsilon)\\ &\geq \mathbf{P}(\hat{Y}_\cdot^{\eta,\epsilon} \in G_{2\epsilon}) - \mathbf{P}(\boldsymbol{d}_{L_p}^{[0,T]}(\hat{Y}_\cdot^{\eta,\epsilon}, Y_\cdot^\eta) \geq 2\epsilon).\end{aligned}$$

For small enough $\epsilon > 0$, using part (i) of Condition 1 we get $\liminf_{\eta\downarrow 0} \mathbf{P}(\hat{Y}_\cdot^{\eta,\epsilon} \in G_{2\epsilon}) \geq \mathbf{P}(Y_\cdot^* \in G_{2\epsilon})$, and by part (ii) of Condition 1 we have $\lim_{\eta\downarrow 0} \mathbf{P}(\boldsymbol{d}_{L_p}^{[0,T]}(\hat{Y}_\cdot^{\eta,\epsilon}, Y_\cdot^\eta) \geq 2\epsilon) = 0$. Therefore, $\liminf_{\eta\downarrow 0} \mathbf{P}(Y_\cdot^\eta \in G) \geq \mathbf{P}(Y_\cdot^* \in G_{2\epsilon})$. Driving $\epsilon \downarrow 0$, we conclude the proof for the L_p convergence. The proof for the f.d.d. convergence is almost identical and hence we omit the details. □

In light of Lemma A.3, a natural approach to Theorem 3.2 is to identify some $\hat{Y}_t^{\eta,\epsilon}$ that converges to ${Y_t^*}^{|b}$ while staying close enough to $X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta)\rfloor}(x)$. To this end, we introduce the next key component of our framework, i.e., a technical tool for establishing the weak convergence of jump processes. Inspired by the approach in [10], Lemma A.5 shows that the convergence of jump processes can be established by verifying the convergence of the inter-arrival times and destinations of jumps. Specifically, we introduce the following mapping Φ for constructing jump processes.

Definition A.4. *Let random variables $((U_j)_{j\geq 1},(V_j)_{j\geq 1})$ be such that $V_j\in\mathbb{R}$ $\forall j\geq 1$ and*

$$U_j\in[0,\infty)\quad\forall j\geq 1,\qquad \lim_{i\to\infty}\mathbf{P}\Big(\sum_{j=1}^{i}U_j>t\Big)=1\quad\forall t>0. \tag{A.6}$$

Let mapping $\Phi(\cdot)$ be defined as follows: the image $Y_\cdot=\Phi\big((U_j)_{j\geq 1},(V_j)_{j\geq 1}\big)$ is a stochastic process taking values in $\mathbb{R}$ such that (under the convention $V_0\equiv 0$)

$$Y_t=V_{\mathcal{J}(t)}\ \forall t\geq 0\qquad \textit{where}\qquad \mathcal{J}(t)\triangleq\max\{J\geq 0:\ \sum_{j=1}^{J}U_j\leq t\}. \tag{A.7}$$

Remark 2. *We add two remarks regarding Definition A.4. First, $(U_j)_{j\geq 1}$ and $(V_j)_{j\geq 1}$ can be viewed as the inter-arrival times and destinations of jumps in Y_t, respectively. It is worth noticing that we allow for instantaneous jumps, i.e., $U_j=0$. Nevertheless, the condition $\lim_{i\to\infty}\mathbf{P}(\sum_{j=1}^{i}U_j>t)=1$ $\forall t>0$ prevents the concentration of infinitely many instantaneous jumps before any finite time $t\in(0,\infty)$, thus ensuring that the process $Y_t=V_{\mathcal{J}(t)}$ is almost surely well defined. In case that $U_j>0$ $\forall j\geq 1$, the process Y_t admits a more standard expression and satisfies $Y_t=V_i$ for all $t\in[\sum_{j=1}^{i}U_j,\sum_{j=1}^{i+1}U_j)$. Second, to account for the scenario where the process Y_t stays constant after a (possibly random) timestamp T, one can introduce dummy jumps that keep landing at the same location. For instance, suppose that after hitting $w\in\mathbb{R}$ the process Y_t is absorbed at w, then a representation compatible with Definition A.4 is that, conditioning on $V_j=w$, we set U_k as iid Exp(1) RVs and $V_k\equiv w$ for all $k\geq j+1$.*

As mentioned above, Lemma A.5 states that the convergence of jump processes in f.d.d. follows from the convergence in distributions of the inter-arrival times and destinations of jumps.

Lemma A.5. *Let mapping Φ be specified as in Definition A.4. Let $Y_\cdot=\Phi\big((U_j)_{j\geq 1},(V_j)_{j\geq 1}\big)$ and, for each $n\geq 1$, $Y_\cdot^n=\Phi\big((U_j^n)_{j\geq 1},(V_j^n)_{j\geq 1}\big)$. Suppose that*

- *$(U_1^n,V_1^n,U_2^n,V_2^n,\cdots)$ converges in distribution to $(U_1,V_1,U_2,V_2,\cdots)$ as $n\to\infty$;*
- *For any $u>0$ and any $j\geq 1$, $\mathbf{P}(U_1+\cdots+U_j=u)=0$;*
- *For any $u>0$, $\lim_{j\to\infty}\mathbf{P}(U_1+U_2+\cdots U_j>u)=1$.*

Then $\{Y_t^n:\ t>0\}\overset{f.d.d.}{\to}\{Y_t^*:\ t>0\}$ as $n\to\infty$.

Proof. Fix some $k\in\mathbb{N}$ and $0<t_1<t_2<\cdots<t_k<\infty$. Set $t=t_k$. Pick some $\epsilon>0$. By assumption, one can find some $J(\epsilon)>0$ such that $\mathbf{P}(\sum_{j=1}^{J(\epsilon)}U_j\leq t)<\epsilon$, and hence $\mathbf{P}(\sum_{j=1}^{J(\epsilon)}U_j^n\leq t)<\epsilon$ for all n large enough. Also, we can fix $\Delta(\epsilon)>0$ such that $\mathbf{P}\Big(\sum_{i=1}^{j}U_i\in\bigcup_{l\in[k]}[t_l-\Delta(\epsilon),t_l+\Delta(\epsilon)]$ for some $j\leq J(\epsilon)\Big)<\epsilon$. Throughout the proof, we may abuse the notation slightly and write $J=J(\epsilon)$ and $\Delta=\Delta(\epsilon)$ when there is no ambiguity.

For any probability measure μ, let $\mathscr{L}_\mu(X)$ be the law of the random element X under μ. Applying Skorokhod's representation theorem, we can construct a probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\mathbf{Q})$ that supports random variables $(\widetilde{U}_1^n,\widetilde{V}_1^n,\widetilde{U}_2^n,\widetilde{V}_2^n\cdots)_{n\geq 1}$ and $(\widetilde{U}_1,\widetilde{V}_1,\widetilde{U}_2,\widetilde{V}_2,\cdots)$ such that (i) $\mathscr{L}_{\mathbf{P}}(U_1^n,V_1^n,U_2^n,V_2^n,\cdots)=$

$\mathscr{L}_{\mathbf{Q}}(\widetilde{U}_1^n, \widetilde{V}_1^n, \widetilde{U}_2^n, \widetilde{V}_2^n \cdots)$ for all $n \geq 1$, (ii) $\mathscr{L}_{\mathbf{P}}(U_1, V_1, U_2, V_2, \cdots) = \mathscr{L}_{\mathbf{Q}}(\widetilde{U}_1, \widetilde{V}_1, \widetilde{U}_2, \widetilde{V}_2, \cdots)$, and (iii) $\widetilde{U}_j^n \xrightarrow{\mathbf{Q}-a.s.} \widetilde{U}_j$ and $\widetilde{V}_j^n \xrightarrow{\mathbf{Q}-a.s.} \widetilde{V}_j$ as $n \to \infty$ for all $j \geq 1$. This allows us to construct a coupling between processes Y_t and Y_t^n on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \mathbf{Q})$ by setting $Y = \Phi\Big((\widetilde{U}_j)_{j\geq 1}, (\widetilde{V}_j)_{j\geq 1}\Big)$ and (for each $n \geq 1$) $Y^n = \Phi\Big((\widetilde{U}_j^n)_{j\geq 1}, (\widetilde{V}_j^n)_{j\geq 1}\Big)$. Next, for each $i \in [k]$, we define

$$\mathcal{I}_i^{\leftarrow}(\Delta) = \max\{j \geq 0 :\ \widetilde{U}_1 + \cdots \widetilde{U}_j \leq t_i - \Delta\}, \qquad \mathcal{I}_i^{\rightarrow}(\Delta) = \min\{j \geq 0 :\ \widetilde{U}_1 + \cdots \widetilde{U}_j \geq t_i + \Delta\}.$$

That is, $\mathcal{I}_i^{\leftarrow}(\Delta)$ is the index of the last jump in Y_s before time $t_i - \Delta$, and $\mathcal{I}_i^{\rightarrow}(\Delta)$ is the index of the first jump after time $t_i + \Delta$. Recall that we have fixed $0 < t_1 < \cdots < t_k = t < \infty$. On event

$$A_n = \Big\{ \sum_{i=1}^{j} \widetilde{U}_i \notin \bigcup_{l\in[k]} [t_l - \Delta, t_l + \Delta] \ \forall j \leq J \Big\} \cap \Big\{ \sum_{j=1}^{J} \widetilde{U}_j > t,\ \sum_{j=1}^{J} \widetilde{U}_j^n > t \Big\},$$

we have $\mathcal{I}_i^{\rightarrow}(\Delta) = \mathcal{I}_i^{\leftarrow}(\Delta) + 1 \leq J$ for all $i \in [k]$. Then, on A_n it holds $\mathbf{Q}$-a.s. that (for all $i \in [k]$)

$$\lim_{n\to\infty} \sum_{j=1}^{\mathcal{I}_i^{\leftarrow}(\Delta)} \widetilde{U}_j^n = \sum_{j=1}^{\mathcal{I}_i^{\leftarrow}(\Delta)} \widetilde{U}_j \leq t_i - \Delta, \qquad \lim_{n\to\infty} \sum_{j=1}^{\mathcal{I}_i^{\leftarrow}(\Delta)+1} \widetilde{U}_j^n = \sum_{j=1}^{\mathcal{I}_i^{\leftarrow}(\Delta)+1} \widetilde{U}_j \geq t_i + \Delta,$$

As a result, on A_n it holds for all n large enough that $\sum_{j=1}^{\mathcal{I}_i^{\leftarrow}(\Delta)} \widetilde{U}_j^n < t_i$ and $\sum_{j=1}^{\mathcal{I}_i^{\leftarrow}(\Delta)+1} \widetilde{U}_j^n > t_i$ for all $i \in [k]$, implying that $Y_{t_i}^n = \widetilde{V}_{\mathcal{I}_i^{\leftarrow}(\Delta)}$ $\forall i \in [k]$. Furthermore, due to $\widetilde{V}_j^n \to \widetilde{V}_j$ $\mathbf{Q}$-a.s. for all $j \leq J$, it holds $\mathbf{Q}$-a.s. that $\lim_{n\to\infty} |\widetilde{V}_{I_i^{\leftarrow}(\Delta)}^n - \widetilde{V}_{I_i^{\leftarrow}(\Delta)}| \leq \lim_{n\to\infty} \max_{j\leq J} |\widetilde{V}_j^n - \widetilde{V}_j| = 0$. Therefore, on A_n it holds $\mathbf{Q}$-a.s. that $\lim_{n\to\infty} Y_{t_i}^n = \lim_{n\to\infty} \widetilde{V}_{I_i^{\leftarrow}(\Delta)}^n = \widetilde{V}_{I_i^{\leftarrow}(\Delta)} = Y_{t_i}$ for all $i \in [k]$. Then, for any $g : \mathbb{R}^k \to \mathbb{R}$ that is bounded and continuous, note that (let $\boldsymbol{Y}^n = (Y_{t_1}^n, \cdots, Y_{t_k}^n)$, $\boldsymbol{Y} = (Y_{t_1}, \cdots, Y_{t_k})$, and $\|g\| = \sup_{\boldsymbol{y}\in\mathbb{R}^k} |g(\boldsymbol{y})|$)

$$\begin{aligned}
\limsup_{n\to\infty} \Big|\mathbf{E}g(\boldsymbol{Y}^n) - \mathbf{E}g(\boldsymbol{Y})\Big| &\leq \limsup_{n\to\infty} \mathbf{E}_{\mathbf{Q}}\Big|g(\boldsymbol{Y}^n) - g(\boldsymbol{Y})\Big| \\
&= \limsup_{n\to\infty} \mathbf{E}_{\mathbf{Q}}\Big|g(\boldsymbol{Y}^n) - g(\boldsymbol{Y})\Big|\mathbf{I}_{A_n} + \limsup_{n\to\infty} \mathbf{E}_{\mathbf{Q}}\Big|g(\boldsymbol{Y}^n) - g(\boldsymbol{Y})\Big|\mathbf{I}_{(A_n)^c} \\
&\leq 0 + 2\,\|g\| \limsup_{n\to\infty} \mathbf{Q}\big((A_n)^c\big) \qquad \text{due to } \boldsymbol{Y}^n \xrightarrow{\mathbf{Q}-a.s.} \boldsymbol{Y} \text{ on } A_n \\
&\leq 2\,\|g\| \cdot \Bigg(\limsup_{n\to\infty} \mathbf{Q}(\sum_{i=1}^{J} \widetilde{U}_j \leq t) + \limsup_{n\to\infty} \mathbf{Q}(\sum_{i=1}^{J} \widetilde{U}_j^n \leq t) \\
&\qquad + \limsup_{n\to\infty} \mathbf{Q}\Big(\sum_{i=1}^{j} \widetilde{U}_i \in \bigcup_{l\in[k]} [t_l - \Delta, t_l + \Delta] \text{ for some } j \leq J \Big) \Bigg) \\
&\leq 6\,\|g\| \cdot \epsilon.
\end{aligned}$$

The last inequality follows from our choice of $J = J(\epsilon)$ and $\Delta = \Delta(\epsilon)$ at the beginning. From the arbitrariness of the mapping g and $\epsilon > 0$, we conclude the proof using Portmanteau theorem. □

B Proof of Theorems 3.2 and 3.3

In this section, we apply the framework developed in Section A and prove Theorems 3.2 and 3.3. In particular, the verification of part (i) of Condition 1 hinges on the chioce of the approximator $\hat{Y}_t^{\eta,\epsilon}$. Here, we construct a process $\hat{X}_t^{\eta,\epsilon|b}(x)$ as follows. Let $\hat{\tau}_0^{\eta,\epsilon|b}(x) \triangleq 0$,

$$\hat{\tau}_1^{\eta,\epsilon|b}(x) \triangleq \min\Big\{ j \geq 0 :\ X_j^{\eta|b}(x) \in \bigcup_{i\in[n_{\min}]} (m_i - \epsilon, m_i + \epsilon) \Big\}, \tag{B.1}$$

and

$$\hat{\mathcal{I}}_1^{\eta,\epsilon|b}(x) \triangleq i \iff X^{\eta|b}_{\hat{\tau}_1^{\eta,\epsilon|b}(x)}(x) \in I_i. \tag{B.2}$$

For $k \geq 2$,

$$\hat{\tau}_k^{\eta,\epsilon|b}(x) \triangleq \min\Big\{ j \geq \hat{\tau}_{k-1}^{\eta,\epsilon|b}(x) : \ X_j^{\eta|b}(x) \in \bigcup_{i \neq \hat{\mathcal{I}}_{k-1}^{\eta,\epsilon|b}(x)} (m_i - \epsilon, m_i + \epsilon) \Big\} \qquad \forall k \geq 2. \tag{B.3}$$

and

$$\hat{\mathcal{I}}_k^{\eta,\epsilon|b}(x) \triangleq i \iff X^{\eta|b}_{\hat{\tau}_k^{\eta,\epsilon|b}(x)}(x) \in I_i. \tag{B.4}$$

Essentially, $\hat{\tau}_k^{\eta,\epsilon|b}(x)$ records the k-th time $X_j^{\eta|b}(x)$ visits (the ϵ-neighborhood of) a local minimum and $\hat{\mathcal{I}}_k^{\eta,\epsilon|b}(x)$ denotes the index of the visited local minimum. Let

$$\hat{X}_\cdot^{\eta,\epsilon|b}(x) \triangleq \Phi\Big(\Big((\hat{\tau}_k^{\eta,\epsilon|b}(x) - \hat{\tau}_{k-1}^{\eta,\epsilon|b}(x)) \cdot \lambda_b^*(\eta) \Big)_{k\geq 1}, \big(m_{\hat{\mathcal{I}}_k^{\eta,\epsilon|b}(x)} \big)_{k \geq 1} \Big).$$

By definition, $\hat{X}_t^{\eta,\epsilon|b}(x)$ keeps track of how $X_j^{\eta|b}(x)$ traverses the potential U and makes transitions between the different local minima, under a time scaling of $\lambda_b^*(\eta)$.

Using Lemma A.5, the convergence of $\hat{X}_\cdot^{\eta,\epsilon|b}(x)$ follows directly from the convergence of $\hat{\tau}_k^{\eta,\epsilon|b}(x) - \hat{\tau}_{k-1}^{\eta,\epsilon|b}(x)$ and $m_{\hat{\mathcal{I}}_k^{\eta,\epsilon|b}(x)}$, i.e., the inter-arrival times and destinations of the transitions in $X_j^{\eta|b}(x)$ between different local minima over the potential U. This is exactly the content of the first exit time analysis. In particular, based on a straightforward adaptation of the first exit time analysis in Section 2.2 to the current setup, we obtain Proposition B.1.

Proposition B.1. *Let Assumptions 1, 2, 3, 6, and 7 hold. Let $i \in [n_{min}]$ and $x \in I_i$. For any $\epsilon > 0$ small enough, the following claims hold.*

(i) $\{\hat{X}_t^{\eta,\epsilon|b}(x) : \ t > 0\} \overset{f.d.d.}{\to} \{Y_t^{*|b} : \ t > 0\}$ *as* $\eta \downarrow 0$*;*

(ii) Given any $T \in (0,\infty)$, $p \in [1,\infty)$, and any sequence of strictly positive reals η_n's such that $\lim_{n\to\infty} \eta_n = 0$, the laws of $\hat{X}_\cdot^{\eta_n,\epsilon|b}$ are tight in $(\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$.

Proposition B.2 then verifies part (ii) of Condition 1 in Lemma A.3, under the choice of $Y_t^\eta = X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta) \rfloor}(x)$ and $\hat{Y}_t^{\eta,\epsilon} = \hat{X}_t^{\eta,\epsilon|b}(x)$. We give the proof in Section C.

Proposition B.2. *Let Assumptions 1, 2, 3, 6, and 7 hold. Let $x \in \bigcup_{i\in[n_{min}]} I_i$. Given any $T > 0$ and $p \in [1,\infty)$, it holds for all $\epsilon > 0$ small enough that*

$$\lim_{\eta\downarrow 0} \mathbf{P}\left(\boldsymbol{d}_{L_p}^{[0,T]}\Big(X^{\eta|b}_{\lfloor \cdot/\lambda_b^*(\eta)\rfloor}(x), \hat{X}_\cdot^{\eta,\epsilon|b}(x) \Big) \geq 2\epsilon \right) = 0, \qquad \lim_{\eta\downarrow 0} \mathbf{P}\Big(\Big| X_T^{\eta|b}(x) - \hat{X}_T^{\eta,\epsilon|b}(x) \Big| \geq \epsilon \Big) = 0.$$

Now, we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. Fix some $i \in [n_{\min}]$ and $x \in I_i$. From Lemma A.2 and Proposition B.1, we verify part (i) of Condition 1, i.e., given any $T > 0$, the claim

$$\{\hat{X}_t^{\eta,\epsilon|b}(x) : \ t > 0\} \overset{f.d.d.}{\to} \{Y_t^{*|b} : \ t > 0\} \quad \text{and} \quad \hat{X}_\cdot^{\eta,\epsilon|b}(x) \Rightarrow Y_\cdot^{*|b} \text{ in } (\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]}) \text{ as } \eta \downarrow 0$$

holds for all $\epsilon > 0$ small enough. Meanwhile, given any $T \in (0,\infty)$ and $p \in [1,\infty)$, Proposition B.2 verifies part (ii) of Condition 1 under the choice of $Y_t^\eta = X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta)\rfloor}(x)$, $\hat{Y}_t^{\eta,\epsilon} = \hat{X}_t^{\eta,\epsilon|b}(x)$, and $Y_t^* = Y_t^{*|b}$. Applying Lemma A.3, we obtain that (for any $T \in (0,\infty)$ and $p \in [1,\infty)$)

$$\{X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta)\rfloor}(x) :\ t>0\} \overset{f.d.d.}{\to} \{Y_t^{*|b} :\ t>0\} \quad \text{and} \quad X^{\eta|b}_{\lfloor \cdot/\lambda_b^*(\eta)\rfloor}(x) \Rightarrow Y_\cdot^{*|b} \text{ in } (\mathbb{D}[0,T], \boldsymbol{d}_{L_p}^{[0,T]})$$

as $\eta \downarrow 0$. This allows us to conclude the proof using Lemma A.1. □

To conclude, Theorem 3.3 follows almost immediately from Theorem 3.2.

