Large Deviations and Metastability Analysis for Heavy-Tailed Dynamical Systems

Xingyu Wang
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
xingyuwang2017@u.northwestern.edu

Chang-Han Rhee
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
chang-han.rhee@northwestern.edu

Abstract

We study large deviations and metastability of heavy-tailed stochastic dynamical systems and provide the heavy-tailed counterparts of the classical Freidlin-Wentzell and Eyring-Kramers theory. Our findings address the rare-event analysis for sufficiently general events and heavy-tailed dynamical systems. We also unveil an intricate phase transitions in the first exit problems under truncated heavy-tailed noises. Furthermore, our results provide tools to systematically study the connection between the global dynamics of the stochastic gradient descent (SGD) under heavy-tailed noises and the generalization mystery of deep learning.

1 Introduction

The study of large deviations and metastability for stochastic systems has a long and illustrious history. For instance, the Freidlin-Wentzell theorem (see Ventsel' and Freidlin [1970]) characterized the large deviations for Itô diffusions at the sample-path level. During the last few decades, the Freidlin-Wentzell theorem has received several significant extensions, including the discrete-time results (see, e.g., Kifer [1990]), results under relaxed assumptions (see, e.g., Donati-Martin [2008], Dupuis and Ellis [2011]), and generalizations to infinite dimensional processes (see, e.g., Budhiraja and Dupuis [2000]). When it comes to metastability analysis, the Eyring–Kramers law (refer to Eyring [1935], Kramers [1940]) is arguably the most renowned result. It provides a comprehensive characterization of the asymptotics of first exit times for dynamics driven by Brownian motion. See also theoretical advancements such as the characterization of the most likely exit path for Brownian particles in more complicated settings (e.g., M. I. Freidlin [1998]).

Given the prevalence of heavy-tailed phenomena in various deep learning tasks (for example, see Simsekli et al. [2019], Gurbuzbalaban et al. [2021]), it becomes particularly important to investigate the fundamental properties of algorithms in the heavy-tailed contexts, including convergence rates (see, e.g., Wang et al. [2021], Zhang et al. [2020]) and stationary distributions (see, e.g., Hodgkinson and Mahoney [2020], Gurbuzbalaban et al. [2020]). In particular, it is worth noticing that, in sharp contrast to the classical light-tailed analyses, stochastic dynamical systems exhibit fundamentally different large deviations and metastability behaviors under heavy-tailed perturbations. As shown in Imkeller and Pavlyukevich [2006], Imkeller et al. [2010, 2009], Imkeller, Peter and Pavlyukevich, Ilya [2008], when the continuous-time stochastic differential equations (SDEs) are driven by heavy-tailed

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

dynamics, the exits are typically caused by large perturbations of a small number of components rather than smooth tilting of the dynamics. Recently, Nguyen et al. [2019] established that, when viewed as a discretized version of SDEs driven by α-stable Lévy motions, heavy-tailed stochastic gradient descents (SGDs) exhibit first exit behaviors similar to their continuous-time counterparts (i.e., heavy-tailed SDEs). Besides, Wang et al. [2022] made observations of a much richer mathematical structure in the first exit times and global dynamics of SGD algorithms when heavy-tailed noises are truncated using gradient clipping, which is commonly employed in many contexts; see, e.g., Engstrom et al. [2020], Gorbunov et al. [2020].

In this article, we characterize the sample-path large deviations and metastability of heavy-tailed stochastic difference equations, thus offering the heavy-tailed counterparts of Freidlin–Wentzell and Eyring–Kramers theorems. Specifically, we (i) establish a version of sample-path large deviations that is uniform w.r.t. the initial values, (ii) obtain the sharp asymptotics of the joint law of (scaled) exit times and exit locations for heavy-tailed dynamical systems, and (iii) study the global dynamics of heavy-tailed SGDs over a multi-well potential. Our findings systematically characterize a fascinating phenomena that, under truncated heavy-tailed noises, SGD can almost always avoid narrow minima of the potential. In particular, such global behavior is known to improve the test performance of deep neural networks trained with SGD (see, e.g., Keskar et al. [2017], Jiang et al. [2020], Wang et al. [2022]). Compared to existing works, our results bring several substantial theoretical advancements. First, our results accommodate general and state-dependent noise structure in heavy-tailed SGDs (see the diffusion coefficient $\sigma(\cdot)$ in (1)). This level of generality greatly surpasses the scope of SGDs driven by iid noises as studied in Wang et al. [2022]. It offers a much more realistic representation of the actual dynamics in deep learning. Second, our results can address the continuous-time stochastic differential equations driven by heavy-tailed perturbations. The formal statements are omitted in this paper due to space limit, but are collected in the preprint Wang and Rhee [2023]. Next, Theorem 1 develops sample-path large deviations for heavy-tailed dynamical systems, which is a fundamental framework for studying how rare events arise in heavy-tailed systems. It is applicable to fairly general events and processes beyond the first exit problem in the machine learning context. Besides, the first exit analysis in Theorem 2 does not require any condition on the second-order derivative of the potential U. Technical proofs are all collected in the preprint Wang and Rhee [2023]. For clarity of the exposition, we focus on results in $\mathbb{R}^1$, but we stress that our approach can be adapted to provide multi-dimensional analogues.