Proof of Theorem 3.3. For any $b > \max_{i\in[n_{\min}],\ j\in[n_{\min}-1]} |m_i - s_j|$, by definitions in (3.2) we have $\mathcal{J}_b^*(i,j) = \mathcal{J}_b^*(i) = 1$ for all $i \in [n_{\min}]$ and $j \in [n_{\min}-1]$. Therefore, for such $b > 0$ large enough, we also have $\lambda_b^*(\eta) = \eta \cdot \lambda(\eta) = H(\eta^{-1})$. Henceforth in this proof, we only consider such large b.

Pick some closed set $A \subseteq \mathbb{D}[0,T]$ (w.r.t. L_p topology), and observe that

$$\begin{aligned}
\mathbf{P}\Big(X^\eta_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \in A\Big) &= \mathbf{P}\Big(X^\eta_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \in A;\ X_j^{\eta|b}(x) = X_j^\eta(x)\ \forall j \le \lfloor T/H(\eta^{-1})\rfloor\Big) \\
&\quad + \mathbf{P}\Big(X^\eta_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \in A;\ X_j^{\eta|b}(x) \neq X_j^\eta(x) \text{ for some } j \le \lfloor T/H(\eta^{-1})\rfloor\Big) \\
&\le \underbrace{\mathbf{P}\Big(X^{\eta|b}_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \in A\Big)}_{\text{(I)}} + \underbrace{\mathbf{P}\Big(X_j^{\eta|b}(x) \neq X_j^\eta(x) \text{ for some } j \le \lfloor T/H(\eta^{-1})\rfloor\Big)}_{\text{(II)}}.
\end{aligned} \tag{B.5}$$

For term (I), it follows from Theorem 3.2 that $\limsup_{\eta\downarrow 0}$ (I) $\le \mathbf{P}\big(Y_\cdot^{*|b}(m_i) \in A\big)$. For term (II), we make two observations. First, recall that $C \in [1,\infty)$ is the constant in Assumption 4 such that $\sup_{x\in\mathbb{R}} |a(x)| \vee \sigma(x) \le C$. Under any $\eta \in (0, \frac{b}{2C})$, on the event $\{\eta|Z_j| \le \frac{b}{2C}\ \forall j \le \lfloor T/H(\eta^{-1})\rfloor\}$ the step-size (before truncation) $\eta a\big(X_{j-1}^{\eta|b}(x)\big) + \eta\sigma\big(X_{j-1}^{\eta|b}(x)\big)Z_j$ of $X_j^{\eta|b}$ is less than b for each $j \le \lfloor T/H(\eta^{-1})\rfloor$. Therefore, $X_j^{\eta|b}(x)$ and $X_j^\eta(x)$ coincide for such j's. In other words, for any $\eta \in (0, \frac{b}{2C})$, we have $\{\eta|Z_j| \le \frac{b}{2C}\ \forall j \le \lfloor T/H(\eta^{-1})\rfloor\} \subseteq \{X_j^{\eta|b}(x) = X_j^\eta(x)\ \forall j \le \lfloor T/H(\eta^{-1})\rfloor\}$. which leads to (recall that $H(\cdot) = \mathbf{P}(|Z_1| > \cdot)$)

$$\begin{aligned}
\limsup_{\eta\downarrow 0} \text{(II)} &\le \limsup_{\eta\downarrow 0} \mathbf{P}\left(\exists j \le \lfloor T/H(\eta^{-1})\rfloor \ s.t.\ \eta|Z_j| > \frac{b}{2C}\right) \\
&\le \limsup_{\eta\downarrow 0} \frac{T}{H(\eta^{-1})} \cdot H(\eta^{-1}\cdot\frac{b}{2C}) = T\cdot\left(\frac{2C}{b}\right)^\alpha \qquad \text{due to } H(x) \in \mathcal{RV}_{-\alpha}(x).
\end{aligned}$$

In summary, $\limsup_{\eta\downarrow 0} \mathbf{P}\big(X^\eta_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \in A\big) \le \mathbf{P}\big(Y_\cdot^{*|b}(m_i) \in A\big) + T\cdot\left(\frac{2C}{b}\right)^\alpha$. Furthermore, note that for all b large enough, we have $q_b(i,j) = q(i,j)$ for all $i,j \in [n_{\min}]$ with $i \neq j$. To see why, we fix some $i,j \in [n_{\min}]$ with $i \neq j$. For all b large enough, we have $\mathcal{J}_b^*(i,j) = 1$, and hence (see (3.10) and (3.12) for definitions of $q_b(i,j)$ and $q(i,j)$)

$$q(i,j) = \nu_\alpha\Big(\big\{w\in\mathbb{R} :\ m_i + \sigma(m_i)\cdot w \in I_j\big\}\Big),\quad q_b(i,j) = \nu_\alpha\Big(\big\{w\in\mathbb{R} :\ m_i + \varphi_b\big(\sigma(m_i)\cdot w\big) \in I_j\big\}\Big).$$

Suppose that I_j has bounded support (i.e., $j = 2,3,\cdots,n_{\min}-1$ so that I_j is not the leftmost or the rightmost attraction field), then it holds for all b large enough that $m_i - b \notin I_j$ and $m_i + b \notin I_j$. Under such large b, for $m_i + \varphi_b\big(\sigma(m_i)\cdot w\big) \in I_j$ to hold we must have $|\sigma(m_i)\cdot w| < b$, thus implying $m_i + \varphi_b\big(\sigma(m_i)\cdot w\big) = m_i + \sigma(m_i)\cdot w$ and hence $q_b(i,j) = q(i,j)$. Next, consider the case where $j = 1$ so $I_j = I_1 = (-\infty, s_1)$ is the leftmost attraction field. For any b large enough we must have $m_i - z \in (-\infty, s_1) = I_1$ for all $z \ge b$. This also implies $m_i + \varphi_b\big(\sigma(m_i)\cdot w\big) \in I_1 \iff m_i + \sigma(m_i)\cdot w \in I_1$. The same argument can be applied to the case with $j = n_{\min}$ (that is, $I_j = (s_{n_{\min}-1}, \infty)$ is the rightmost attraction field).

Now that we know $q_b(i,j) = q(i,j)$ for all b large enough, the claim $Y_t^{*|b}(m_i) = Y_t^*(m_i)\ \forall t \geq 0$ must hold for all b large enough as both CTMCs have the same infinitesimal generator. Therefore, $\lim_{b\to\infty} \mathbf{P}\big(Y_\cdot^{*|b}(m_i) \in A\big) = \mathbf{P}\big(Y_\cdot^*(m_i) \in A\big)$. Together with the fact that $\lim_{b\to\infty} \big(\frac{2C}{b}\big)^\alpha = 0$, in (B.5) we obtain $\limsup_{\eta\downarrow 0} \mathbf{P}\big(X^\eta_{\lfloor \cdot/H(\eta^{-1})\rfloor}(x) \in A\big) \leq \mathbf{P}\big(Y_\cdot^*(m_i) \in A\big)$. From the arbitrariness of the closed set A, we conclude the proof with Portmanteau theorem. □

C Proof of Propositions B.1 and B.2

This section is devoted to proving Propositions B.1 and B.2. Henceforth in Section C, we fix some $b \in (0,\infty)$ be such that Assumption 7 holds. In particular,

$$|s_j - m_i|/b \notin \mathbb{Z} \qquad \forall i \in [n_{\min}],\ j \in [n_{\min} - 1]. \tag{C.1}$$

This allows us to fix some $\bar\epsilon \in (0, 1 \wedge b)$ such that

$$l_i > \big(\mathcal{J}_b^*(i) - 1\big)b + 3\bar\epsilon \text{ and } [m_i - \bar\epsilon, m_i + \bar\epsilon] \subseteq [s_{i-1} + \bar\epsilon, s_i - \bar\epsilon] \qquad \forall i \in [n_{\min}] \tag{C.2}$$

with l_i and $\mathcal{J}_b^*(i)$ defined in (3.1) and (3.2), respectively. In other words, we fix some $\bar\epsilon$ small enough such that, even with $\bar\epsilon$-shrinkage, the number of jumps required to exit from (ϵ-shrunken) I_i remains $\mathcal{J}_b^*(i)$.

We start by highlighting a few properties of the limiting Markov jump process $Y^{*|b}$ in Theorem 3.2, using results for the measure $\check{\mathbf{C}}^{(k)|b}$ collected in Section E. Recall the definitions of $q_b(i)$ and $q_b(i,j)$ in (3.10). First, by definition,

$$q_b(i) = \sum_{j\in[n_{\min}]:\ j\neq i} q_b(i,j) + \sum_{j\in[n_{\min}-1]} \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(\{s_j\}; m_i).$$

From (C.2), we have $|s_j - m_i| > (\mathcal{J}_b^*(i) - 1)\cdot b + \bar\epsilon$. Then, by applying Lemma E.1, we get $\sum_{j\in[n_{\min}-1]} \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(\{s_j\}; m_i) = 0$. Together with Lemma E.2, we yield that

$$\sum_{j\in[n_{\min}]:\ j\neq i} q_b(i,j) = q_b(i) \in (0,\infty). \tag{C.3}$$

Furthermore, Lemma E.3 verifies that

$$q_b(i,j) > 0 \qquad \Longleftrightarrow \qquad \mathcal{J}_b^*(i,j) = \mathcal{J}_b^*(i). \tag{C.4}$$

As a result, in Definition 3.1 we know that the typical transition graph associated with threshold b contains an edge $(m_i \to m_j)$ if and only if $q_b(i,j) > 0$.

Next, we stress that the law of the Markov jump process $Y^{*|b}$ can be expressed using the mapping Φ introduced in Definition A.4. Given any $m_{\text{init}} \in \{m_1, m_2, \ldots, m_{n_{\min}}\}$, we set $V_1 = m_{\text{init}}$, $U_1 = 0$, and (for any $t > 0$, $l \geq 1$, and $i, j \in [n_{\min}]$ with $i \neq j$)

$$\begin{aligned}
&\mathbf{P}\Big(U_{l+1} < t,\ V_{l+1} = m_j \ \Big|\ V_l = m_i, (V_j)_{j=1}^{l-1},\ (U_j)_{j=1}^l\Big) = \mathbf{P}\Big(U_{l+1} < t,\ V_{l+1} = m_j \ \Big|\ V_l = m_i\Big) \\
&= \begin{cases} \frac{q_b(i,j)}{q_b(i)} & \text{if } m_i \notin V_b^*, \\ \frac{q_b(i,j)}{q_b(i)} \cdot \Big(1 - \exp\big(-q_b(i)t\big)\Big) & \text{if } m_i \in V_b^*. \end{cases}
\end{aligned} \tag{C.5}$$

In other words, conditioning on $V_l = m_i$, we have $V_{l+1} = m_j$ with probability $q_b(i,j)/q_b(i)$; as for U_{l+1}, we set $U_{l+1} \equiv 0$ if $m_i \notin V_b^*$ (i.e., the current value m_i is not a widest minimum), and set U_{l+1} as an Exponential RV with rate $q_b(i)$ otherwise. We claim that

$$Y_\cdot^{*|b} \overset{d}{=} \Phi\Big((U_j)_{j\geq 1}, (V_j)_{j\geq 1}\Big). \tag{C.6}$$

In fact, under the conditions in Theorem 3.2, it is straightforward to show that

(*i*) For any $t > 0$, $\lim_{i\to\infty} \mathbf{P}(\sum_{j\le i} U_j > t) = 1$;

(*ii*) For any $u > 0$ and $i \ge 1$, $\mathbf{P}(U_1 + \cdots + U_i = u) = 0$;

(*iii*) $Y_\cdot^{*|b} \overset{d}{=} \Phi\Big((U_j)_{j\ge1}, (V_j)_{j\ge1}\Big)$; that is, it is a continuous-time Markov chain with state space V_b^*, generator

$$\mathbf{P}(Y_{t+h}^{*|b} = m_j \mid Y_t^{*|b} = m_i) = h \cdot \sum_{j'\in[n_{\min}]:\ j'\neq i} q_b(i,j')\theta_b(m_j|m_{j'}) + \boldsymbol{o}(h) \qquad \text{as } h \downarrow 0,$$

and initial distribution $\mathbf{P}(Y_0^{*|b} = m_j) = \theta_b(m_j|m_{\text{init}})$; see (3.10) and (3.11) for the definitions of $q_b(i,j)$ and $\theta_b(m_j|m_i)$, respectively.

For the sake of completeness, we collect the proof in Section D. The representation (C.6) and the properties stated above will greatly facilitate the proofs below.

The proofs of Propositions B.1 and B.2 hinge on the first exit analysis in Result 2. Note that Result 2 focuses on some bounded interval I. In contrast, regarding the potential U characterized in Assumption 6, while for all $i = 2, \cdots, n_{\min}$ the attraction field I_i is indeed bounded, for $i = 1$ or $n_{\min}$ (that is, the leftmost or the rightmost attraction field) we have $I_1 = (-\infty, s_1)$ and $I_{n_{\min}} = (s_{n_{\min}-1}, \infty)$, both of which are unbounded. Besides, our analysis below involves $S(\delta) \triangleq \bigcup_{i\in[n_{\min}-1]}[s_i - \delta, s_i + \delta]$ (i.e., the union of the δ-neighborhood of any boundary point s_i). As a result, we will frequently consider sets of form

$$I_{i;\delta,M} = (s_{i-1} + \delta, s_i - \delta) \cap (-M, M) = (I_i)_\delta \cap (-M, M) \tag{C.7}$$

for some $\delta, M > 0$. For any $M > 0$ large enough such that $-M < m_1 < s_1 < \cdots < s_{n_{\min}-1} < m_{n_{\min}} < M$, we have $I_{i;\delta,M} = (s_{i-1} + \delta, s_i - \delta) \cap (-M, M) = (s_{i-1} + \delta, s_i - \delta)$ for all $i = 2, 3, \cdots, n_{\min} - 1$, and we have $I_{1;\delta,M} = (s_0 + \delta, s_1 - \delta) \cap (-M, M) = (-M, s_1 - \delta)$ (due to $s_0 = -\infty$) and $I_{n_{\min};\delta,M} = (s_{n_{\min}-1} + \delta, s_{n_{\min}} - \delta) \cap (-M, M) = (s_{n_{\min}-1} + \delta, M)$ (due to $s_{n_{\min}} = \infty$). We also set

$$\sigma_{i;\epsilon}^{\eta|b}(x) \triangleq \min\Big\{j \ge 0 :\ X_j^{\eta|b}(x) \in \bigcup_{l\neq i}(m_l - \epsilon, m_l + \epsilon)\Big\}, \tag{C.8}$$

$$\tau_{i;\delta,M}^{\eta|b}(x) \triangleq \min\Big\{j \ge 0 :\ X_j^{\eta|b}(x) \notin I_{i;\delta,M}\Big\}. \tag{C.9}$$

In other words, $\tau_{i;\delta,M}^{\eta|b}(x)$ is the first exit time from $I_{i;\delta,M}$, and $\sigma_{i;\epsilon}^{\eta|b}(x)$ is the first time visiting the ϵ-neighborhood of a local minimum different from m_i.

We first state Lemmas C.1 and C.2. To give an overview, we establish these two lemmas by adapting the first exit analysis in Section 2.2 to the slightly more general settings in Propositions B.1 and B.2. First, Lemma C.1 states that it is unlikely to get close to any of the boundary points s_i's or exit a wide enough compact set.

Lemma C.1. *Let Assumptions 1, 2, 3, 4, and 6 hold. Let $b \in (0, \infty)$ be such that (C.1) holds. There exists $M > 0$ such that*

$$\max_{i\in[n_{min}]} \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\big((-M, M)^c; m_i\big) = 0, \tag{C.10}$$

Furthermore, given any $\Delta > 0$ and any $\epsilon \in (0, \bar{\epsilon})$ (with $\bar{\epsilon}$ specified in (C.2)), it holds for all $\delta > 0$ small enough that

$$\limsup_{\eta\downarrow0} \max_{i\in[n_{min}]} \sup_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\Big(\exists j < \sigma_{i;\epsilon}^{\eta|b}(x) \text{ s.t. } X_j^{\eta|b}(x) \in S(\delta) \text{ or } \big|X_j^{\eta|b}(x)\big| \ge M + 1\Big) < \Delta. \tag{C.11}$$

Proof. In light of Lemma E.3, it holds for all $M > 0$ large enough such that

$$\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}((-M,M)^c; m_i) = 0 \qquad \forall i \in [n_{\min}].$$

This concludes the proof of (C.10).

Henceforth in this proof, we fix such large M satisfying $|M - m_i|/b \notin \mathbb{Z}\ \forall i \in [n_{\min}]$ and $M > \max_{i\in[n_{\min}]}(\mathcal{J}_b^*(i) - 1)b + \bar{\epsilon}$, where $\bar{\epsilon} > 0$ is the constant in (C.2). Also, we fix some $\epsilon \in (0, \bar{\epsilon})$ and show that (C.11) holds for such ϵ. Recall the definition of $\tau^{\eta|b}_{i;\delta,M}(x)$ in (C.9) and $I_{i;\delta,M} = (s_{i-1} + \delta, s_i - \delta) \cap (-M, M)$. We make a few observations regarding the stopping time $\tau^{\eta|b}_{i;2\delta,M}(x) = \min\{j \geq 0 :\ X^{\eta|b}_j(x) \notin I_{i;2\delta,M}\}$. First, due to $I_{i;2\delta,M} \subseteq I_{i;\delta,M}$, we must have $\tau^{\eta|b}_{i;2\delta,M}(x) \leq \tau^{\eta|b}_{i;\delta,M}(x) \leq \sigma^{\eta|b}_{i;\epsilon}(x)$. Second, by definition of $\tau^{\eta|b}_{i;2\delta,M}(x)$, we have $X^{\eta|b}_j(x) \notin S(\delta),\ |X^{\eta|b}_j(x)| < M$ for all $j < \tau^{\eta|b}_{i;2\delta,M}(x)$. On event

$$A_0(\eta,\delta,x) \triangleq \{X^{\eta|b}_{\tau^{\eta|b}_{i;2\delta,M}(x)}(x) \in (-M,M);\ X^{\eta|b}_{\tau^{\eta|b}_{i;2\delta,M}(x)}(x) \notin S(2\delta)\},$$

there exists some $j \in [n_{\min}],\ j \neq i$ such that $X^{\eta|b}_{\tau^{\eta|b}_{i;2\delta,M}(x)}(x) \in I_{j;2\delta,M}$. Now define

$$A_1(\eta,\delta,x) \triangleq \{\exists j < \sigma^{\eta|b}_{i;\epsilon}(x)\ s.t.\ X^{\eta|b}_j(x) \in S(\delta)\},\ A_2(\eta,x) \triangleq \{\exists j < \sigma^{\eta|b}_{i;\epsilon}(x)\ s.t.\ |X^{\eta|b}_j(x)| \geq M+1\}.$$

Let $R^{\eta|b}_{j;\epsilon}(x) \triangleq \min\{k \geq 0 :\ X^{\eta|b}_k(x) \in (m_j - \epsilon, m_j + \epsilon)\}$ be the first time entering $(m_j - \epsilon, m_j + \epsilon)$. From the strong Markov property at $\tau^{\eta|b}_{i;2\delta,M}(x)$,

$$\begin{aligned}
&\max_{i\in[n_{\min}]} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\Big(\big(A_1(\eta,\delta,x)\cup A_2(\eta,x)\big)\cap A_0(\eta,\delta,x)\Big)\\
&\leq \max_{i\in[n_{\min}]} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\Big(A_1(\eta,\delta,x)\cup A_2(\eta,x)\ \Big|\ A_0(\eta,\delta,x)\Big)\\
&\leq \max_{j\in[n_{\min}]} \underbrace{\sup_{y\in[s_{j-1}+2\delta,s_j-2\delta]\cap[-M,M]} \mathbf{P}\Big(\Big\{X^{\eta|b}_k(x) \in [s_{j-1}+\delta, s_j-\delta]\cap(-M-1,M+1)\ \forall\exists k < R^{\eta|b}_{j;\epsilon}(x)\Big\}^c\Big)}_{p_j(\eta)}.
\end{aligned}$$

For any $j \in [n_{\min}]$ and any $\delta > 0$ small enough, by applying Lemma E.5 onto $I_j \cap (-M-1, M+1)$ (with parameter ϵ therein set as 2δ) we get $\lim_{\eta\downarrow 0} p_j(\eta) = 0$. In summary, we have shown that $\limsup_{\eta\downarrow 0} \max_{i\in[n_{\min}]} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\big((A_1(\eta,\delta,x)\cup A_2(\eta,x))\cap A_0(\eta,\delta,x)\big) = 0$. Therefore, to establish (C.11), it only remains to show that $\limsup_{\eta\downarrow 0} \max_{i\in[n_{\min}]} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\big(\big(A_0(\eta,\delta,x)\big)^c\big) < \Delta$. Now, it only remains to prove that for all $\delta > 0$ small enough,

$$\limsup_{\eta\downarrow 0} \max_{i\in[n_{\min}]} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\Big(X^{\eta|b}_{\tau^{\eta|b}_{i;2\delta,M}(x)}(x) \in S(2\delta)\Big) < \Delta, \tag{C.12}$$

$$\limsup_{\eta\downarrow 0} \max_{i\in[n_{\min}]} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\Big(X^{\eta|b}_{\tau^{\eta|b}_{i;2\delta,M}(x)}(x) \notin (-M,M)\Big) = 0. \tag{C.13}$$

Note that

$$\limsup_{\eta\downarrow 0} \sup_{x\in[m_i-\epsilon,m_i+\epsilon]} \mathbf{P}\Big(X^{\eta|b}_{\tau^{\eta|b}_{i;2\delta,M}(x)}(x) \in S(2\delta)\Big) \leq \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\Big(S(2\delta); m_i\Big)\Big/\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\Big(I^c_{i;2\delta,M}; m_i\Big)$$

can be established using part (a) of Result 2. From Lemma E.1, we get $\lim_{\delta\downarrow 0} \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\big(S(2\delta); m_i\big) = \check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\big(\{s_1, \cdots, s_{n_{\min}}\}; m_i\big) = 0$ and verifies claim (C.12). Similarly, claim (C.13) follows directly from part (a) of Result 2 when applied onto $I_{i;\delta,M}$, combined with (C.10). □

Recall the scale function λ_b^* defined in (3.8). Lemma C.2 then provides an analogue of Result 2 for the current setup.

Lemma C.2. *Let Assumptions 1, 2, 3, 4 and 6 hold. Let $b \in (0,\infty)$ be such that (C.1) holds. Let $\bar{\epsilon} > 0$ be specified as in (C.2).*

(i) Let $R^{\eta|b}_{i;\epsilon}(x) \triangleq \min\{j \geq 0 :\ X^{\eta|b}_j(x) \in (m_i - \epsilon, m_i + \epsilon)\}$. For any $\epsilon \in (0,\bar{\epsilon})$, $t > 0$ and $i \in [n_{min}]$,

$$\liminf_{\eta\downarrow 0} \inf_{x\in[s_{i-1}+\epsilon, s_i-\epsilon]} \mathbf{P}\left(R^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_b(\eta) \leq t,\ X^{\eta|b}_j(x) \in I_i\ \forall j \leq R^{\eta|b}_{i;\epsilon}(x)\right) = 1.$$

(ii) Let $i, j \in [n_{min}]$ be such that $i \neq j$. Let $\sigma^{\eta|b}_{i;\epsilon}(x) \triangleq \min\{j \geq 0 :\ X^{\eta|b}_j(x) \in \bigcup_{l\neq i}(m_l - \epsilon, m_l + \epsilon)\}$. If $m_i \in V^*_b$, then for any $\epsilon \in (0,\bar{\epsilon})$ and any $t > 0$,

$$\liminf_{\eta\downarrow 0} \inf_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\left(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_b(\eta) > t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x) \in I_j\right) \geq \exp\big(-q_b(i)\cdot t\big)\cdot\frac{q_b(i,j)}{q_b(i)},$$
$$\limsup_{\eta\downarrow 0} \sup_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\left(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_b(\eta) > t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x) \in I_j\right) \leq \exp\big(-q_b(i)\cdot t\big)\cdot\frac{q_b(i,j)}{q_b(i)}.$$

If $m_i \notin V^*_b$, then for any $\epsilon \in (0,\bar{\epsilon})$ and any $t > 0$,

$$\begin{aligned}\frac{q_b(i,j)}{q_b(i)} &\leq \liminf_{\eta\downarrow 0} \inf_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\left(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_b(\eta) \leq t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x) \in I_j\right)\\ &\leq \limsup_{\eta\downarrow 0} \sup_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\left(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_b(\eta) \leq t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x) \in I_j\right) \leq \frac{q_b(i,j)}{q_b(i)}.\end{aligned}$$

Proof. (i) Fix some $t > 0$ and $\epsilon \in (0,\bar{\epsilon})$. Recall that $\lambda^*_b(\eta) \in \mathcal{RV}_{\mathcal{J}^*_b(V)\cdot(\alpha-1)+1}(\eta)$ as $\eta \downarrow 0$. Due to $\mathcal{J}^*_b(V) \geq 1$ and $\alpha > 1$, we have $\mathcal{J}^*_b(V)\cdot(\alpha-1)+1 \geq \alpha > 1$. This implies that $\lim_{\eta\downarrow 0} \frac{T/\eta}{t/\lambda^*_b(\eta)} = 0\ \forall T > 0$, and hence (given any $T > 0$)

$$\mathbf{P}\big(R^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_b(\eta) \leq t,\ X^{\eta|b}_j(x) \in I_i\ \forall j \leq R^{\eta|b}_{i;\epsilon}(x)\big) \geq \mathbf{P}\big(R^{\eta|b}_{i;\epsilon}(x) \leq T/\eta,\ X^{\eta|b}_j(x) \in I_i\ \forall j \leq R^{\eta|b}_{i;\epsilon}(x)\big)$$

for all η small enough. Now, pick $M > 0$ large enough such that $|M| > \min\{|s_{i-1}+\epsilon|, |s_i-\epsilon|\}$. By picking $T > 0$ large enough, one can apply Lemma E.5 onto $(-M, M)\cap I_j$ to conclude the proof of part (i).

(ii) Let $\lambda^*_{i;b}(\eta) \triangleq \eta\cdot\lambda^{\mathcal{J}^*_b(i)}(\eta)$. It suffices to establish the following upper and lower bounds: for all $i, j \in [n_{\min}]$ such that $i \neq j$, all $\epsilon \in (0,\bar{\epsilon})$, and all $t > 0$,

$$\liminf_{\eta\downarrow 0} \inf_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\big(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_{i;b}(\eta) > t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x) \in I_j\big) \geq \exp\big(-q_b(i)\cdot t\big)\cdot\frac{q_b(i,j)}{q_b(i)}, \tag{C.14}$$
$$\limsup_{\eta\downarrow 0} \sup_{x\in[m_i-\epsilon, m_i+\epsilon]} \mathbf{P}\big(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_{i;b}(\eta) > t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x) \in I_j\big) \leq \exp\big(-q_b(i)\cdot t\big)\cdot\frac{q_b(i,j)}{q_b(i)}. \tag{C.15}$$

Indeed, in case that $m_i \in V^*_b$, claims in part (ii) are equivalent to (C.14) and (C.15) due to $\mathcal{J}^*_b(i) = \mathcal{J}^*_b(V)$ (see (3.4)) and hence $\lambda^*_{i;b}(\eta) = \lambda^*_b(\eta) = \eta\cdot\lambda^{\mathcal{J}^*_b(V)}(\eta)$. In case that $m_i \notin V^*_b$ (i.e., $\mathcal{J}^*_b(i) < \mathcal{J}^*_b(V)$), we have $\lim_{\eta\downarrow 0} \frac{t/\lambda^*_{i;b}(\eta)}{T/\lambda^*_b(\eta)} = 0$ for all $t, T \in (0,\infty)$. We then recover the upper and lower bounds in part (ii) by letting $t \downarrow 0$ in (C.14) and (C.15).

The rest of this proof is devoted to establishing (C.14) and (C.15). Here, we collect a few useful facts about the measure $\check{\mathbf{C}}^{(\mathcal{J}^*_b(i))|b}$. By assumption (C.1), one can apply Lemma E.1 and obtain $\check{\mathbf{C}}^{(\mathcal{J}^*_b(i))|b}\big(\{s_1, \cdots, s_{n_{\min}-1}\}; m_i\big) = 0$. Recall the definition of $q_b(i,j) = \check{\mathbf{C}}^{(\mathcal{J}^*_b(i))|b}(I_j; m_i)$ in (3.10). Due to $I_j = (s_{j-1}, s_j)$,

$$q_b(i,j) = \check{\mathbf{C}}^{(\mathcal{J}^*_b(i))|b}(I_j; m_i) = \check{\mathbf{C}}^{(\mathcal{J}^*_b(i))|b}(I^-_j; m_i) \qquad \forall i, j \in [n_{\min}] \text{ with } i \neq j. \tag{C.16}$$

Combining (C.16) with the continuity of measures, we have $\lim_{\delta\downarrow 0}\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\big((s_{i-1}-\delta,s_i+\delta)^c;m_i\big)=q_b(i)=\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_i^c;m_i)$. Next, throughout the remainder of this proof, we only consider $M\in(0,\infty)$ large enough such that the claim (C.10) of Lemma C.1 holds. Given any $\Delta>0$, regarding the set $I_{i;\delta,M}=(-M,M)\cap(s_{i-1}+\delta,s_i-\delta)$ it holds for all $\delta>0$ small enough that

$$\begin{aligned}\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_{i;\delta,M}^c;m_i)&\le\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}((-M,M)^c;m_i)+\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}((s_{i-1}+\delta,s_i-\delta)^c;m_i)\\&=0+\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}((s_{i-1}+\delta,s_i-\delta)^c;m_i)<(1+\Delta)\cdot q_b(i).\end{aligned}\tag{C.17}$$

Lastly, due to $I_{i;\delta,M}\subseteq I_i$,

$$\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\big(I_{i;\delta,M}^c;m_i\big)\ge\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}\big(I_i^c;m_i\big)=q_b(i).\tag{C.18}$$

Proof of Lower Bound (C.14).

We fix some $i\neq j$ and $t>0$ when proving (C.14). Recall the definitions of $I_{i;\delta,M}$ and $\tau_{i;\delta,M}^{\eta|b}(x)$ in (C.7) and (C.9), respectively. Observe that

$$\begin{aligned}&\big\{\sigma_{i;\epsilon}^{\eta|b}(x)\cdot\lambda_{i;b}^*(\eta)>t,\ X^{\eta|b}_{\sigma_{i;\epsilon}^{\eta|b}(x)}(x)\in I_j\big\}\\&\supseteq\underbrace{\big\{\tau_{i;\delta,M}^{\eta|b}(x)\cdot\lambda_{i;b}^*(\eta)>t;\ X^{\eta|b}_{\tau_{i;\delta,M}^{\eta|b}(x)}(x)\in I_{j;\delta,M+1}\big\}}_{\text{(I)}}\cap\underbrace{\big\{X^{\eta|b}_{\sigma_{i;\epsilon}^{\eta|b}(x)}(x)\in I_j\big\}}_{\text{(II)}}.\end{aligned}$$

We first analyze $\mathbf{P}(\text{(II)}|\text{(I)})$. By strong Markov property at $\tau_{i;\delta,M}^{\eta|b}(x)$, $\inf_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}(\text{(II)}\mid\text{(I)})\ge\inf_{y\in I_{j;\delta,M+1}}\mathbf{P}\big(X_k^{\eta|b}(y)\in I_j\ \forall k\le R_{j;\epsilon}^{\eta|b}(y)\big)$. Here, recall that $R_{j;\epsilon}^{\eta|b}(x)=\min\{j\ge 0:\ X_k^{\eta|b}(x)\in(m_j-\epsilon,m_j+\epsilon)\}$. Applying Lemma E.5, we yield

$$\liminf_{\eta\downarrow 0}\inf_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}(\text{(II)}\mid\text{(I)})=1.\tag{C.19}$$

Next, we move onto the analysis of $\mathbf{P}(\text{(I)})$. Due to $I_{j;\delta,M+1}\subseteq I_j$,

$$\text{(I)}=\underbrace{\{\tau_{i;\delta,M}^{\eta|b}(x)\cdot\lambda_{i;b}^*(\eta)>t;\ X^{\eta|b}_{\tau_{i;\delta,M}^{\eta|b}(x))}(x)\in I_j\}}_{\text{(III)}}\cap\underbrace{\{X^{\eta|b}_{\tau_{i;\delta,M}^{\eta|b}(x)}(x)\in I_{j;\delta,M+1}\}}_{\text{(IV)}}.$$

Given any $\Delta>0$, by applying part (a) of Result 2 onto $I_{i;\delta,M}$, we yield (for any δ small enough)

$$\begin{aligned}\liminf_{\eta\downarrow 0}\inf_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}(\text{(III)})&\ge\exp\big(-\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_{i;\delta,M}^c;m_i)\cdot t\big)\cdot\frac{\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_j;m_i)}{\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_{i;\delta,M}^c;m_i)}\\&>\frac{\exp(-(1+\Delta)q_b(i)\cdot t)}{1+\Delta}\cdot\frac{q_b(i,j)}{q_b(i)}\qquad\text{due to (C.16) and (C.17).}\end{aligned}$$

Meanwhile, note that $(\text{IV})^c=\{X^{\eta|b}_{\tau_{i;\delta,M}^{\eta|b}(x)}(x)\in S(\delta)\}\cup\{|X^{\eta|b}_{\tau_{i;\delta,M}^{\eta|b}(x)}(x)|\ge M+1\}$. Due to $\tau_{i;\delta,M}^{\eta|b}(x)\le\sigma_{i;\epsilon}^{\eta|b}(x)$, the claim $\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((\text{IV})^c)<\Delta$ follows directly from (C.11) of Lemma C.1. In summary, for all $\delta>0$ small enough,

$$\liminf_{\eta\downarrow 0}\inf_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}(\text{(I)})\ge\frac{\exp(-(1+\Delta)q_b(i)\cdot t)}{1+\Delta}\cdot\frac{q_b(i,j)}{q_b(i)}-\Delta.\tag{C.20}$$

Combining (C.19) and (C.20) and then driving $\Delta\downarrow 0$, we conclude the proof of the lower bound (C.14).

Proof of Upper Bound (C.15).

Let $(\mathrm{I}) = \{\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_{i;b}(\eta) > t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x)\in I_j\}$. Arbitrarily pick some $\Delta>0$. Given $\delta>0$, define event $(\mathrm{II}) = \{X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in(-M-1,M+1)\backslash S(\delta)\}$. We start from the decomposition $(\mathrm{I}) = \big((\mathrm{I})\setminus(\mathrm{II})\big)\cup\big((\mathrm{I})\cap(\mathrm{II})\big)$. Applying (C.11) of Lemma C.1, it holds for all $\delta>0$ small enough that

$$\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((\mathrm{I})\setminus(\mathrm{II}))\le\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((\mathrm{II})^c)<\Delta. \tag{C.21}$$

Next, recall that $\tau^{\eta|b}_{i;\delta,M}(x)$ is the first exit time from $I_{i;\delta,M}$, and $\sigma^{\eta|b}_{i;\epsilon}(x)$ is the first time visiting the ϵ-neighborhood of a local minimum different from m_i; see (C.8) and (C.9). By definition of $\tau^{\eta|b}_{i;\delta,M}(x)$, on event $(\mathrm{I})\cap(\mathrm{II})$ there must be some $K\in[n_{\min}]$, $K\neq i$ such that $X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in(-M-1,M+1)\cap(s_{K-1}+\delta,s_K-\delta) = I_{K;\delta,M+1}$. For each $k\in[n_{\min}]$ with $k\neq i$, define event

$$(k) = (\mathrm{I})\cap(\mathrm{II})\cap\{X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in I_{k;\delta,M+1}\}$$

and note that $\bigcup_{k\in[n_{\min}]:\ k\neq i}(k) = (\mathrm{I})\cap(\mathrm{II})$. To proceed, consider the following decomposition

$$(k) = \underbrace{\Big((k)\cap\Big\{\Big(\sigma^{\eta|b}_{i;\epsilon}(x)-\tau^{\eta|b}_{i;\delta,M}(x)\Big)\cdot\lambda^*_{i;b}(\eta)>\Delta\Big\}\Big)}_{(k,1)}\cup\underbrace{\Big((k)\cap\Big\{\Big(\sigma^{\eta|b}_{i;\epsilon}(x)-\tau^{\eta|b}_{i;\delta,M}(x)\Big)\cdot\lambda^*_{i;b}(\eta)\le\Delta\Big\}\Big)}_{(k,2)}.$$

We fix some $k\in[n_{\min}]$ with $k\neq i$ and analyze the probability of events $(k,1)$ and $(k,2)$ separately. First, as has been shown at the beginning of the proof of part (ii), we have $\lim_{\eta\downarrow0}\frac{T/\eta}{\Delta/\lambda^*_{i,b}(\eta)}=0$ for all $T\in(0,\infty)$. Then,

$$\begin{aligned}
&\limsup_{\eta\downarrow0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((k,1))\\
&\le\limsup_{\eta\downarrow0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}\big((k)\cap\{\sigma^{\eta|b}_{i;\epsilon}(x)-\tau^{\eta|b}_{i;\delta,M}(x)>T/\eta\}\big)\\
&\le\limsup_{\eta\downarrow0}\sup_{y\in I_{k;\delta,M+1}}\mathbf{P}(\sigma^{\eta|b}_{i;\epsilon}(y)>T/\eta)\qquad\text{by strong Markov property at }\tau^{\eta|b}_{i;\delta,M}(x)\\
&\le\limsup_{\eta\downarrow0}\sup_{y\in I_{k;\delta,M+1}}\mathbf{P}(X^{\eta|b}_j(y)\notin(m_k-\epsilon,m_k+\epsilon)\ \forall j\le T/\eta)\\
&=0\qquad\text{for all }T>0\text{ large enough due to Lemma E.5.}
\end{aligned}\tag{C.22}$$

Meanwhile,

$$\begin{aligned}
&\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((k,2))\\
&\le\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}\big(\tau^{\eta|b}_{i;\delta,M}(x)\cdot\lambda^*_{i;b}(\eta)>t-\Delta;\ X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in I_{k;\delta,M+1};\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x)\in I_j\big)\\
&\le\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}\big(\tau^{\eta|b}_{i;\delta,M}(x)\cdot\lambda^*_{i;b}(\eta)>t-\Delta;\ X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in I_{k;\delta,M+1}\big)\\
&\quad\cdot\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}\big(X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x)\in I_j\ \big|\ \tau^{\eta|b}_{i;\delta,M}(x)\cdot\lambda^*_{i;b}(\eta)>t-\Delta;\ X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in I_{k;\delta,M+1}\big)\\
&\le\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\underbrace{\mathbf{P}\big(\tau^{\eta|b}_{i;\delta,M}(x)\cdot\lambda^*_{i;b}(\eta)>t-\Delta;\ X^{\eta|b}_{\tau^{\eta|b}_{i;\delta,M}(x)}(x)\in I_k\big)}_{(k,\mathrm{I})}\cdot\sup_{y\in I_{k;\delta,M+1}}\underbrace{\mathbf{P}\big(X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(y)}(y)\in I_j\big)}_{(k,\mathrm{II})}.
\end{aligned}$$

Applying part (a) of Result 2 onto $I_{i;\delta,M}$ and the bound (C.18), we yield (for any δ small enough)

$$\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((k,\mathrm{I}))\le\exp\big(-\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I^c_{i;\delta,M};m_i)\cdot(t-\Delta)\big)\cdot\frac{\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I_k^-;m_i)}{\check{\mathbf{C}}^{(\mathcal{J}_b^*(i))|b}(I^c_{i;\delta,M};m_i)}$$
$$\le\exp\big(-q_b(i)\cdot(t-\Delta)\big)\cdot\frac{q_b(i,k)}{q_b(i)}\quad\text{using (C.18) and (C.16).}\qquad\text{(C.23)}$$

Moving on, we analyze the probability of event (k,II). If $k=j$, we apply the trivial upper bound $\mathbf{P}((k,\mathrm{II}))\le 1$. If $k\neq j$, on event (k,II), $\big(X_n^{\eta|b}(y)\big)_{n\ge 0}$ visited $(m_j-\epsilon,m_j+\epsilon)$ before visiting any other local minima's ϵ-neighborhood, despite the fact that the initial value $X_0^{\eta|b}(y)=y$ belongs to $I_{k;\delta,M+1}\subset I_k$. This implies that $\big(X_n^{\eta|b}(y)\big)_{n\ge 0}$ must have left I_k before visiting its local minimum m_k. Applying Lemma E.5, we obtain $\limsup_{\eta\downarrow 0}\sup_{y\in I_{k;\delta,M+1}}\mathbf{P}((k,\mathrm{II}))=0\ \forall k\neq j$ for all $\delta>0$ small enough In summary, for all δ small enough,

$$\begin{aligned}
&\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}\big(\sigma^{\eta|b}_{i;\epsilon}(x)\cdot\lambda^*_{i;b}(\eta)>t,\ X^{\eta|b}_{\sigma^{\eta|b}_{i;\epsilon}(x)}(x)\in I_j\big)\\
&\le\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((\mathrm{II})^c)\\
&\qquad+\sum_{k\in[n_{\min}]:\ k\neq i}\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((k,\mathrm{I}))\cdot\limsup_{\eta\downarrow 0}\sup_{y\in I_{k;\delta,M}}\mathbf{P}((k,\mathrm{II}))\qquad\text{due to (C.22)}\\
&\le\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((\mathrm{II})^c)+\limsup_{\eta\downarrow 0}\sup_{x\in[m_i-\epsilon,m_i+\epsilon]}\mathbf{P}((j,\mathrm{I}))\\
&\le\Delta+\exp\big(-q_b(i)\cdot(t-\Delta)\big)\cdot\frac{q_b(i,j)}{q_b(i)}\qquad\text{due to (C.21) and (C.23).}
\end{aligned}$$

Let $\Delta\downarrow 0$ and we conclude the proof of the upper bound. □

Now, we are ready to prove Proposition B.1.