The rest of this article is organized as follows. Section 2 studies sample-path large deviations of heavy-tailed dynamical systems. Section 3 conducts first exit analysis, where we characterize a notion of asymptotic atom, develop a general framework of first exit analysis for Markov processes, and discuss the implication of our results in the context of machine learning.

2 Sample-Path Large Deviations

Let $Z_1, Z_2, \ldots$ be the iid copies of some random variable Z. The goal of this section is to study the sample-path large deviations for $\{X_j^{\eta|b}(x) : j \geq 0\}$ defined through the recursion

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

Here, for any $c \in (0,\infty)$, we set $\varphi_c(\cdot)$ as the projection operator from $\mathbb{R}$ onto $[-c,c]$. We also set $\varphi_\infty(x) = x$ as the identity mapping. Besides, we focus on the case where the heavy-tailedness in Z_i's is captured by the notion of regular variation. For any measurable function $\phi : (0,\infty) \to (0,\infty)$, we say that ϕ is regularly varying as $x \to \infty$ with index β (denoted as $\phi \in \mathcal{RV}_\beta$) if $\lim_{x\to\infty} \phi(tx)/\phi(x) = t^\beta$ for all $t > 0$. For details on the definition and properties of regularly varying functions, see, for example, chapter 2 of Resnick [2007]. Let

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

Now, we impose the following conditions on the perturbation Z and $a(\cdot)$ and $\sigma(\cdot)$, i.e., the drift and diffusion coefficients in (1).

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

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

Meanwhile, there exists some $D \in (0,\infty)$ such that

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

Furthermore, $\sigma(x) > 0\ \forall x \in \mathbb{R}$. Lastly, in case that $b = \infty$, there is some $C \in (0,\infty)$ such that

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

In short, we assume that Z_i's are heavy-tailed noises with index $\alpha > 1$, the drift and diffusion coefficients are Lipschitz continuous, and the diffusion coeffcient $\sigma(\cdot)$ is non-degenerate. In case that the truncation threshold b in (1) is set as ∞ (that is, φ_b is simply the identity mapping and no truncation is involved), we further impose the boundedness condition on $a(\cdot)$ and $\sigma(\cdot)$.

To present the main results, we set a few notations. Let $(\mathbb{D}, \boldsymbol{d}_{J_1})$ be a metric space where $\mathbb{D}$ is the space of all RCLL functions on $[0,1]$, and $\boldsymbol{d}_{J_1}$ is the Skorokhold J_1 metric. Given any $A \subseteq \mathbb{R}$, let $A^{k\uparrow} \triangleq \big\{(t_1,\cdots,t_k) \in A^k :\ t_1 < t_2 < \cdots < t_k\big\}$ be the set of sequences of increasing real numbers with length k on A. For any $k \in \mathbb{N}$ and $b \in (0,\infty]$, define mapping $h^{(k)|b} : \mathbb{R} \times \mathbb{R}^k \times (0,1]^{k\uparrow} \to \mathbb{D}$ as follows. Given any $x_0 \in \mathbb{R}$, $\boldsymbol{w} = (w_1,\cdots,w_k) \in \mathbb{R}^k$, and $\boldsymbol{t} = (t_1,\cdots,t_k) \in (0,1]^{k\uparrow}$, let $\xi = h^{(k)|b}(x_0,\boldsymbol{w},\boldsymbol{t}) \in \mathbb{D}$ be the solution to (under the initial condition $\xi_0 = x_0$)

$$\frac{d\xi_t}{dt} = a(\xi_t) \qquad \forall t \in [0,1],\ t \neq t_1,\cdots,t_k, \tag{2}$$

$$\xi_t = \xi_{t-} + \varphi_b\big(\sigma(\xi_{t-})\cdot w_j\big) \qquad \text{if } t = t_j \text{ for some } j \in [k]. \tag{3}$$