Proof of Proposition B.1. We first show that claims (i) and (ii) follow directly from the next claim: for any $\epsilon>0$ small enough,

$$(U_1^{\eta,\epsilon},V_1^{\eta,\epsilon},U_2^{\eta,\epsilon},V_2^{\eta,\epsilon},\cdots)\Rightarrow(U_1,V_2,U_2,V_2,\cdots)\qquad\text{as }\eta\downarrow 0,\qquad\text{(C.24)}$$

where the laws of U_j's and V_j's are defined in (C.5). Specifically, we only consider $\epsilon>0$ small enough such that claim (C.24) holds. In light of Lemma A.5 and Proposition D.1, (C.24) immediately leads to the claims in (i). Regarding claim (ii), note that $\hat{X}_t^{\eta,\epsilon,|b}$ is a step function (i.e., piece-wise constant) that only takes values in $\mathcal{M}\triangleq\{m_j:\ j=1,2,\cdots,n_{\min}\}$, which is a finite set. Let

$$A_N\triangleq\{\xi\in\mathbb{D}[0,T]:\ \xi\text{ is a step function with at most }N\text{ jumps and only takes values in }\mathcal{M}\}.$$

First, the finite-dimensional nature of A_N (i.e., at most N jumps over $[0,T]$, only $n_{\min}$ possible values) implies that A_N is a compact set in $(\mathbb{D}[0,T],\boldsymbol{d}^{[0,T]}_{L_p})$. Besides,

$$\limsup_{n\to\infty}\mathbf{P}(\hat{X}_\cdot^{\eta_n,\epsilon,|b}\notin A_N)=\limsup_{n\to\infty}\mathbf{P}(\sum_{j=1}^{N+1}U_j^{\eta_n,\epsilon}\le T)\le\mathbf{P}(\sum_{j=1}^{N+1}U_j\le T),$$

where the last inequality follows from $(U_1^{\eta_n,\epsilon},\cdots,U_N^{\eta_n,\epsilon})\Rightarrow(U_1,\cdots,U_N)$. Using part (i) of Proposition D.1, we have $\lim_{N\to\infty}\limsup_{n\to\infty}\mathbf{P}(\hat{X}_\cdot^{\eta_n,\epsilon,|b}\notin A_N)=0$, which verifies the tightness of $\hat{X}_\cdot^{\eta_n,\epsilon,|b}$.

Now, it only remains to prove (C.24). This is equivalent to proving that, for each $N\ge 1$, $(U_1^{\eta,\epsilon},V_1^{\eta,\epsilon},\cdots,U_N^{\eta,\epsilon},V_N^{\eta,\epsilon})$ converges in distribution to $(U_1,V_1,\cdots,U_N,V_N)$ as $\eta\downarrow 0$. Fix some

$N = 1, 2, \cdots$. First, note that $U_1 = 0$ and $V_1 = m_i$. From part (i) of Lemma C.2, we get $(U_1^{\eta,\epsilon}, V_1^{\eta,\epsilon}) \Rightarrow (0, m_i) = (U_1, V_1)$ as $\eta \downarrow 0$. Next, for any $n \geq 1$, any $t_l \in (0, \infty)$, any $v_l \in \{m_i : \ i \in [n_{\min}]\}$, and any $t > 0$, $i, j \in [n_{\min}]$ with $i \neq j$, it follows directly from part (ii) of Lemma C.2 that

$$\begin{aligned}
&\lim_{\eta \downarrow 0} \mathbf{P}\Big(U_{n+1}^{\eta,\epsilon} \leq t, \ V_{n+1}^{\eta,\epsilon} = m_j \ \Big| \ V_n^{\eta,\epsilon} = m_i, \ V_l^{\eta,\epsilon} = v_l \ \forall l \in [n-1], \ U_l^{\eta,\epsilon} \leq t_l \ \forall l \in [n] \Big) \\
&= \begin{cases} \frac{q_b(i,j)}{q_b(i)} & \text{if } m_i \notin V_b^*, \\ \frac{q_b(i,j)}{q_b(i)} \cdot \Big(1 - \exp\big(-q_b(i)t\big)\Big) & \text{if } m_i \in V_b^*. \end{cases}
\end{aligned}$$

This coincides with the conditional law of $\mathbf{P}\Big(U_{n+1} < t, \ V_{n+1} = m_j \ \Big| \ V_n = m_i, \ (V_j)_{j=1}^{n-1}, \ (U_j)_{j=1}^{n}\Big)$ specified in (C.5). By arguing inductively, we conclude the proof. □

Moving onto the proof of Proposition B.2, we first prepare a lemma that establishes the weak convergence from $X^{\eta|b}_{\lfloor \cdot/\lambda_b^*(\eta) \rfloor}(x)$ to $\hat{X}^{\eta,\epsilon|b}_{\cdot}(x)$ in terms of finite dimensional distributions.

Lemma C.3. *Let Assumptions 1, 2, 3, 4, 6, and 7 hold. Given any $t > 0$ and $x \in \bigcup_{i \in [n_{min}]} I_i$,*

(i) $\lim_{\eta \downarrow 0} \mathbf{P}\Big(X_j^{\eta|b}(x) \notin (-M, M) \text{ for some } j \leq t/\lambda_b^*(\eta)\Big) = 0$ *for the constant $M > 0$ specified in Lemma C.1;*

(ii) $\lim_{\eta \downarrow 0} \mathbf{P}\Big(\Big|X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta) \rfloor}(x) - \hat{X}_t^{\eta,\epsilon|b}(x)\Big| \geq \epsilon\Big) = 0$ *for all $\epsilon > 0$ small enough.*

Proof. Throughout this proof, let $\bar{\epsilon}$ be specified as in (C.2).

(i) We prove a stronger result. Let $I_{M,\delta} = (-M, M) \backslash S(\delta)$ where $S(\delta) = \bigcup_{i \in [n_{\min}-1]} [s_i - \delta, s_i + \delta]$. Recall the definition of $\hat{\tau}_j^{\eta,\epsilon|b}(x)$ in (B.1) and (B.3). For any $N \in \mathbb{Z}_+$, on event

$$\Bigg(\bigcap_{k=1}^{N-1} \underbrace{\Big\{ X_j^{\eta|b}(x) \in I_{M,\delta} \ \forall j \in \big[\hat{\tau}_k^{\eta,\epsilon|b}(x), \hat{\tau}_{k+1}^{\eta,\epsilon|b}(x)\big] \Big\}}_{A_k(\eta,\delta)} \Bigg) \cap \underbrace{\Big\{ \hat{\tau}_1^{\eta,\epsilon|b}(x) \leq t/\lambda_b^*(\eta) \Big\}}_{B_1(\eta)} \cap \underbrace{\Big\{ \hat{\tau}_N^{\eta,\epsilon|b} > t/\lambda_b^*(\eta) \Big\}}_{B_2(\eta)}$$

we have $X_j^{\eta|b}(x) \in I_{M,\delta}$ for all $j \in [\hat{\tau}_1^{\eta,\epsilon|b}(x), \hat{\tau}_N^{\eta,\epsilon|b}(x)]$ and $\hat{\tau}_1^{\eta,\epsilon|b}(x) \leq t/\lambda_b^*(\eta) < \hat{\tau}_N^{\eta,\epsilon|b}(x)$. Therefore, it suffices to show that for any $\Delta > 0$, there are some positive integer N and $\delta > 0$ such that

$$\limsup_{\eta \downarrow 0} \Big[\mathbf{P}\big(B_1^c(\eta)\big) + \mathbf{P}\big(B_2^c(\eta)\big) + \sum_{k=1}^{N-1} \mathbf{P}\big(A_k^c(\eta, \delta)\big) \Big] < \Delta. \tag{C.25}$$

Let $i \in [n_{\min}]$ be such that $x \in I_i$ and let $R_{i;\epsilon}^{\eta|b}(x) = \min\{j \geq 0 : \ X_j^{\eta|b}(x) \in [m_i - \epsilon, m_i + \epsilon]\}$. Since $\hat{\tau}_1^{\eta,\epsilon|b}(x)$ is the first visit time to $\bigcup_{l \in [n_{\min}]} (m_l - \epsilon, m_l + \epsilon)$, we have $\hat{\tau}_1^{\eta,\epsilon|b}(x) \leq R_{i;\epsilon}^{\eta|b}(x)$ and hence

$$\begin{aligned}
\limsup_{\eta \downarrow 0} \mathbf{P}\Big(B_1^c(\eta)\Big) &\leq \limsup_{\eta \downarrow 0} \mathbf{P}\Big(\hat{\tau}_1^{\eta,\epsilon|b}(x) > t/\lambda_b^*(\eta)\Big) \leq \limsup_{\eta \downarrow 0} \mathbf{P}\Big(\lambda_b^*(\eta) \cdot R_{i;\epsilon}^{\eta|b}(x) > t\Big) \\
&= 0 \qquad \text{using Lemma C.2 } (i).
\end{aligned} \tag{C.26}$$

We move onto the analysis of event $B_2(\eta)$ and the choice of N. Recall that $Y_t^{*|b}(x)$ is the irreducible, continuous-time Markov chain over V_b^* with important properties summarized in Section D. In particular, we can fix some N large enough such that $\mathbf{P}(U_1 + \cdots + U_N \leq t) < \Delta/2$. From part (i) of Proposition B.1, we now get

$$\begin{aligned}
\limsup_{\eta \downarrow 0} \mathbf{P}\Big(B_2^c(\eta)\Big) &\leq \limsup_{\eta \downarrow 0} \mathbf{P}\Big(\sum_{n=1}^{N} \big(\tau_n^{\eta,\epsilon|b}(x) - \tau_{n-1}^{\eta,\epsilon|b}(x)\big) \cdot \lambda_b^*(\eta) \leq t \Big) \\
&\leq \mathbf{P}(U_1 + \cdots + U_N \leq t) < \Delta/2.
\end{aligned} \tag{C.27}$$

Meanwhile, recall that $\sigma^{\eta|b}_{k;\epsilon}(x) = \min\{j \geq 0 : \ X^{\eta|b}_j(x) \in \bigcup_{l\neq k}(m_l - \epsilon, m_l + \epsilon)\}$ (i.e., the first time $X^{\eta|b}_j(x)$ visits the ϵ-neighborhood of some m_l that is different from m_k); also, for all $j \geq 2$, $\hat{\tau}^{\eta,\epsilon|b}_j(x)$ is the first time since $\hat{\tau}^{\eta,\epsilon|b}_{j-1}(x)$ that $X^{\eta|b}_\cdot(x)$ visits the ϵ-neighborhood of some m_l that is different from the one visited at $\hat{\tau}^{\eta,\epsilon|b}_{j-1}(x)$. From the strong Markov property at $\hat{\tau}^{\eta,\epsilon|b}_k(x)$,

$$\sup_{k\geq 1}\mathbf{P}\Big(A^c_k(\eta)\Big) \leq \max_{l\in[n_{\min}]}\sup_{y\in[m_l-\epsilon,m_l+\epsilon]}\mathbf{P}\Big(\exists j < \sigma^{\eta|b}_{l;\epsilon}(y) \ s.t. \ X^{\eta|b}_j(y) \in S(\delta) \text{ or } \big|X^{\eta|b}_j(y)\big| \geq M\Big).$$

Applying Lemma C.1, we are able to fix some $M > 0$ and $\delta \in (0, \epsilon/2)$ such that $\limsup_{\eta\downarrow 0}\mathbf{P}\big(A^c_k(\eta)\big) \leq \frac{\Delta}{2N}$ $\forall k \in [N-1]$. Combining this bound with (C.26) and (C.27), we finish the proof of (C.25).

(ii) If $X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor}(x) \in \bigcup_{l\in[n_{\min}]}(m_l-\epsilon, m_l+\epsilon)$, then due to the definition of $\hat{X}^{\eta,\epsilon|b}_t(x)$ as the marker of the last visited local minimum (see (B.1)–(B.4) for the definition of the process $\hat{X}^{\eta,\epsilon|b}_t(x)$), we must have $|X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor}(x) - \hat{X}^{\eta,\epsilon|b}_t(x)| < \epsilon$. Therefore, it suffices to show that for any $\epsilon \in (0, \bar{\epsilon})$

$$\lim_{\eta\downarrow 0}\mathbf{P}\Big(X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor}(x) \in \bigcup_{l\in[n_{\min}]}(m_l-\epsilon, m_l+\epsilon)\Big) = 1.$$

Pick some $\delta_t \in (0, \frac{t}{3})$, $\delta > 0$. Recall that $H(\cdot) = \mathbf{P}(|Z_1| > \cdot)$, and define event

$$\text{(I)} = \Big\{X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor - \lfloor 2\delta_t/H(\eta^{-1})\rfloor}(x) \in I_{M,\delta}\Big\}.$$

Let $t_1(\eta) = \lfloor t/\lambda^*_b(\eta)\rfloor - \lfloor 2\delta_t/H(\eta^{-1})\rfloor$. On event (I), let $R^\eta \triangleq \min\{j \geq t_1(\eta) : \ X^{\eta|b}_j(x) \in \bigcup_{l\in[n_{\min}]}(m_l - \frac{\epsilon}{2}, m_l + \frac{\epsilon}{2})\}$ and set $\hat{\mathcal{I}}^\eta$ by the rule $\hat{\mathcal{I}}^\eta = j \iff X^{\eta|b}_{R^\eta}(x) \in I_j$. Now, define event

$$\text{(II)} = \Big\{R^\eta - t_1(\eta) \leq \delta_t/H(\eta^{-1})\Big\}.$$

On event (I) $\cap$ (II) we have $\lfloor t/\lambda^*_b(\eta)\rfloor - \lfloor 2\delta_t/H(\eta^{-1})\rfloor \leq R^\eta \leq \lfloor t/\lambda^*_b(\eta)\rfloor$. Let $\tau^\eta \triangleq \min\{j \geq R^\eta : X^{\eta|b}_j(x) \notin (m_{\hat{\mathcal{I}}^\eta} - \epsilon, m_{\hat{\mathcal{I}}^\eta} + \epsilon)\}$, and define event

$$\text{(III)} = \Big\{\tau^\eta - R^\eta > 2\delta_t/H(\eta^{-1})\Big\}.$$

On event (I) $\cap$ (II) $\cap$ (III), we have $\tau^\eta > \lfloor t/\lambda^*_b(\eta)\rfloor \geq R^\eta$, and hence $X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor}(x) \in \bigcup_{l\in[n_{\min}]}(m_l - \epsilon, m_l + \epsilon)$. Furthermore, we claim that for any $\Delta > 0$ there exist $\delta_t \in (0, \frac{t}{3})$ and $\delta > 0$ such that

$$\liminf_{\eta\downarrow 0}\mathbf{P}\Big(\text{(I)}\Big) \geq 1 - \Delta, \tag{C.28}$$

$$\liminf_{\eta\downarrow 0}\mathbf{P}\Big(\text{(II)} \ \Big| \ \text{(I)}\Big) \geq 1, \tag{C.29}$$

$$\liminf_{\eta\downarrow 0}\mathbf{P}\Big(\text{(III)} \ \Big| \ \text{(I)} \cap \text{(II)}\Big) \geq 1 - \Delta. \tag{C.30}$$

An immediate consequence is that $\liminf_{\eta\downarrow 0}\mathbf{P}(\text{(I)}\cap\text{(II)}\cap\text{(III)}) \geq (1-\Delta)^2$. Let $\Delta \downarrow 0$ and we conclude the proof. Now it only remains to establish (C.28) (C.29) (C.30). Throughout the remainer of this proof, we fix some $\epsilon \in (0, \bar{\epsilon})$ and $\Delta > 0$.

Proof of (C.28). This has been established in the proof for part *(i)*.

Proof of (C.29). We show that the claim holds for all $\delta_t \in (0, t/3)$. Due to $H(x) \in \mathcal{RV}_{-\alpha}(x)$ and $\alpha > 1$, given any $T > 0$ we have $T/\eta < \delta_t/H(\eta^{-1})$ eventually for all η small enough. Recall that

$I_{j;\delta,M} = (s_{j-1}+\delta, s_j-\delta)\cap(-M,M)$. By Markov property at $t_1(\eta)$, for any $T>0$ it holds for all $\eta>0$ small enough that

$$\begin{aligned}\mathbf{P}\Big((\mathrm{II})^c \;\Big|\; (\mathrm{I})\Big) &\le \max_{k\in[n_{\min}]}\sup_{y\in I_{k;\delta,M}}\mathbf{P}\Big(X_j^{\eta|b}(y)\notin\bigcup_{l\in[n_{\min}]}(m_l-\frac{\epsilon}{2},m_l+\frac{\epsilon}{2})\ \forall j\le \delta_t/H(\eta^{-1})\Big)\\ &\le \max_{k\in[n_{\min}]}\sup_{y\in I_{k;\delta,M}}\mathbf{P}\Big(R^{\eta|b}_{k;\epsilon/2}(y)>\delta_t/H(\eta^{-1})\Big)\\ &\le \max_{k\in[n_{\min}]}\sup_{y\in I_{k;\delta,M}}\mathbf{P}\Big(R^{\eta|b}_{k;\epsilon/2}(y)>T/\eta\Big)\end{aligned}$$

where $R^{\eta|b}_{k;\epsilon/2}(y)=\min\{j\ge 0:\ X_j^{\eta|b}(y)\in(m_k-\frac{\epsilon}{2},m_k+\frac{\epsilon}{2})\}$.

Let $\boldsymbol{t}_k(x,\epsilon)\triangleq\inf\{t\ge 0:\ \boldsymbol{y}_t(x)\in(m_k-\epsilon,m_k+\epsilon)\}$. By Assumption 6, $\boldsymbol{t}_k(x,\frac{\epsilon}{4})<\infty$ for all $x\in[-M-1,M+1]\cap[s_{k-1}+\frac{\delta}{2},s_k-\frac{\delta}{2}]$, with $\boldsymbol{t}_k(\,\cdot\,,\frac{\epsilon}{4})$ being continuous over $[-M-1,M+1]\cap[s_{k-1}+\frac{\delta}{2},s_k-\frac{\delta}{2}]$. As a result, we can fix $T\in(0,\infty)$ large enough such that

$$T>\sup\Big\{\boldsymbol{t}_k(x,\frac{\epsilon}{4}):\ x\in[-M-1,M+1]\cap[s_{k-1}+\frac{\delta}{2},s_k-\frac{\delta}{2}]\Big\}\qquad\forall k\in[n_{\min}].$$

For each $k\in[n_{\min}]$, by applying Lemma E.5 onto $(-M-1,M+1)\cap(s_{k-1},s_k)$, we are able to show that $\limsup_{\eta\downarrow 0}\sup_{y\in I_{k;\delta,M}}\mathbf{P}\Big(R^{\eta|b}_{k;\epsilon/2}(y)>T/\eta\Big)=0$. This concludes the proof of claim (C.29).

Proof of (C.30). We prove the claim for all δ_t small enough. By strong Markov property at R^η,

$$\mathbf{P}\Big((\mathrm{III})^c\;\Big|\;(\mathrm{I})\cap(\mathrm{II})\Big)\le\max_{k\in[n_{\min}]}\sup_{y\in[m_k-\epsilon/2,m_k+\epsilon/2]}\mathbf{P}\Big(\exists j\le\frac{2\delta_t}{H(\eta^{-1})}\ s.t.\ X_j^{\eta|b}(y)\notin(m_k-\epsilon,m_k+\epsilon)\Big).$$

Also, note that $\epsilon<\bar\epsilon<b$; see (C.2). For each $k\in[n_{\min}]$, by applying part (a) of Result 2 onto $(m_k-\epsilon,m_k+\epsilon)$, we obtain some $c_{k,\epsilon}\in(0,\infty)$ such that for any $u>0$,

$$\limsup_{\eta\downarrow0}\sup_{y\in[m_k-\epsilon/2,m_k+\epsilon/2]}\mathbf{P}\Big(\exists j\le\frac{u}{H(\eta^{-1})}\ s.t.\ X_j^{\eta|b}(y)\notin(m_k-\epsilon,m_k+\epsilon)\Big)\le 1-\exp(-c_{k,\epsilon}\cdot u).$$

By picking δ_t small enough, we ensure that $1-\exp(-c_{k,\epsilon}\cdot 2\delta_t)<\Delta$ for all $k\in[n_{\min}]$, thus completing the proof of claim (C.30). □

The next result provides a bound over the proportion of time that $X_j^{\eta|b}(x)$ is not close enough to a local minimum.

Lemma C.4. *Let Assumptions 1, 2, 3, 4, 6, and 7 hold. Given any $\epsilon\in(0,\bar\epsilon)$, it holds for all $t\in(0,1)$ small enough that*

$$\limsup_{\eta\downarrow0}\max_{i:\ m_i\in V_b^*}\sup_{x\in(m_i-\frac{\epsilon}{2},m_i+\frac{\epsilon}{2})}\mathbf{P}\Big(\int_0^t\mathbf{I}\Big\{X^{\eta|b}_{\lfloor s/\lambda_b^*(\eta)\rfloor}(x)\notin(m_i-\epsilon,m_i+\epsilon)\Big\}ds>t^2\Big)<q^*t,$$

where $q^*\in(0,\infty)$ is a constant that does not vary with t.

Proof. There are only finitely many elements in V_b^*. Therefore, it suffices to fix some $m_i\in V_b^*$ (recall that $I_i=(s_{i-1},s_i)$ is the attraction field associated with m_i, and w.l.o.g. we assume $m_i=0$) and prove that

$$\limsup_{\eta\downarrow0}\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\Big(\int_0^t\mathbf{I}\Big\{X^{\eta|b}_{\lfloor s/\lambda_b^*(\eta)\rfloor}(x)\notin(-\epsilon,\epsilon)\Big\}ds>t^2\Big)<q^*t\tag{C.31}$$

holds for all $t>0$ small enough, where $q^*\in(0,\infty)$ is a constant that does not vary with ϵ,Δ, or t.

Let $T_0^\eta=0$, and (for all $i\geq 1$)

$$S_i^\eta\triangleq\min\{j>T_{i-1}^\eta:\ X_j^{\eta|b}(x)\notin(-\epsilon,\epsilon)\},\qquad T_i^\eta\triangleq\min\{j>S_i^\eta:\ X_j^{\eta|b}(x)\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})\}.$$

Note that if $X_j^{\eta|b}(x)\notin(-\epsilon,\epsilon)$, then there is some $i\geq 1$ such that $j\in[S_i^\eta,T_i^\eta-1]$. Next, let $N^\eta\triangleq\max\{i\geq 0:\ S_i^\eta\leq t/\lambda_b^*(\eta)\}$, and note that $\#\{j\leq\lfloor t/\lambda_b^*(\eta)\rfloor:\ X_j^{\eta|b}(x)\notin(-\epsilon,\epsilon)\}\leq\sum_{i=1}^{N^\eta}T_i^\eta-S_i^\eta$.

Now, recall that $\alpha>1$ is the heavy-tailed index in Assumption 1, and the scale function $\lambda_b^*(\eta)$ is defined in (3.8) with $\lambda_b^*(\eta)\in\mathcal{RV}_{\mathcal{J}_b^*(V)\cdot(\alpha-1)+1}(\eta)$. Let $\beta\in(0,\alpha-1)$ and $k(\eta)=1/\eta^{(J_b^*(V)-1)(\alpha-1)+\beta}$. For any $x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})$, define events

$$A_{t,M}^\eta(x)\triangleq\big\{X_j^{\eta|b}(x)\in(s_{i-1}+\frac{\epsilon}{2},s_i-\frac{\epsilon}{2})\cap(-M,M)\text{ for all }j\leq\lfloor t/\lambda_b^*(\eta)\rfloor\big\},$$
$$B_{t,\delta}^\eta(x)\triangleq\big\{\text{for each }i\leq k(\eta),\exists j\in[T_{i-1}^\eta+1,S_i^\eta]\text{ s.t. }\eta|Z_j|>\delta\big\}.$$

On $B_{t,\delta}^\eta(x)$, we must have $N^\eta\leq\#\{j\leq\lfloor t/\lambda_b^*(\eta)\rfloor:\ \eta|Z_j|>\delta\}$. Furthermore, given some constant $T\in(0,\infty)$, let $E_{t,T}^\eta(x)\triangleq\{T_i^\eta\wedge\lfloor t/\lambda_b^*(\eta)\rfloor-S_i^\eta\leq T/\eta\ \forall i\geq 1\}$. On event $B_{t,\delta}^\eta(x)\cap E_{t,T}^\eta(x)$ we get

$$\begin{aligned}\#\{j\leq\lfloor t/\lambda_b^*(\eta)\rfloor:\ X_j^{\eta|b}(x)\notin(-\epsilon,\epsilon)\}&\leq\sum_{i=1}^{N^\eta}T_i^\eta\wedge\lfloor t/\lambda_b^*(\eta)\rfloor-S_i^\eta\\&\leq k(\eta)\cdot T/\eta=T/\eta^{1+\beta+(J_b^*(V)-1)(\alpha-1)},\end{aligned}$$

and hence

$$\int_0^t\mathbf{I}\Big\{X^{\eta|b}_{\lfloor s/\lambda_b^*(\eta)\rfloor}(x)\notin(-\epsilon,\epsilon)\Big\}ds\leq\frac{T/\eta^{1+\beta+(J_b^*(V)-1)(\alpha-1)}+1}{\lfloor t/\lambda_b^*(\eta)\rfloor}.$$

However, due to $\lambda_b^*(\eta)\in\mathcal{RV}_{\mathcal{J}_b^*(V)\cdot(\alpha-1)+1}(\eta)$ and $J_b^*(V)\cdot(\alpha-1)+1=(J_b^*(V)-1)\cdot(\alpha-1)+\alpha>(J_b^*(V)-1)\cdot(\alpha-1)+1+\beta$, we have (for any $t,T>0$ and $\beta\in(0,\alpha-1)$)

$$\lim_{\eta\downarrow 0}\frac{T/\eta^{1+\beta+(J_b^*(V)-1)(\alpha-1)}+1}{\lfloor t/\lambda_b^*(\eta)\rfloor}=0.$$

The discussion above implies the following: to prove (C.31), it suffices to find some $t,T,M,\delta\in(0,\infty)$ such that

$$\limsup_{\eta\downarrow 0}\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\Big((A_{t,M}^\eta(x))^c\Big)<q^*t,\tag{C.32}$$

$$\lim_{\eta\downarrow 0}\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\Big(\big(B_{t,\delta}^\eta(x)\big)^c\Big)=0,\tag{C.33}$$

$$\lim_{\eta\downarrow 0}\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\Big(A_{t,M}^\eta(x)\cap B_{t,\delta}^\eta(x)\cap\big(E_{t,T}^\eta(x)\big)^c\Big)=0.\tag{C.34}$$

In particular, $q^*\in(0,\infty)$ is a constant that does not vary with ϵ,Δ,M,δ, or t.

Proof of (C.32). This follows immediately from the first exit time analysis. Specifically, recall that we have assumed w.l.o.g. that the local minimum $m_i\in V_b^*$ at hand is located at the origin, i.e., $m_i=0$. This implies $\mathcal{J}_b^*(V)=\lceil\min\{|s_{i-1}|,s_i\}/b\rceil$; that is, starting from the local minimum, it requires at least $\mathcal{J}_b^*(V)$ jumps (each bounded by b) to escape from the attraction field (s_{i-1},s_i). Furthermore, by our choice of $\bar\epsilon$ in (C.2) (which is essentially due to the assumption that $|s_j-m_i|/b\notin\mathbb{Z}$ for all $i\in[n_{\min}]$ and $j\in[n_{\min}-1]$), it holds for all $\epsilon\in(0,\bar\epsilon)$ that $\mathcal{J}_b^*(V)=\lceil\min\{|s_{i-1}+\frac{\epsilon}{2}|,s_i-\frac{\epsilon}{2}\}/b\rceil$. For any $M\in(0,\infty)$ large enough, we then have $\mathcal{J}_b^*(V)=\lceil\min\{|s_{i-1}+\frac{\epsilon}{2}|,s_i-\frac{\epsilon}{2},M\}/b\rceil$, thus implying that, starting from

the origin, it also requires at least $\mathcal{J}_b^*(V)$ jumps to escape from $(s_{i-1}+\frac{\epsilon}{2}, s_i-\frac{\epsilon}{2})\cap(-M,M)$. By applying part (a) of Result 2 onto $(s_{i-1}+\frac{\epsilon}{2}, s_i-\frac{\epsilon}{2})\cap(-M,M)$, we can find $q\in(0,\infty)$ such that

$$\limsup_{\eta\downarrow 0}\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\Big((A_{t,M}^{\eta}(x))^c\Big)\le 1-\exp(-qt)\qquad\forall t>0.$$

For all $t>0$ small enough, we have $1-\exp(-qt)\le 2qt$. By picking $q^*=2q$, we conclude the proof.

Proof of (C.33). By strong Markov property at T_i^η,

$$\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\Big(\big(B_{t,\delta}^{\eta}(x)\big)^c\Big)\le k(\eta)\cdot\sup_{y\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\big(X_j^{\eta|b}(y)\notin(-\epsilon,\epsilon)\text{ for some }j<\tau_1^{>\delta}(\eta)\big)$$

Here, $\tau_1^{>\delta}(\eta)\triangleq\min\{n\ge 1:\ \eta|Z_n|>\delta\}$. That is, $\tau_1^{>\delta}(\eta)$ is the arrival time of the first Z_j with $\eta|Z_j|>\delta$. Applying Lemma E.6, it holds for all $\delta>0$ small enough that $\sup_{y\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\big(X_j^{\eta|b}(y)\notin(-\epsilon,\epsilon)\text{ for some }j<\tau_1^{>\delta}(\eta)\big)=\boldsymbol{o}\big(1/k(\eta)\big)$. This concludes the proof of claim (C.33).

Proof of (C.34). On $A_{t,M}^{\eta}(x)\cap B_{t,\delta}^{\eta}(x)$, we have $T_i^\eta\wedge\lfloor t/\lambda_b^*(\eta)\rfloor=\tilde{T}_i^\eta\wedge\lfloor t/\lambda_b^*(\eta)\rfloor$ for each $i\ge 1$, where

$$\tilde{T}_i^\eta\triangleq\min\big\{j>S_i^\eta:\ X_j^{\eta|b}(x)\notin\big((s_{i-1}+\frac{\epsilon}{2},s_i-\frac{\epsilon}{2})\cap(-M,M)\big)\setminus[-\frac{\epsilon}{2},\frac{\epsilon}{2}]\big\}.$$

Furthermore, it has been shown above that, on $B_{t,\delta}^{\eta}(x)$ we have $N^\eta\le k(\eta)$. Therefore,

$$\begin{aligned}
&\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\big(A_{t,M}^{\eta}(x)\cap B_{t,\delta}^{\eta}(x)\cap\big(E_{t,T}^{\eta}(x)\big)^c\big)\\
&\le\sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})}\mathbf{P}\big(\tilde{T}_i^\eta-S_i^\eta>T/\eta\text{ for some }i\le k(\eta)\big)\\
&\le k(\eta)\cdot\underbrace{\sup_{y\in(-\epsilon,\epsilon)}\mathbf{P}\big(X_j^{\eta|b}(x)\in\big((s_{i-1}+\frac{\epsilon}{2},s_i-\frac{\epsilon}{2})\cap(-M,M)\big)\setminus[-\frac{\epsilon}{2},\frac{\epsilon}{2}]\text{ for all }j\le\lfloor T/\eta\rfloor\big)}_{\triangleq p_T^*(\eta)}.
\end{aligned}$$

The last step is due to the strong Markov property at $\{S_i^\eta:i\in[k(\eta)]\}$. Applying Lemma E.4, we can find T large enough such that $p_T^*(\eta)=\boldsymbol{o}\big(1/k(\eta)\big)$ as $\eta\downarrow 0$ and complete the proof. □

Now, we are ready to prove Proposition B.2.

Proof of Proposition B.2. The claim $\lim_{\eta\downarrow 0}\mathbf{P}(|X_T^{\eta|b}(x)-\hat{X}_T^{\eta,\epsilon|b}(x)|\ge\epsilon)=0$ has already been proved in part (ii) of Lemma C.3. In the remainder of this proof, we focus on establishing the claim $\lim_{\eta\downarrow 0}\mathbf{P}\big(\boldsymbol{d}_{L_p}^{[0,T]}(X_{\lfloor\cdot/\lambda_b^*(\eta)\rfloor}^{\eta|b}(x),\hat{X}_\cdot^{\eta,\epsilon|b}(x))\ge 2\epsilon\big)=0$. For simplicity of notations, we focus on the case where $T=1$. Nevertheless, the proof below can be easily generalized for arbitrary $T>0$.

By definition of $\hat{X}_t^{\eta,\epsilon|b}(x)$, we have $|X_{\lfloor t/\lambda_b^*(\eta)\rfloor}^{\eta|b}(x)-\hat{X}_t^{\eta,\epsilon|b}(x)|<\epsilon$ whenever $X_{\lfloor t/\lambda_b^*(\eta)\rfloor}^{\eta|b}(x)\in\bigcup_{i\in[n_{\min}]}(m_i-\epsilon,m_i+\epsilon)$. Now, we make a few observations. For any $\eta>0$ and any positive integer N, let $\mathcal{I}_N^{(\eta)}(n)\triangleq\mathbf{I}\big\{\boldsymbol{i}_N^{(\eta)}(n)>1/N^2\big\}$ where

$$\boldsymbol{i}_N^{(\eta)}(n)\triangleq\int_{n/N}^{(n+1)/N}\mathbf{I}\Big\{X_{\lfloor t/\lambda_b^*(\eta)\rfloor}^{\eta|b}(x)\notin\bigcup_{i\in[n_{\min}]}(m_i-\epsilon,m_i+\epsilon)\Big\}dt\qquad\forall n=0,1,\cdots,N-1.$$

That is, $\boldsymbol{i}_N^{(\eta)}(n)$ is the amount of time over $[\frac{n}{N},\frac{n+1}{N})$ that $X_{\lfloor t/\lambda_b^*(\eta)\rfloor}^{\eta|b}(x)$ is not close enough to any local minima, and $\mathcal{I}_N^{(\eta)}(n)$ is the indicator that $\boldsymbol{i}_N^{(\eta)}(n)>1/N^2$.

Let $K_N^{(\eta)}\triangleq\sum_{n=1}^{N-1}\mathcal{I}_N^{(\eta)}(n)$. The proof hinges on the following claims: there exist some $C\in(0,\infty)$, a family of events A_N^η, and some constant $M>0$ such that

(*i*) on A_N^η, we have $X_j^{\eta|b}(x) \in [-M, M]$ for all $j \le \lfloor 1/\lambda_b^*(\eta) \rfloor$;

(*ii*) for all positive integer N large enough, $\lim_{\eta \downarrow 0} \mathbf{P}(A_N^\eta) = 1$;

(*iii*) for all positive integer N large enough, there exists $\bar{\eta} = \bar{\eta}(N) > 0$ such that under any $\eta \in (0, \bar{\eta})$,

$$\mathbf{P}(K_N^{(\eta)} \ge j \mid A_N^\eta) \le \mathbf{P}\Big(\mathrm{Binom}(N, \frac{2C}{N}) \ge j\Big) \qquad \forall j = 1, 2, \cdots, N.$$

Here, $\mathrm{Binom}(n, p)$ is the RV denoting the number of successful trials among n independent Bernoulli trials, each with success rate p. W.l.o.g., in claim (*i*) we can assume that M is sufficiently large such that $x \in [-M, M]$ and $m_j \in [-M, M]$ for all $j \in [n_{\min}]$. As a result, we must have $\hat{X}_t^{\eta,\epsilon|b}(x) \in [-M, M]$ as well for all $t \ge 0$. To see how we apply these claims, let

$$\boldsymbol{d}_p^{(\eta)}(n) \triangleq \int_{n/N}^{(n+1)/N} \Big| X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta) \rfloor}(x) - \hat{X}_t^{\eta,\epsilon|b}(x) \Big|^p dt.$$

On event A_N^η, for any $n = 0, 1, \cdots, N-1$, if $\boldsymbol{i}_N^{(\eta)}(n) \le 1/N^2$, we have $\boldsymbol{d}_p^{(\eta)}(n) \le \epsilon^p \cdot \frac{1}{N} + (2M)^p \cdot \frac{1}{N^2}$; Otherwise, we have the trivial bound $\boldsymbol{d}_p^{(\eta)}(n) \le (2M)^p \cdot \frac{1}{N}$. Therefore, on A_N^η,

$$\begin{aligned}
\boldsymbol{\Delta}(\eta) &\triangleq \Big(\boldsymbol{d}_{L_p}\Big(X^{\eta|b}_{\lfloor \cdot/\lambda_b^*(\eta) \rfloor}(x), \hat{X}_\cdot^{\eta,\epsilon|b}(x) \Big) \Big)^p \\
&= \sum_{n=0}^{N-1} \int_{n/N}^{(n+1)/N} \Big| X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta) \rfloor}(x) - \hat{X}_t^{\eta,\epsilon|b}(x) \Big|^p dt \\
&\le (2M)^p \cdot \frac{1}{N} + \sum_{n=1}^{N-1} \int_{n/N}^{(n+1)/N} \Big| X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta) \rfloor}(x) - \hat{X}_t^{\eta,\epsilon|b}(x) \Big|^p dt \\
&\le (2M)^p \cdot \frac{1}{N} + K_N^{(\eta)} \cdot \frac{(2M)^p}{N} + (N - 1 - K_N^{(\eta)}) \cdot \Big(\frac{\epsilon^p}{N} + \frac{(2M)^p}{N^2} \Big) \le (2M)^p \cdot \frac{1 + K_N^{(\eta)} + \frac{1}{N}}{N} + \epsilon^p.
\end{aligned}$$

Then, given any N large enough, $\eta \in (0, \bar{\eta}(N))$ and any $\beta \in (0, 1)$,

$$\begin{aligned}
\boldsymbol{r}(\eta) &\triangleq \mathbf{P}\Big(\boldsymbol{\Delta}(\eta) \ge \underbrace{\frac{1 + \frac{1}{N} + 2C + \sqrt{N^\beta}}{N}}_{\triangleq \delta(N,\beta)} \cdot (2M)^p + \epsilon^p \Big) \\
&\le \mathbf{P}(K_N^{(\eta)} \ge 2C + \sqrt{N^\beta}) = \mathbf{P}\big(\{K_N^{(\eta)} \ge 2C + \sqrt{N^\beta}\} \cap A_N^\eta \big) + \mathbf{P}\big(\{K_N^{(\eta)} \ge 2C + \sqrt{N^\beta}\} \setminus A_N^\eta \big) \\
&\le \mathbf{P}\Big(\mathrm{Binom}(N, \frac{2C}{N}) \ge 2C + \sqrt{N^\beta} \Big) + \mathbf{P}\big((A_N^\eta)^c \big) \qquad \text{by claim } (iii) \\
&\le \frac{\mathrm{var}\Big[\mathrm{Binom}(N, \frac{2C}{N}) \Big]}{N^\beta} + \mathbf{P}\big((A_N^\eta)^c \big) \le \frac{2C}{N^\beta} + \mathbf{P}\big((A_N^\eta)^c \big).
\end{aligned}$$

Driving $\eta \downarrow 0$, it follows from claim (*ii*) that $\limsup_{\eta \downarrow 0} \boldsymbol{r}(\eta) \le 2C/N^\beta$ for all N large enough. Lastly, note that $C/N^\beta \to 0$ as $N \to \infty$; also, due to $\beta \in (0, 1)$ we have $\lim_{N \to \infty} \delta(N, \beta) = 0$, and hence $\delta(N, \beta) \cdot (2M)^p + \epsilon^p < 2^p \epsilon^p$ eventually for all N large enough. In summary, we get $\lim_{\eta \downarrow 0} \mathbf{P}(\boldsymbol{\Delta}(\eta) > 2^p \epsilon^p) = 0$ and conclude the proof. Now, it only remains to verify claims (*i*), (*ii*), and (*iii*).