Here, recall that the truncation operator $\varphi_c(w)$ simply projects $w \in \mathbb{R}$ onto $[-c,c]$; for any $\xi \in \mathbb{D}$ and $t \in (0,1]$, we use $\xi_{t-} = \lim_{s\uparrow t}\xi_s$ to denote the left limit of ξ at t. In essence, the mapping $h^{(k)|b}(x_0,\boldsymbol{w},\boldsymbol{t})$ produces the ODE path perturbed by jumps $w_1,\cdots,w_k$ (modulated by the diffusion coefficient $\sigma(\cdot)$ and with sizes truncated under b) at times $t_1,\cdots,t_k$. We also adopt the convention that $\xi = h^{(0)|b}(x_0)$ is the solution to the ODE $\frac{d\xi_s}{ds} = a(\xi_s)\ \forall s \in [0,1]$. For any $A \subseteq \mathbb{R}$, let $\mathbb{D}_A^{(k)|b} \triangleq h^{(k)|b}\big(A \times \mathbb{R}^k \times (0,1]^{k\uparrow}\big)$.

Let $\boldsymbol{X}^{\eta|b}(x) \triangleq \big\{X^{\eta|b}_{\lfloor t/\eta\rfloor}(x) :\ t \in [0,1]\big\}$ with $\lfloor x\rfloor \triangleq \max\{n \in \mathbb{Z} :\ n \le x\}$ and $\lceil x\rceil \triangleq \min\{n \in \mathbb{Z} :\ n \ge x\}$. Note that $\boldsymbol{X}^{\eta}(x)$ is a $\mathbb{D}$-valued random element. Also, recall that $H(\cdot) = \mathbf{P}(|Z_1| > \cdot)$, and let $\lambda(\eta) \triangleq \eta^{-1}H(\eta^{-1})$. Over a metric space $(\mathbb{S},\boldsymbol{d})$, we use A^- and A° to denote the closure and interior of as set $A \subseteq \mathbb{S}$, and we say that set $A \subseteq \mathbb{S}$ is bounded away from another set $B \subseteq \mathbb{S}$ if $\inf_{x\in A, y\in B}\boldsymbol{d}(x,y) > 0$. Throughout this paper, all measurable functions and sets are understood as Borel measurable. We establish the following theorem in Wang and Rhee [2023], where we also present the precise definition of measures $\mathbf{C}^{(k)|b}(\,\cdot\,;\,x)$.

Theorem 1. *Let Assumption 1 hold. There exists a family of (Borel) measures $\big\{\mathbf{C}^{(k)|b}(\,\cdot\,;\,x) :\ k \ge 0,\ x \in \mathbb{R}\big\}$, with each $\mathbf{C}^{(k)|b}(\,\cdot\,;\,x)$ supported on $\mathbb{D}^{(k)|b}_{\{x\}}$, such that the following claims hold. Given any $k \in \mathbb{N}$, any compact $A \subseteq \mathbb{R}$, and any measurable set $B \in \mathbb{D}$ that is bounded away from $\mathbb{D}_A^{(k-1)|b}$,*

$$\begin{aligned}\inf_{x\in A}\mathbf{C}^{(k)|b}\big(B^\circ;x\big) &\le \liminf_{\eta\downarrow 0}\frac{\inf_{x\in A}\mathbf{P}\big(\boldsymbol{X}^{\eta|b}(x)\in B\big)}{\lambda^k(\eta)}\\ &\le \limsup_{\eta\downarrow 0}\frac{\sup_{x\in A}\mathbf{P}\big(\boldsymbol{X}^{\eta|b}(x)\in B\big)}{\lambda^k(\eta)} \le \sup_{x\in A}\mathbf{C}^{(k)|b}\big(B^-;x\big) < \infty.\end{aligned} \tag{4}$$

Furthermore, let $\boldsymbol{X}^{\eta|b}_{|B}(x)$ be a process having the conditional law of $\boldsymbol{X}^{\eta|b}(x)$ given that $\boldsymbol{X}^{\eta|b}(x) \in B$. If B is bounded away from $\mathbb{D}^{(k-1)|b}_{\{x\}}$ with $k = \min\{j \ge 0 :\ \mathbb{D}^{(j)|b}_{\{x\}} \cap B \neq \emptyset\}$, then $\boldsymbol{X}^{\eta|b}_{|B}(x) \Rightarrow \boldsymbol{X}^{*|b}_{|B}(x)$ as $\eta \downarrow 0$, where the law of $\boldsymbol{X}^{*|b}_{|B}(x)$, denoted as $\mathbf{P}_{|B}(\,\cdot\,;x)$, is given by

$$\mathbf{P}_{|B}(\,\cdot\,;x) \triangleq \frac{\mathbf{C}^{(k)|b}(\cdot\cap B;x)}{\mathbf{C}^{(k)|b}(B;x)}.$$