Proof of Claims (*i*) **and** (*ii*). We start by specifying events A_N^η. Let $t_N(n) = n/N$ and

$A_N^\eta(n)$

$$\triangleq \underbrace{\Big\{X^{\eta|b}_{\lfloor t_N(k)/\lambda^*_b(\eta)\rfloor}(x) \in \bigcup_{i:\ m_i\in V_b^*}\Big(m_i-\frac{\epsilon}{2},m_i+\frac{\epsilon}{2}\Big)\ \forall k\in[n]\Big\}}_{\triangleq A^\eta_{N,1}(n)} \cap \underbrace{\Big\{X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor}(x)\in[-M,M]\ \forall t\le t_N(k)\Big\}}_{\triangleq A^\eta_{N,2}(n)}$$

and let $A^\eta_N = A^\eta_N(N)$. Note that $A^\eta_N(1)\supseteq A^\eta_N(2)\supseteq\cdots\supseteq A^\eta_N(N)=A^\eta_N$. Furthermore, $\{X^{\eta|b}_{\lfloor t/\lambda^*_b(\eta)\rfloor}: t>0\} \overset{f.d.d.}{\to} \{Y^{*|b}_t:\ t>0\}$ due to $\{\hat{X}^{\eta,\epsilon|b}_t:\ t>0\}\overset{f.d.d.}{\to}\{Y^{*|b}_t:\ t>0\}$ and $\lim_{\eta\downarrow 0}\mathbf{P}(|X^{\eta|b}_T(x)-\hat{X}^{\eta,\epsilon|b}_T(x)|\ge\epsilon)=0$ for any $T>0$; this is the content of Lemma A.3. Next, by definition, $Y^{*|b}_t$ only visits states in V_b^*. Combining this fact with the weak convergence in f.d.d. we get $\lim_{\eta\downarrow0}\mathbf{P}\big(A^\eta_{N,1}\big)=1$ for any $N\ge1$. On the other hand, part (i) of Lemma C.3 gives $\lim_{\eta\downarrow0}\mathbf{P}\big(A^\eta_{N,2}\big)=1\ \forall N\ge1$ for any M large enough. This verifies claims (i) and (ii).

Proof of Claim (iii). Consider a random vector $\big(\widetilde{\mathcal{I}}^\eta_N(n)\big)_{n\in[N-1]}$ with law $\mathcal{L}\Big(\big(\mathcal{I}^\eta_N(n)\big)_{n\in[N-1]}\ \Big|\ A^\eta_N\Big)$. It suffices to find some $C\in(0,\infty)$ such that for all N large enough, there is $\bar\eta=\bar\eta(N)>0$ for the following claim to hold: Given any $n\in[N-1]$ and any sequence $i_j\in\{0,1\}\ \forall j\in[n-1]$,

$$\mathbf{P}\Big(\widetilde{\mathcal{I}}^\eta_N(n)=1\ \Big|\ \widetilde{\mathcal{I}}^\eta_N(j)=i_j\ \forall j\in[n-1]\Big)<2C/N\qquad\forall\eta\in(0,\bar\eta). \tag{C.35}$$

To see why, under condition (C.35) and for any $\eta\in(0,\bar\eta(N))$, there exists a coupling between iid Bernoulli RVs $(\mathcal{Z}_N(n))_{n\in[N-1]}$ with success rate $2C/N$ and $(\widetilde{\mathcal{I}}^\eta_N(n))_{n\in[N-1]}$ such that $\widetilde{\mathcal{I}}^\eta_N(n)\le\mathcal{Z}_N(n)\ \forall n\in[N-1]$ almost surely. This stochastic dominance between $(\mathcal{Z}_N(n))_{n\in[N-1]}$ and $(\widetilde{\mathcal{I}}^\eta_N(n))_{n\in[N-1]}$ immediately verifies claim (iii).

To prove condition (C.35) note that given any N, any $n\in[N-1]$, and any sequence $i_j\in\{0,1\}\ \forall j\in[n-1]$,

$$\begin{aligned}
&\mathbf{P}\Big(\widetilde{\mathcal{I}}^\eta_N(n)=1\ \Big|\ \widetilde{\mathcal{I}}^\eta_N(j)=i_j\ \forall j\in[n-1]\Big)\\
&=\frac{\mathbf{P}\Big(\widetilde{\mathcal{I}}^\eta_N(n)=1;\ \widetilde{\mathcal{I}}^\eta_N(j)=i_j\ \forall j\in[n-1]\Big)}{\mathbf{P}\Big(\widetilde{\mathcal{I}}^\eta_N(j)=i_j\ \forall j\in[n-1]\Big)}\\
&=\frac{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(n)=1;\ \mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N\Big)\Big/\mathbf{P}(A^\eta_N)}{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N\Big)\Big/\mathbf{P}(A^\eta_N)}\qquad\text{by definition of }\big(\widetilde{\mathcal{I}}^\eta_N(n)\big)_{n\in[N-1]}\\
&\le\frac{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(n)=1;\ \mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N(n)\Big)}{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N\Big)}\qquad\text{due to } A^\eta_N(n)\supseteq A^\eta_N\\
&=\frac{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(n)=1;\ \mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N(n)\Big)}{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N(n)\Big)}\cdot\frac{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N(n)\Big)}{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N\Big)}\\
&=\underbrace{\mathbf{P}\Big(\mathcal{I}^\eta_N(n)=1\ \Big|\ \{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N(n)\Big)}_{\triangleq p^\eta_1(N)}\cdot\underbrace{\frac{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N(n)\Big)}{\mathbf{P}\Big(\{\mathcal{I}^\eta_N(j)=i_j\ \forall j\in[n-1]\}\cap A^\eta_N\Big)}}_{\triangleq p^\eta_2(N)}.
\end{aligned}$$

For term $p^\eta_1(N)$, note that on $A^\eta_N(n)$ we have $X^{\eta|b}_j(x)\in\bigcup_{i:\ m_i\in V_b^*}(m_i-\frac{\epsilon}{2},m_i+\frac{\epsilon}{2})$ at $j=\lfloor t_N(n)/\lambda^*_b(\eta)\rfloor$, and hence (using Markov property)

$$p^\eta_1(N)\le\max_{i:\ m_i\in V_b^*}\ \sup_{y\in(m_i-\frac{\epsilon}{2},m_i+\frac{\epsilon}{2})}\mathbf{P}\Big(\int_0^{1/N}\mathbf{I}\Big\{X^{\eta|b}_{\lfloor s/\lambda^*_b(\eta)\rfloor}(y)\notin(m_i-\epsilon,m_i+\epsilon)\Big\}ds>1/N^2\Big).$$

Applying Lemma C.4, for all N large enough there exist $\bar{\eta} = \bar{\eta}(N) > 0$, such that $p_1^\eta \le C/N \; \forall \eta \in (0, \bar{\eta})$, where $C \in (0, \infty)$ is independent of the value of N and η. As for term p_2^η, note that for any event B with $\mathbf{P}(B) > 0$, we have

$$\frac{\mathbf{P}(B \cap A_N^\eta(n))}{\mathbf{P}(B \cap A_N^\eta)} \le \frac{\mathbf{P}(B)}{\mathbf{P}(B) - \mathbf{P}\big((A_N^\eta)^c\big)} \to 1 \qquad \text{as } \eta \downarrow 1 \text{ due to } \lim_{\eta \downarrow 0} \mathbf{P}(A_N^\eta) = 1. \tag{C.36}$$

In the definition of p_2^η, note that there are only finitely many choices of $n \in [N-1]$ and finitely many combinations for $i_j \in \{0, 1\} \; \forall j \in [n-1]$. By considering each of the finitely many choices for $B = \{\mathcal{I}_N^\eta(j) = i_j \; \forall j \in [n-1]\}$ in (C.36), we can find some $\bar{\eta} = \bar{\eta}(N)$ such that $p_2^\eta < 2 \; \forall \eta \in (0, \bar{\eta})$ uniformly for all those choices. Combining the bounds $p_1^\eta < C/N$ and $p_2^\eta < 2$, we verify condition (C.35) and conclude the proof. □

D Properties of the Markov Jump Process $Y^{*|b}$

Proposition D.1. *Let Assumptions 6 and 7 hold. Given any $m_{init} \in \{m_1, \dots, m_{n_{min}}\}$, the following claims hold for $((U_j)_{j\ge1}, (V_j)_{j\ge1})$ defined in (C.5):*

(i) For any $t > 0$, $\lim_{i\to\infty} \mathbf{P}(\sum_{j\le i} U_j > t) = 1$;

(ii) For any $u > 0$ and $i \ge 1$, $\mathbf{P}(U_1 + \cdots + U_i = u) = 0$;

(iii) $Y_\cdot^{*|b} \overset{d}{=} \Phi((U_j)_{j\ge1}, (V_j)_{j\ge1})$ holds for the mapping Φ defined in (A.4), ; that is, it is a continuous-time Markov chain with initial distribution (3.9) and generator

$$\mathbf{P}(Y_{t+h}^{*|b} = m_j \mid Y_t^{*|b} = m_i) = h \cdot \sum_{j' \in [n_{min}]:\; j' \ne i} q_b(i, j')\theta_b(m_j|m_{j'}) + \boldsymbol{o}(h) \qquad \text{as } h \downarrow 0; \tag{D.1}$$

see (3.10) and (3.11) for the definitions of $q_b(i,j)$ and $\theta_b(m_j|m_i)$, respectively.

Proof. (i) Recall the definitions of $q_b(i)$ and $q_b(i,j)$ in (3.10). Let $(S_n)_{n\ge0}$ be a discrete-time Markov chain with state space $\{m_1, \dots, m_{n_{\min}}\}$ and one-step transition kernel $\mathbf{P}(S_{n+1} = m_j | S_n = m_i) = q_b(i,j)/q_b(i)$. Note that the chain is well-defined due to (C.3). We also introduce notations $S_n(v)$ for the chain initialized under $S_0(v) = v$. For each $n \ge 0$, set $I_n^S(v) = i$ if and only if $S_n(v) = m_i$; that is, the sequence of indices $(I_n^S(v))_{n\ge0}$ indicates the state of the chain at time n.

Let $(E_i)_{i\ge0}$ be a sequence of iid Exponential RVs with rate 1, which is also independent of $(S_n(m_{\text{init}}))_{n\ge0}$. For any $i \ge 2$, the law of $(U_j)_{j\ge1}, (V_j)_{j\ge1}$ defined in (C.5) then indicates that (recall that $U_1 = 0$ and $V_1 = m_{\text{init}}$)

$$\begin{aligned}
\sum_{j\in[i]} U_j &\overset{d}{=} \sum_{j=0,1,\cdots,i-2} \frac{E_j}{q_b\big(I_j^S(m_{\text{init}})\big)} \cdot \mathbf{I}\big\{S_j(m_{\text{init}}) \in V_b^*\big\} \qquad (D.2)\\
&\ge \frac{1}{q^*} \cdot \sum_{j=0,1,\cdots,i-2} E_j \cdot \mathbf{I}\big\{S_j(m_{\text{init}}) \in V_b^*\big\} \qquad \text{where } q^* \triangleq \max_{i\in[n_{\min}]:\; m_i \in V_b^*} q_b(i) \in (0,\infty)\\
&\overset{d}{=} \sum_{j=0}^{N_{i-2}} \frac{E_j}{q^*} \qquad \text{where } N_i \triangleq \sum_{j=0}^{i} \mathbf{I}\big\{S_j(m_{\text{init}}) \in V_b^*\big\}.
\end{aligned}$$

To proceed with the proof of part (i), we fix some $t > 0$. Note that for any positive integer n, it holds on event $\{\sum_{j=0}^n E_j/q^* > t\} \cap \{N_{i-2} > n\}$ that $\sum_{j=0}^{N_{i-2}} E_j/q^* > t$. This implies $\mathbf{P}(\sum_{j\le i} U_j > t) \ge \mathbf{P}(\sum_{j=0}^n E_j/q^* > t) \cdot \mathbf{P}(N_{i-2} > n)$, due to the independence between $(E_j)_{j\ge1}$ and $(S_j)_{j\ge0}$ (and

hence N_i). Therefore, it suffices to show that, for any $\epsilon > 0$, there exists some positive integer $n = n(\epsilon)$ such that

$$\mathbf{P}(\sum_{j=0}^{n} E_j/q^* > t) > 1 - \epsilon, \qquad \lim_{i\to\infty} \mathbf{P}(N_i > n) = 1. \tag{D.3}$$

Furthermore, the inequality $\mathbf{P}(\sum_{j=0}^{n} E_j/q^* > t) > 1 - \epsilon$ holds for any n large enough due to $q^* \in (0, \infty)$; see (C.3). Meanwhile, since S_n's is irreducible, the chain will visit V_b^* (more generally, any subset of its state space) infinitely often. In other words, for any fixed n, we have $\lim_{i\to\infty} \mathbf{P}(N_i > n) = \lim_{i\to\infty} \mathbf{P}(\#\{j = 0, 1, \ldots, i : \ S_j(m_{\text{init}}) \in V_b^*\} > n) = 1$. This concludes the proof of (D.3).

(*ii*) Fix some $u > 0$ and positive integer i. Representation (D.2) implies $U_1 + \cdots + U_i \overset{d}{=} \sum_{j=1}^{i} C_j \cdot E_j$ for an iid sequence $(E_i)_{i\geq 1}$ of Exponential RVs with rate 1 and another sequence of RVs $(C_i)_{i\geq 1}$ that is independent of $(E_i)_{i\geq 1}$. In particular, C_i's only take values in the set $\mathscr{C} = \{0\} \cup \{1/q_b(i) : \ m_i \in V_b^*\}$, which has finitely many elements. Therefore,

$$\mathbf{P}(U_1 + \cdots + U_i = u) = \sum_{(c_1,\cdots,c_i)\in\mathscr{C}^i} \mathbf{P}(c_1E_1 + \cdots + c_iE_i = u)\mathbf{P}(C_j = c_j \ \forall j \in [i]) = 0$$

due to the absolutely continuity of Exponential RVs.

(*iii*) Recall that $U_1 \equiv 0$. We start by stating a useful property of the mapping Φ. Set $\hat{T}_0 = 1$. For any $k \geq 1$, define (under the convention $U_0 \equiv 0$)

$$\hat{T}_k \triangleq \min\{j > \hat{T}_{k-1} : \ U_j \neq 0\}, \qquad \hat{V}_k \triangleq V_{-1+\hat{T}_k}, \qquad \hat{U}_k \triangleq \sum_{j=\hat{T}_{k-1}}^{-1+\hat{T}_k} U_j = U_{\hat{T}_{k-1}}. \tag{D.4}$$

Note that due to $U_1 \equiv 0$, we have $\hat{T}_1 \geq 2$ and hence $-1+\hat{T}_1 \geq 1$. This confirms that $\hat{V}_1$ is well-defined. In simple terms, $((\hat{U}_k)_{k\geq 1}, (\hat{V}_k)_{k\geq 1})$ can be interpreted as a transformation of $((U_j)_{j\geq 1}, (V_j)_{j\geq 1})$ with consecutive instantaneous jumps grouped together. As a result,

$$\Phi\Big((U_j)_{j\geq 1}, (V_j)_{j\geq 1}\Big) = \Phi\Big((\hat{U}_k)_{k\geq 1}, (\hat{V}_k)_{k\geq 1}\Big). \tag{D.5}$$

To proceed, we consider another representation of the Markov jump process $Y^{*|b}$ based on the following straightforward observation: the law of the process would remain the same if we allow the process to jump from any state m_i to itself at any exponential rate (i.e., by including Markovian "dummy" jumps where the process does not move at all). Specifically, $Y_\cdot^{*|b} \overset{d}{=} \Phi((\tilde{U}_k)_{k\geq 1}, (\tilde{V}_k)_{k\geq 1})$ with $\tilde{U}_k$'s and $\tilde{V}_k$'s defined as follows. Let $\tilde{V}_1$ be sampled from the distribution $\theta_b(\cdot|m_{\text{init}})$ defined in (3.11), and set $\tilde{U}_1 \equiv 0$. Note that so far, we have $(\tilde{U}_1, \tilde{V}_1) \overset{d}{=} (\hat{U}_1, \hat{V}_1)$. Furthermore, for any $t > 0$, $l \geq 1$, and $m_i, \ m_j \in V_b^*$ (with possibly $m_i = m_j$),

$$\begin{aligned}&\mathbf{P}(\tilde{U}_{l+1} < t, \ \tilde{V}_{l+1} = m_j \mid \tilde{V}_l = m_i, (\tilde{V}_j)_{j=1}^{l-1}, \ (\tilde{U}_j)_{j=1}^{l}) = \mathbf{P}(\tilde{U}_{l+1} < t, \ \tilde{V}_{l+1} = m_j \mid \tilde{V}_l = m_i)\\ &= r^{*|b}(i,j) \cdot \big(1 - \exp(-q_b(i)t)\big),\end{aligned} \tag{D.6}$$

where

$$r^{*|b}(i,j) \triangleq \sum_{j'\in[n_{\min}]: \ j'\neq i} \frac{q_b(i,j')}{q_b(i)} \cdot \theta_b(m_j|m_{j'}) \tag{D.7}$$

with $q_b(i)$ and $q_b(i,j)$ defined in (3.10). To see why $Y_\cdot^{*|b} \overset{d}{=} \Phi((\tilde{U}_k)_{k\geq 1}, (\tilde{V}_k)_{k\geq 1})$, note that the process $\Phi((\tilde{U}_k)_{k\geq 1}, (\tilde{V}_k)_{k\geq 1})$ is initialized under $\tilde{V}_1 \sim \theta_b(\cdot|m_{\text{init}})$, which is the same initial distribution of $Y_\cdot^{*|b}$. Moreover, any jump in $\Phi((\tilde{U}_k)_{k\geq 1}, (\tilde{V}_k)_{k\geq 1})$ from $m_i \in V_b^*$ to $m_j \in V_b^*$ (with possibly

$m_i = m_j$) is Markovian and occurs with exponential rate $\sum_{j' \neq i} q_b(i,j')\theta_b(m_j|m_{j'})$. In other words, $\Phi((\tilde{U}_k)_{k\geq 1}, (\tilde{V}_k)_{k\geq 1})$ is simply a reformulation of $Y^{*|b}$ where we include "dummy" jumps from $m_i \in V_b^*$ to itself with exponential rate $\sum_{j' \neq i} q_b(i,j')\theta_b(m_i|m_{j'})$.

In light of (D.5), it only remains to show that

$$\Big((\hat{U}_k)_{k\geq 1}, (\hat{V}_k)_{k\geq 1}\Big) \stackrel{d}{=} \Big((\tilde{U}_k)_{k\geq 1}, (\tilde{V}_k)_{k\geq 1}\Big). \tag{D.8}$$

Specifically, fix some $k \geq 1$, $m_i, m_j \in V_b^*$, and some $t > 0$. Observe that

$$\begin{aligned}
&\mathbf{P}(\hat{U}_{k+1} < t,\ \hat{V}_{k+1} = m_j,\ \hat{V}_k = m_i) \\
&= \sum_{N\geq 1}\sum_{n\geq 1} \mathbf{P}(\hat{U}_{k+1} < t,\ V_{N+n} = m_j,\ \hat{T}_{k+1} - 1 = N + n,\ V_N = m_i,\ \hat{T}_k - 1 = N) \qquad \text{by (D.4)} \\
&= \sum_{N\geq 1}\sum_{n\geq 1} \mathbf{P}(U_{N+1} < t,\ V_p \notin V_b^*\ \forall N+1 \leq p \leq N+n-1; \\
&\qquad\qquad V_{N+n} = m_j,\ \hat{T}_{k+1} - 1 = N + n,\ V_N = m_i,\ \hat{T}_k - 1 = N) \qquad \text{by (D.4) and (C.5)} \\
&= \sum_{N\geq 1}\sum_{n\geq 1}\sum_{(i_1,\cdots,i_{n-1})\in\mathscr{I}(i,n-1)} \mathbf{P}(U_{N+1} < t,\ V_{N+p} = m_{i_p}\ \forall p \in [n-1]; \\
&\qquad\qquad V_{N+n} = m_j,\ \hat{T}_{k+1} - 1 = N + n,\ V_N = m_i,\ \hat{T}_k - 1 = N) \\
&\text{where } \mathscr{I}(i,n-1) \triangleq \big\{(i_1,\ldots,i_{n-1}):\ i_p \neq i_{p-1} \text{ and } m_{i_p} \notin V_b^*\ \forall p \in [n-1]\big\} \text{ with convention } i_0 = i \\
&= \sum_{N\geq 1} \mathbf{P}(V_N = m_i,\ \hat{T}_k - 1 = N) \\
&\quad \cdot \sum_{n\geq 1}\sum_{(i_1,\cdots,i_{n-1})\in\mathscr{I}(i,n-1)} \frac{q_b(i,i_1)}{q_b(i)}\Big(1 - \exp\big(-q_b(i)t\big)\Big)\frac{q_b(i_1,i_2)}{q_b(i_1)}\cdots\frac{q_b(i_{n-2},i_{n-1})}{q_b(i_{n-2})}\frac{q_b(i_{n-1},j)}{q_b(i_{n-1})} \\
&\qquad\qquad \text{using (C.5)} \\
&= \sum_{N\geq 1} \mathbf{P}(V_N = m_i,\ \hat{T}_k - 1 = N) \\
&\quad \cdot \sum_{i_1\neq i} \frac{q_b(i,i_1)}{q_b(i)}\Big(1 - \exp\big(-q_b(i)t\big)\Big) \cdot \sum_{n\geq 1} \mathbf{P}(\tau^S(m_1) = n-1,\ S_{\tau^S(m_1)}(m_1) = m_j).
\end{aligned}$$

In the last line of the display above, we adopt the notations in part (i) and let $S_n(v)$ be a Markov chain with initial value $S_0(v) = v$ and transition kernel $\mathbf{P}(S_{n+1} = m_j | S_n = m_i) = q_b(i,j)/q_b(i)$. Furthermore, let $\tau^S(v) = \min\{n \geq 0:\ S_n(v) \in V_b^*\}$ be the hitting time of any state in V_b^*. Now, observe that

$$\begin{aligned}
&\mathbf{P}(\hat{U}_{k+1} < t,\ \hat{V}_{k+1} = m_j,\ \hat{V}_k = m_i) \\
&= \sum_{N\geq 1} \mathbf{P}(V_N = m_i,\ \hat{T}_k - 1 = N) \cdot \sum_{i_1\in[n_{\min}]:\ i_1\neq i} \frac{q_b(i,i_1)}{q_b(i)}\Big(1 - \exp\big(-q_b(i)t\big)\Big)\theta_b(m_j|m_{i_1}) \\
&= \sum_{N\geq 1} \mathbf{P}(V_N = m_i,\ \hat{T}_k - 1 = N) \cdot r^{*|b}(i,j) \cdot \Big(1 - \exp\big(-q_b(i)t\big)\Big) \qquad \text{with } r^{*|b}(\cdot,\cdot) \text{ defined in (D.7)} \\
&= r^{*|b}(i,j) \cdot \Big(1 - \exp\big(-q_b(i)t\big)\Big) \cdot \mathbf{P}(\hat{V}_k = m_i).
\end{aligned}$$

This verifies $\mathbf{P}(\hat{U}_{k+1} < t,\ \hat{V}_{k+1} = m_j \mid \hat{V}_k = m_i) = r^{*|b}(i,j) \cdot \big(1 - \exp(-q_b(i)t)\big)$. Through (D.6) we conclude the proof of (D.8). □

E Technical Lemmas for Measures $\check{\mathbf{C}}^{(k)|b}$ and First Exit Analysis

This section collects useful results established in [34] for the measure $\check{\mathbf{C}}^{(k)|b}(\cdot)$ defined in (2.14). Throughout this section, we impose Assumptions 2, 3, and 5 on some $I = (s_{\text{left}}, s_{\text{right}})$ where $s_{\text{left}} < 0 < s_{\text{right}}$, and fix some $b > 0$ such that $s_{\text{left}}/b \notin \mathbb{Z}$ and $s_{\text{right}}/b \notin \mathbb{Z}$. With this, for $l = \inf_{x \in I^c} |x| = |s_{\text{left}}| \wedge s_{\text{right}}$ we have $l > (\mathcal{J}_b^* - 1)b$. This allows us to fix, throughout this section, some $\bar{\epsilon} > 0$ small enough such that

$$\bar{\epsilon} \in (0, 1 \wedge b), \qquad l > (\mathcal{J}_b^* - 1)b + 3\bar{\epsilon}. \tag{E.1}$$

Next, for any $\epsilon \in (0, \bar{\epsilon})$, let

$$\boldsymbol{t}(\epsilon) \triangleq \min\left\{t \geq 0 : \ \boldsymbol{y}_t(s_{\text{left}} + \epsilon) \in [-\epsilon, \epsilon] \text{ and } \boldsymbol{y}_t(s_{\text{right}} - \epsilon) \in [-\epsilon, \epsilon]\right\} \tag{E.2}$$

for the ODE $\boldsymbol{y}_t(x)$ defined in (2.11). Also, recall that $I_\epsilon \triangleq (s_{\text{left}} + \epsilon, s_{\text{right}} - \epsilon)$ is the ϵ-shrinkage of set I. We use $I_\epsilon^- = [s_{\text{left}} + \epsilon, s_{\text{right}} - \epsilon]$ to denote the closure of I_ϵ. Then, the definition of $\boldsymbol{t}(\cdot)$ implies

$$\boldsymbol{y}_t(y) \in [-\epsilon, \epsilon] \qquad \forall y \in I_\epsilon^-, \ t \geq \boldsymbol{t}(\epsilon). \tag{E.3}$$

Lemma E.1 (Lemma D.2 of [34]). *Let Assumptions 2, 3, and 5 hold. Let $\bar{\epsilon} \in (0, b)$ be defined as in (E.1). For any $|\gamma| > (\mathcal{J}_b^* - 1)b + \bar{\epsilon}$ such that $\gamma/b \notin \mathbb{Z}$,*

$$\check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(\{\gamma\}) = 0.$$

Lemma E.2 (Lemma D.3 of [34]). *If Assumptions 2, 3, and 5 hold, then $\check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(I^c) \in (0, \infty)$.*

Lemma E.3 (Lemma D.4 of [34]). *Let Assumptions 2 and 5 hold. Let $\bar{\epsilon} \in (0, b)$ be defined as in (E.1). Given any open interval $S \subseteq \mathbb{R}$, let*

$$r_S \triangleq \inf\{|x| : \ x \in S\}, \qquad d_S \triangleq \lceil r_S/b \rceil.$$

If $d_S \geq k$ and $r_S - (d_S - 1) \cdot b > \bar{\epsilon}$ for some positive integer k, then

$$\check{\mathbf{C}}^{(k)|b}(S) > 0 \qquad \Longleftrightarrow \qquad d_S = k.$$

Next, we collect a few useful results in [34] regarding the behavior of $X_j^{\eta|b}(x)$ before exiting from (some subset of) I or returning to the neighborhood of the origin. The proofs hinge on the sample path large deviations developed in Result 1. Specifically, let $\tau_1^{>\delta}(\eta) = \min\{n \geq 0 : \ \eta|Z_n| > \delta\}$ be the arrival time of the first Z_n with $\eta|Z_n| > \delta$ (i.e., the first "large" noise). The next result states that it is unlikely for $X_j^\eta(x)$ to take long before exiting from I_ϵ or returning to $(-\epsilon, \epsilon)$.

Lemma E.4 (Lemma 4.4 of [34]). *Let Assumptions 1, 2, 3, and 5 hold. Given any $k \geq 1$ and $\epsilon \in (0, \bar{\epsilon})$, it holds for all $T \geq k \cdot \boldsymbol{t}(\epsilon/2)$ that*

$$\lim_{\eta \downarrow 0} \sup_{x \in I_\epsilon^-} \frac{1}{\big(\lambda(\eta)\big)^{k-1}} \mathbf{P}\Big(X_j^{\eta|b}(x) \in I_\epsilon \setminus (-\epsilon, \epsilon) \quad \forall j \leq T/\eta\Big) = 0$$

where $\lambda(\eta) = \eta^{-1}\mathbf{P}(|Z| > \eta^{-1})$.

Let $R_\epsilon^{\eta|b}(x) \triangleq \min\left\{j \geq 0 : \ X_j^{\eta|b}(x) \in (-\epsilon, \epsilon)\right\}$ be the first time $X_j^{\eta|b}(x)$ returned to the ϵ-neighborhood of the origin. The next result verifies that, initialized within the attraction field I, $X_j^{\eta|b}(x)$ would return to $(-\epsilon, \epsilon)$ efficiently with high probability.

Lemma E.5 (Lemma 4.5 of [34]). *Let Assumptions 1, 2, 3, and 5 hold. Let $\boldsymbol{t}(\cdot)$ be defined as in (E.2) and*

$$E(\eta,\epsilon,x) \triangleq \Big\{ R_\epsilon^{\eta|b}(x) \le \frac{\boldsymbol{t}(\epsilon/2)}{\eta};\ X_j^{\eta|b}(x) \in I_{\epsilon/2}\ \forall j \le R_\epsilon^{\eta|b}(x) \Big\}.$$

For each $\epsilon \in (0,\bar\epsilon)$ we have $\lim_{\eta\downarrow 0} \sup_{x\in I_\epsilon^-} \mathbf{P}\Big(\big(E(\eta,\epsilon,x)\big)^c \Big) = 0$.

Lastly, we show that it is unlikely for $X_j^{\eta|b}(x)$ to deviate far from the origin without any "large" Z_n. Again, the proof makes heavy use of results in [34].

Lemma E.6. *Let Assumptions 1, 2, 3, and 5 hold. Given any $\epsilon \in (0,\bar\epsilon)$ and positive integer N, there is some $\bar\delta > 0$ such that*

$$\lim_{\eta\downarrow 0} \sup_{x\in(-\frac{\epsilon}{2},\frac{\epsilon}{2})} \mathbf{P}\Big(X_j^{\eta|b}(x) \notin (-\epsilon,\epsilon) \text{ for some } j < \tau_1^{>\delta}(\eta) \Big) \Big/ \eta^N = 0 \qquad \forall \delta \in (0,\bar\delta).$$

Proof. Note that the values of $a(\cdot)$ and $\sigma(\cdot)$ outside of $[-\epsilon,\epsilon] \subseteq [-\bar\epsilon,\bar\epsilon]$ has no impact on the first exit time from $(-\epsilon,\epsilon)$ when starting from $(-\epsilon/2,\epsilon/2)$. In light of Assumption 3, by modifying the values of $a(\cdot)$ and $\sigma(\cdot)$ outside of $[-\bar\epsilon,\bar\epsilon]$ we can assume w.l.o.g. the existence of some $0 < c \le C < \infty$ that

$$\inf_{x\in\mathbb{R}} \sigma(x) \ge c, \qquad \sup_{x\in\mathbb{R}} \sigma(x) \vee |a(x)| \le C. \tag{E.4}$$

For any $r > 0$, let $T_r^\eta(x) \triangleq \min\{j \ge 0 :\ X_j^{\eta|b}(x) \notin (-r,r)\}$. Due to the monotonicity in $\tau_1^{>\delta'}(\eta) \le \tau_1^{>\delta}(\eta)$ for any $0 < \delta' < \delta$, it suffices to show that for any positive integer N and any small enough $\epsilon > 0$, there is some $\delta = \delta(N,\epsilon) > 0$ such that

$$\limsup_{\eta\downarrow 0} \sup_{x\in(-\epsilon,\epsilon)} \mathbf{P}\big(T_{2\epsilon}^\eta(x) < \tau_1^{>\delta}(\eta)\big) \big/ \eta^N = 0. \tag{E.5}$$

Fix some $\beta > \alpha$ where $\alpha > 1$ is specified in Assumption 1. Also, pick some $\theta \in (0,\beta-\alpha)$. Applying Lemma 4.6 (i) of [34], we see that the claim $\mathbf{P}(\tau_1^{>\delta}(\eta) > 1/\eta^\beta) = \boldsymbol{o}\big(\exp(-1/\eta^\theta)\big)$ (as $\eta \downarrow 0$) holds for any $\delta > 0$. Also, note that $\tau_1^{>\delta}(\eta)$ only takes integer values, and observe that

$$\big\{T_{2\epsilon}^\eta(x) < \tau_1^{>\delta}(\eta)\big\} \subseteq \big\{T_{2\epsilon}^\eta(x) < \tau_1^{>\delta}(\eta) \le 1/\eta^\beta\big\} \cup \big\{\tau_1^{>\delta}(\eta) > 1/\eta^\beta\big\}.$$

Therefore, to prove (E.5) we only need to find some $\delta > 0$ such that

$$\sup_{x\in(-\epsilon,\epsilon)} \mathbf{P}\big(T_{2\epsilon}^\eta(x) < \tau_1^{>\delta}(\eta) \le \lfloor 1/\eta^\beta \rfloor\big) = \boldsymbol{o}(\eta^N) \quad \text{ as } \eta \downarrow 0. \tag{E.6}$$

Recall the definition of $\boldsymbol{t}(\epsilon)$ in (E.2). Let $t \triangleq \boldsymbol{t}(\epsilon/2) < \infty$ and $K(\eta) \triangleq \lceil \frac{\lfloor 1/\eta^\beta \rfloor}{\lfloor t/\eta \rfloor} \rceil$. Note that $K(\eta) = \boldsymbol{O}(1/\eta^{\beta-1})$. Next, we fix some $\widetilde\epsilon > 0$ small enough such that $2\exp(tD)\widetilde\epsilon < \epsilon/2$, with $D < \infty$ being the Lipschitz constant in Assumption 2. Define events

$$\widetilde A_k(\eta,x) \triangleq \Big\{ \max_{(k-1)\lfloor \frac{t}{\eta} \rfloor + 1 \le j \le k\lfloor \frac{t}{\eta} \rfloor \wedge \left(\tau_1^{>\delta}(\eta)-1\right)} \eta \Big| \sum_{i=(k-1)\lfloor \frac{t}{\eta} \rfloor + 1}^{j} \sigma\big(X_{i-1}^{\eta|b}(x)\big) Z_i \Big| \le \widetilde\epsilon \Big\}.$$

For any $x \in (-\epsilon,\epsilon)$, any $\delta \in (0, \frac{b}{2C})$ and any $\eta \in (0, \frac{\widetilde\epsilon}{C} \wedge \frac{b\wedge 1}{2C})$ (where C is specified in (E.4)), on event $\widetilde A_1(\eta,x)$ we observe the following facts. First, from part (b) of Lemma 3.7 in [34],

$$\sup_{s \le \frac{t}{\eta} \wedge \left(\tau_1^{>\delta}(\eta)-1\right)} \Big| \boldsymbol{y}_{\eta s}(x) - X_{\lfloor s \rfloor}^{\eta|b}(x) \Big| < \exp(tD)\widetilde\epsilon + \exp(tD)\eta C < 2\exp(tD)\widetilde\epsilon < \epsilon/2$$

due to our choice of η and $\widetilde{\epsilon}$ above. Next, by Assumption 5, we have $\boldsymbol{y}_s(x) \in (-\epsilon, \epsilon)$ $\forall s \geq 0$ and $\boldsymbol{y}_t(x) \in (-\epsilon/2, \epsilon/2)$. Combining these two facts, we get that $X_s^{\eta|b}(x)$ for all $s \leq \lfloor t/\eta \rfloor \wedge \big(\tau_1^{>\delta}(\eta) - 1\big)$ and, in case that $\tau_1^{>\delta}(\eta) > \lfloor t/\eta \rfloor$, we must have $X_{\lfloor t/\eta \rfloor}^{\eta|b}(x) \in (-\epsilon, \epsilon)$. By repeating this argument inductively for $k = 2, 3, \cdots, K(\eta)$, we can see that for any $x \in (-\epsilon, \epsilon)$, any $\delta \in (0, \frac{b}{2C})$, and any $\eta \in (0, \frac{\widetilde{\epsilon}}{C} \wedge \frac{b \wedge 1}{2C})$, it holds on event $\bigcap_{k=1}^{K(\eta)} \widetilde{A}_k(\eta, x)$ that

$$X_j^{\eta|b}(x) \in (-2\epsilon, 2\epsilon) \qquad \forall j \leq \lfloor 1/\eta^\beta \rfloor \wedge \big(\tau_1^{>\delta}(\eta) - 1\big) \leq K(\eta) \lfloor t/\eta \rfloor \wedge \big(\tau_1^{>\delta}(\eta) - 1\big).$$

As a result, for any $x \in (-\epsilon, \epsilon)$, any $\delta \in (0, \frac{b}{2C})$, and any $\eta \in (0, \frac{\widetilde{\epsilon}}{C} \wedge \frac{b \wedge 1}{2C})$,

$$\sup_{x \in (-\epsilon, \epsilon)} \mathbf{P}\Big(T_{2\epsilon}^{\eta}(x) < \tau_1^{>\delta}(\eta)\Big) \leq \sup_{x \in (-\epsilon, \epsilon)} \mathbf{P}\Big(\bigcup_{k=1}^{K(\eta)} \Big(\widetilde{A}_k(\eta, x)\Big)^c \Big).$$

Lastly, due to part (a) of Lemma 3.3 of [34], the claim $\sup_{k \in [K(\eta)]} \sup_{x \in (-\epsilon, \epsilon)} \mathbf{P}\Big(\big(\widetilde{A}_k(\eta, x)\big)^c\Big) = \boldsymbol{o}(\eta^{N+\beta-1})$ holds for all $\delta > 0$ small enough, which leads to

$$\sup_{x \in (-\epsilon, \epsilon)} \mathbf{P}\Big(T_{2\epsilon}^{\eta}(x) < \tau_1^{>\delta}(\eta)\Big) \leq K(\eta) \cdot \boldsymbol{o}(\eta^{N+\beta-1}) \leq \boldsymbol{O}(1/\eta^{\beta-1}) \cdot \boldsymbol{o}(\eta^{N+\beta-1}) = \boldsymbol{o}(\eta^N).$$

This verifies claim (E.6) and concludes the proof. □

F Details of Experiments

F.1 Details of the $\mathbb{R}^1$ simulation experiment

The function f used in the experiments is

$$\begin{aligned} f(x) = {} & (x + 1.6)(x + 1.3)^2 (x - 0.2)^2 (x - 0.7)^2 (x - 1.6) \big(0.05|1.65 - x|\big)^{0.6} \\ & \cdot \Big(1 + \frac{1}{0.01 + 4(x - 0.5)^2}\Big)\Big(1 + \frac{1}{0.1 + 4(x + 1.5)^2}\Big)\Big(1 - \frac{1}{4}\exp(-5(x + 0.8)(x + 0.8))\Big). \end{aligned} \tag{F.1}$$

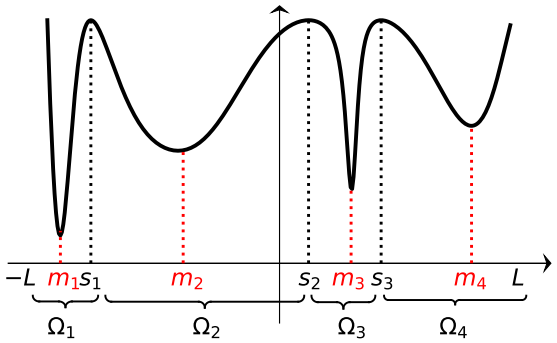

Figure F.1: Illustration of the test function f used in the $\mathbb{R}^1$ experiment.

As shown in Figure F.1, the four isolated local minimizers of f are $m_1 = -1.51, s_1 = -1.3, m_2 = -0.66, s_2 = 0.2, m_3 = 0.49, s_3 = 0.7, m_4 = 1.32$, and in our experiment we restrict the iterates on

$[-L, L]$ with $L = 1.6$. The heavy-tailed noises we used in the experiment were $Z_n = 0.1 U_n W_n$ where W_n were sampled from Pareto Type II distribution (aka Lomax distribution) with shape parameter $\alpha = 1.2$, and the signs U_n were iid RVs such that $\mathbf{P}(U_n = 1) = \mathbf{P}(U_n = -1) = 1/2$.

In the first exit time experiment, we tested three different settings: (a) $b = 0.28$ (so that $l^* = 3$); (b) $b = 0.5$ (so that $l^* = 2$); (c) no gradient clipping (so that $l_2^* = 1$). For the first case, we tested learning rates $\{0.1, 0.05, 0.03, 0.02, 0.01, 0.005, 0.003, 0.001\}$, while for the other two cases, we tested learning rates $\{0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001\}$. For each case, we ran the simulation 20 times and plotted the average of the 20 exit times. Lastly, to prevent excessively long running time of the experiment, the simulation was terminated when the iteration number reached 5×10^7. This threshold was reached only in the setting with $\eta = 0.001, b = 0.28$.

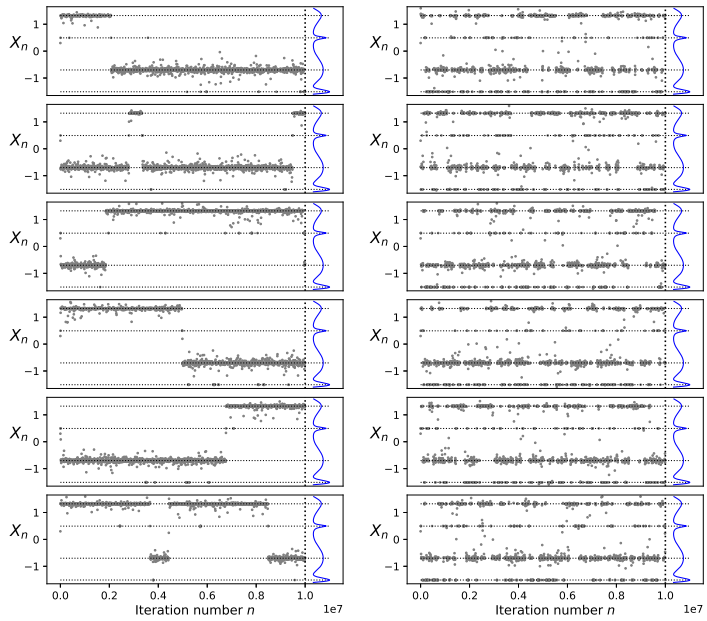

Figure F.2: Five sample paths of SGD under heavy-tailed noises with gradient clipping (left) and without gradient clipping (right). Note that in each case, SGD sample paths exhibit similar patters: SGD almost completely avoided sharp minima with gradient clipping, whereas SGD spent significant amount of time at the sharp minima without gradient clipping.

Next, we present extra sample paths of SGD when applied to function f in (F.1) in Figure F.2 and F.3. The blue curve on the right side of each plot shows f rotated by 90 degrees, and the dashed lines indicate the locations of local minima. For better readability of the figures, we plotted X_n for every 5,000 iterations. To generate these plots, we initialized the SGD iterates at 0.3 (so that it is in $\Omega_3 = (0.2, 0.7)$) and fixed the learning rate as $\eta = 0.001$. Again, we tested both with gradient clipping (with $b = 0.5$) and without gradient clipping. Moreover, we also tested **light-tailed** noises where we use $N(0, 1)$ as the distribution for noises Z_n. For each sample path of X_n, we run $10,000,000$ iterations. In the left plots of Figure F.2, one can see that with clipped heavy-tailed stochastic gradients, the SGD iterates almost always stay around the wide attraction fields, and the sharp minima are almost completely eliminated from the trajectories of SGD. In comparison, in the right plots of Figure F.2 one can see that without gradient clipping, the heavy-tailed noises will drive SGD to spend substantial amount of time in all the different local minima, including the sharp ones. Lastly, from Figures F.3, one can see that under light-tailed noises and small learning rates, SGD

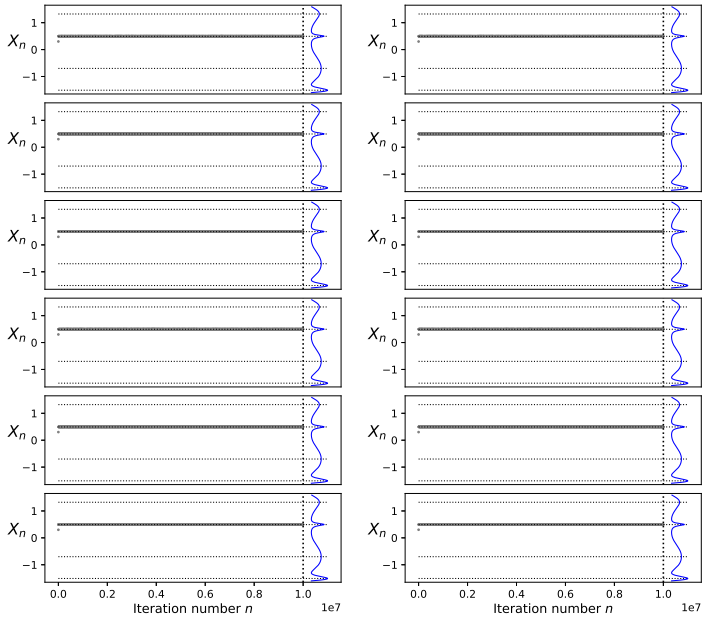

Figure F.3: Five sample paths of SGD under light-tailed noises with gradient clipping (left) and without gradient clipping (right). Note that regardless of the use of gradient clipping, SGD never manages to escape the local minimum that it started from.

cannot escape a sharp minima once trapped there.

F.2 Details of the $\mathbb{R}^d$ simulation experiment

As illustrated in the contour plot in Figure 4.1 (a), the function f in this experiment is a modified version of Himmelblau function, a commonly used test function for optimization algorithm. The modifications serve two purposes. First, as shown in Figure 4.1 (b), for the modified function the four attraction fields $\Omega_1, \Omega_2, \Omega_3, \Omega_4$ have different sizes; in particular, under gradient clipping threshold $b = 2.15$, from the local minimizers of Ω_1 and Ω_2 (indicated by red dots in the corresponding area) at least two jumps are required to escape from the attraction field, while from the local minimizer in Ω_3 or Ω_4 it is possible to escape with one jump. Therefore, for the minimum jump number required to escape, we have $l_1^* = l_2^* = 2 > l_3^* = l_4^* = 1$ in this case. Second, for the modified test function f, the local minimizer in Ω_2 is not a single point but a connected line segment, which is indicated by the dark line in bottom-left region in Figure 4.1 (a) and the red line segment in in Figure 4.1 (b). Therefore, the modification allows us to test the heavy-tailed SGD methods on a more general loss landscape.

Now we describe the construction of the test function f. Let h be the Himmelblau function with expression $h(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$. Next, define the following transformation for coordinates: $\phi(x, y) = \Big(x(\exp(c_0(x - c_x) + 1)), y(\exp(c_0(x - c_x) + 1)\Big)$. Let the composition be $h_\phi(x, y) = h\Big(\phi(x - a_x, y)\Big)$. To create the connected region of local minimizers, define the following locally "cut" version of h_ϕ:

$$
\begin{aligned}
i(x, y) &= \mathbb{1}\{x \in [b_l, b_r],\ |y - a_y| < b_y\}, \\
h^*(x, y) &= (1 - i(x, y))h_\phi(x, y) + i(x, y) \min\{h_\phi(x, y),\ c_1|y - a_y|^{1.1}\}.
\end{aligned}
$$

In other words, by taking minimum of the original h_ϕ and a polynomial function w.r.t. y around the original local minimizer of Ω_2, we obtain a function h^* that attains local minimum on an entire line segment with $y = a_y$. Lastly, the test function we use in the experiment is $f = 0.1h^*$, with $a_x = 1.5, a_y = -2.9, b_l = -5.5, b_r = -0.5, b_y = 2.0, c_0 = 0.4, c_1 = 12$.

In the experiment, we initialize the SGD iterates X_k at $X_0 = (2.9, 1.0)$, which is very close to the local minimizer in the small attraction field Ω_3. For both the clipped and unclipped SGD, we perform updates for 3×10^7 steps, under learning rate 5×10^{-4} and heavy-tailed noise $Z_k = 0.75W_k$ where the iid samples W_k are isotropic and the law of $\|W_k\|$, the size of the noise, is Pareto(1.2). For clipped SGD, we use threshold $b = 2.15$. To prevent the iterates from drifting to infinity, after each update X_k is projected back to the L_2 ball centered at origin with radius 4.2 whenever X_k leaves this ball.

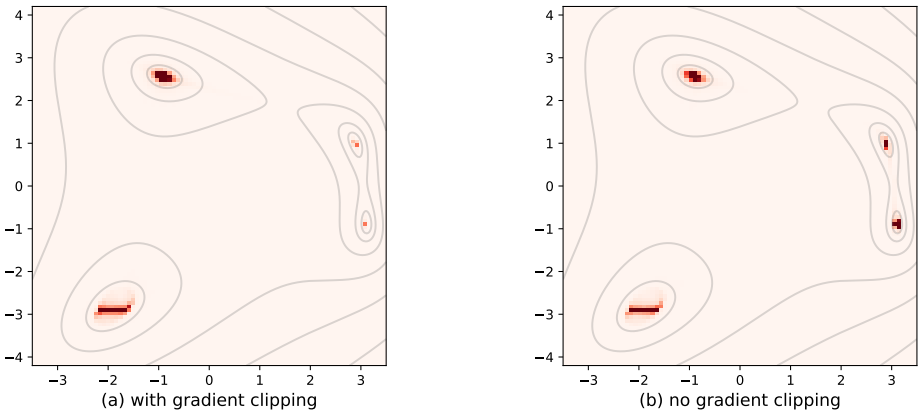

Figure F.4: Heat map of SGD iterates when optimizing the modified Himmelblau function.

In Figure F.4, we use the 3×10^7 steps of SGD iterates (for both the clipped and unclipped case) to create heat maps showing locations of SGD iterates. From this figure, two points can be made clear: first, the heavy-tailed SGD does spend much less time at the two small attraction fields when gradient clipping is applied; second, in Ω_2 (the bottom-left attraction field) the SGD iterates frequent the entire connected region of local minima instead of a certain point on this line segment.

F.3 Details of the ablation study

We first mention that the all experiments using neural networks are conducted on Nvidia GeForce GTX 1080 Ti. For the ablation study, the experiments and scripts are adapted from the ones in [37].[2].

In Figure F.5, we display the gradient noise distribution in the three tasks of the ablation study after the model is randomly initialized.

The experiment setting and choice of hyperparameters are mostly adapted from the experiment in [37]. We consider three different tasks: (1) training LeNet on corrupted FashionMNIST dataset; specifically, we use a 1200-sample subset of the original FashionMNIST training dataset, and for 200 samples points in the training set we randomly assign a label instead of using the correct ones; (2) VGG11 on SVHN dataset, where we use a 25000-sample subset of the training dataset; (3) VGG11 on CIFAR10, where we use the entire training set. For all tasks we use the entire test dataset when evaluating test accuracy.

The heavy-tailed multipliers Z_n used in this experiment, whenever heavy-tailed noise is needed, are $Z_n = cW_n$ where W_n are iid Pareto(α) RVs. For each task, we first randomly initialize each model,

[2]https://github.com/uuujf/SGDNoise

Table F.1: Test accuracy (percentage) and expected sharpness of different methods across different tasks. The reported numbers are the averages and 95%CI over 5 replications.

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
FMNIST, LeNet	68.7±0.4	69.2±0.8	68.8±0.6	64.4±3.4	69.5±0.8	**70.1±0.4**
SVHN, VGG11	82.9±0.4	85.9±0.2	85.9±0.2	38.9±24.1	**88.4±0.2**	**88.4±0.2**
CIFAR10, VGG11	69.4±0.5	74.4±0.4	74.4±0.8	40.5±25.1	75.7±1.1	**75.9±0.7**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
FMNIST, LeNet	0.032±0.006	0.008±0.001	0.009±0.001	0.047±0.02	0.003±0.0003	**0.002±0.0002**
SVHN, VGG11	0.694±0.048	0.037±0.007	0.041±0.006	0.012±0.009	**0.002±0.0007**	0.005±0.004
CIFAR10, VGG11	2.043±0.083	0.050±0.013	0.039±0.019	2.046±2.4	**0.024**±0.005	0.037±0.007

Table F.2: Hyperparameters for training in the ablation study

Hyperparameters	FashionMNIST, LeNet	SVHN, VGG11	CIFAR10, VGG11
learning rate	0.05	0.05	0.05
batch size for g_{SB}	100	100	100
training iterations	10,000	30,000	30,000
gradient clipping threshold	5	20	20
c	0.5	0.5	0.5
α	1.4	1.4	1.4

Table F.3: Sharpness of different methods across different tasks. The reported numbers are the averages over 5 replications.

PAC-Bayes Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
FashionMNIST, LeNet	5.9×10^3	3×10^3	3.3×10^3	3.1×10^3	1.9×10^3	$\mathbf{1.6 \times 10^3}$
SVHN, VGG11	2.97×10^4	6.9×10^3	7.3×10^3	7.76×10^4	$\mathbf{2.1 \times 10^3}$	2.3×10^3
CIFAR10, VGG11	4.87×10^4	7.2×10^3	6.8×10^3	6.74×10^4	$\mathbf{4.8 \times 10^3}$	5.8×10^3
Maximal Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
FashionMNIST, LeNet	1.01×10^4	4.9×10^3	5.4×10^3	5.4×10^3	3.2×10^3	$\mathbf{2.5 \times 10^3}$
SVHN, VGG11	3.78×10^4	9.1×10^3	9.3×10^3	1.19×10^5	$\mathbf{2.5 \times 10^3}$	2.8×10^3
CIFAR10, VGG11	5.46×10^4	8.5×10^3	8×10^3	1.18×10^5	$\mathbf{5.8 \times 10^3}$	6.5×10^3

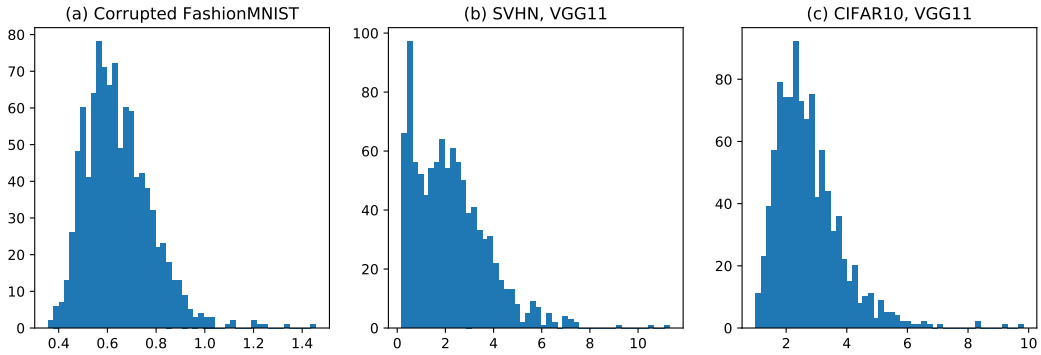

Figure F.5: Distribution of gradient noise in different tasks of the ablation study.

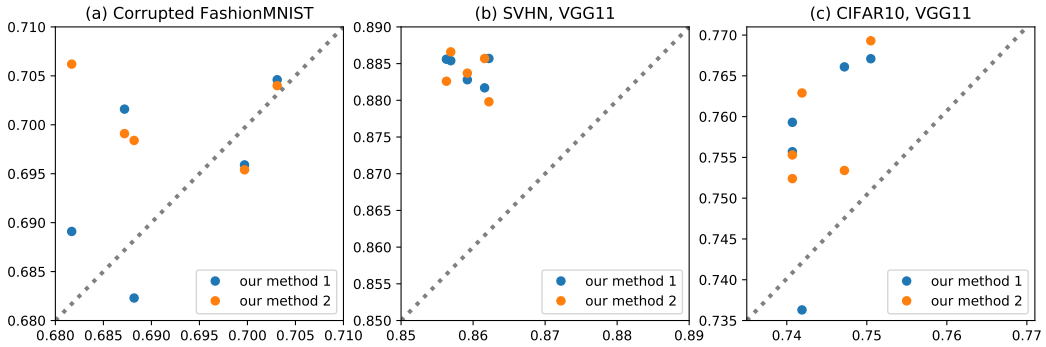

Figure F.6: Test accuracy of the proposed clipped heavy-tailed methods vs. test accuracy of vanilla SGD in the ablation study.

and then run the 6 candidate methods in parallel starting from the same randomly initialized model weights for a fair comparison.

The hyperparameters in training for each task are listed in Table F.2. The same set of hyperparameters is used for all methods in the same task. Whenever gradient clipping scheme is applied, we clip the gradient if its L_2 norm exceeds the threshold given in Table F.2. The exception here is the "*SB + Noise*" method: we use learning rate $\eta = 0.005$; for FashionMNIST task we train for 100,000 iterations and the heavy-tailed noise is removed for the final 50,000 iterations; for SVHN and CIFAR10 tasks, we train for 150,000 iterations and heavy-tailed noise is removed for the last 70,000 iterations. Besides, for this method we always clip the model weights if its L_∞ norm exceeds 1. The reason for the extra tuning and extended training in "*SB + Noise*" method is that, without the said modifications, in all three tasks we observed that the model weights quickly drift to infinity and explodes; even with the weight clipping implemented, the model performance stays at random level with no signs of improvements if we do not tune down learning rate.

In Table F.3, we also report the sharpness of solutions under different shaprness metrics. First, the *PAC-Bayes Sharpness* metric (see equation (53) in [13]) is defined as $1/\sigma^2$ where σ is equal to the smallest δ that induces a 0.1 expected sharpness, and reflects the sharpness/flatness parameter used in studies on generalization gaps under the PAC-Bayes framework (see [23]). Besides, the *Maximal Sharpness* metric (see equation (54) in [13]) is defined as $1/\sigma^2$ where σ is equal to the smallest radius

Table F.4: Results and 95% CI in the experiments with data augmentation.

Test Accuracy	SB + Clip	Our 1	Our 2
CIFAR10, VGG11	89.5±0.2	**90.7±0.1**	90.5±0.2
CIFAR100, VGG16	56.3±0.3	**65.4±1.2**	63.0±2.5
Expected Sharpness	SB + Clip	Our 1	Our 2
CIFAR10, VGG11	0.17±0.005	**0.09±0.004**	0.10±0.003
CIFAR100, VGG16	0.86±0.02	**0.44±0.05**	0.48±0.07

Table F.5: Sharpness of solutions obtained by different methods in CIFAR10/100 tasks with data augmentation. Numbers reported here are the average of 5 replications.

CIFAR10-VGG11	SB + Clip	Our 1	Our 2
Expected Sharpness	0.167	**0.085**	0.096
PAC-Bayes Sharpness	1.31×10^4	$\mathbf{9 \times 10^3}$	10^4
Maximal Sharpness	1.66×10^4	1.29×10^4	$\mathbf{1.22 \times 10^4}$
CIFAR100-VGG16	SB + Clip	Our 1	Our 2
Expected Sharpness	0.857	**0.441**	0.479
PAC-Bayes Sharpness	2.49×10^4	$\mathbf{1.9 \times 10^4}$	1.98×10^4
Maximal Sharpness	2.75×10^4	$\mathbf{2.12 \times 10^4}$	2.16×10^4

δ that makes $\max_{\|\nu\|_\infty \leq \delta} |L(\theta^* + \nu) - L(\theta^*)| \geq 0.1$, and metrics of form $\max_{\|\nu\| \leq \delta} |L(\theta^* + \nu) - L(\theta^*)|$ can be considered as a proxy for the spectral norm of the Hessian at the solution (see [2]). It worth noticing that, for all three sharpness metrics, the smaller the value is the "flatter" the loss landscape is around the solution. Lastly, for evaluation of the PAC-Bayes Sharpness and Maximal Sharpness metrics, we conduct binary search as in Algorithm 2 of [13] with $\epsilon_d = 0.01, \epsilon_\sigma = 0, M_1 = 10$ and $M_2 = 100$; in our setting we always evaluate the training loss using one sweep of the entire training set, so M_3 is a case-specific and is equal to the number of batches of the training set under the batch size for the task at hand.

In Figure F.6, we plot the test accuracy of our method against that of the SGD for all 5 replications and 3 tasks.

F.4 Details of CIFAR10/100 experiments with data augmentation

For both methods, we train the model for 300 epochs and set the initial learning rate as 0.1. In our method, the training can be partitioned into two phases. In the first phase (the first 200 epochs), the learning rate is kept at a constant. In the second phase, for every 30 epoch we reduce the learning rate by half. Also, an L_2 weight decaying with coefficient 5×10^{-4} is enforced. As for parameters for heavy-tailed noises in (5.1), we use $c = 0.5$ and $\alpha = 1.4$ in the first phase, and in the second phase we remove heavy-tailed noise and use SB to update weights. In both methods for the small-batch direction g_{SB} the batch size is 128, while for g_{LB} we evaluate the gradient on a large sample batch of size 1,024. Under the epoch number 300 and batch size 128, the count of total iterations performed during training is 1.17×10^5. To augment the dataset, random horizontal flipping and cropping with padding size 4 is applied for each training batch. Lastly, gradient clipping scheme is applied for both methods, and we fix $b = 0.5$. In other words, when the learning rate is η (note that due to the scheduling of learning rates, η will be changing throughout the training), the gradient is clipped if its L_2 norm is larger than b/η. The scripts are adapted from the ones in https://github.com/chengyangfu/pytorch-vgg-cifar10.

These results are presented in Table 5.2. Furthermore, in Table F.5 we see that our truncated heavy-tailed method also manages to find solutions with a flatter geometry.