This result presents the heavy-tailed counterpart of the classical Freidlin-Wentzell theorem. It is also a clear manifestation of the catastrophe principle in heavy-tailed dynamical systems, which has

been established at sample-path level in Rhee et al. [2019] but only for Lévy processes and random walks. Specifically, the weak convergence for the conditional law $\boldsymbol{X}^{\eta|b}_{|B}(x)$ indicates the most likely causes of a rare event $\{\boldsymbol{X}^{\eta|b}(x) \in B\}$, that is, through k large perturbations in $\boldsymbol{X}^{\eta|b}(x)$ where k is the smallest number of perturbations (truncated under threshold b) needed for the nominal path (i.e., the ODE path initialized at x under drift coefficient $a(\cdot)$) to fall into set B; the sample-path large deviations result (4) gives us the polynomial rate of decay $\lambda^k(\eta)$ for $\mathbf{P}(\boldsymbol{X}^{\eta|b}(x) \in B)$. In particular, due to the regularly varying nature of $H(\cdot)$, we know that $\lambda^k(\eta)$ is roughly of order $\eta^{k(\alpha-1)}$ for small η, which essentially tells us how rare the events $\{\boldsymbol{X}^{\eta|b}(x) \in B\}$ are.

3 First Exit Analysis

Throughout this section, we fix an open interval $I \triangleq (s_{\text{left}}, s_{\text{right}})$ where $s_{\text{left}} < 0 < s_{\text{right}}$, and impose the following assumption on $a(\cdot)$.

Assumption 2. $a(0) = 0$. *Besides, it holds for all* $x \in I \setminus \{0\}$ *that* $a(x)x < 0$.

Henceforth in this article, we interpret $a(\cdot)$ as the derivative of a potential $U \in \mathcal{C}^1(\mathbb{R})$ with $a(\cdot) = -U'(\cdot)$. Note that Assumption 2 then implies that U has a unique local minimum at $x = 0$ over the domain I. Moreover, since $U'(x)x = -a(x)x > 0$ for all $x \in I \setminus \{0\}$, we know that the domain I is a subset of the attraction field of the origin in the following sense: the limit $\lim_{t\to\infty} \boldsymbol{y}_t(x) = 0$ holds for all $x \in I$ where $\boldsymbol{y}_t(x)$ is the solution of ODE $\frac{d\boldsymbol{y}_t(x)}{dt} = a\big(\boldsymbol{y}_t(x)\big)$ under initial condition $\boldsymbol{y}_0(x) = x$. We add a remark that Assumption 2 is at least as flexible as standard assumptions in related works. For instance, compared to Imkeller, Peter and Pavlyukevich, Ilya [2008], Imkeller and Pavlyukevich [2006], our results accommodate non-constant diffusion coefficient $\sigma(\cdot)$, allow for truncation in the dynamics and remove conditions such as $U \in \mathcal{C}^3(\mathbb{R})$ or non-degeneracy of $U''(\cdot)$ at the boundary of I.

Specifically, our goal is to characterize (after a proper scaling) the asymptotic laws of

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

i.e., the first exit time of $X_j^{\eta|b}(x)$ from I, as $\eta \downarrow 0$. Let

$$r \triangleq |s_{\text{left}}| \wedge s_{\text{right}}, \qquad \mathcal{J}_b^* \triangleq \lceil r/b \rceil \ \forall b \in (0,\infty). \tag{5}$$

Here, r is the distance between the origin and $I^{\complement}$, and $\mathcal{J}_b^*$ is the minimum number of jumps required to exit from I if the size of each jump is bounded by b. As a convention, we set $\mathcal{J}_\infty^* = 1$. Recall that $H(\cdot) = \mathbf{P}(|Z_1| > \cdot)$ and $\lambda(\eta) = \eta^{-1}H(\eta^{-1})$. Let $\mathbb{Z}$ be the collection of all integers. We obtain the following result regarding the first exit times and exit locations of $X_j^{\eta|b}(x)$ from I, which provides the heavy-tailed counterparts of the classical Erying-Kramers theorem. In Wang and Rhee [2023] we present a detailed version of the theorem with the precise definition of measures $\check{\mathbf{C}}^{(k)|b}$.

Theorem 2. *Let Assumptions 1 and 2 hold. There exists a family of (Borel) measures* $\big\{\check{\mathbf{C}}^{(k)|b}(\cdot)\big\}_{k\geq 1}$ *on* $\mathbb{R}$ *such that the following claims hold.*

1. *Suppose that* $b \in (0,\infty)$ *such that* $s_{left}/b \notin \mathbb{Z}$ *and* $s_{right}/b \notin \mathbb{Z}$. *For any* $\epsilon > 0$, $t > 0$, *and measurable set* $B \subseteq I^{\complement}$,

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

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

where $C_b^* \triangleq \check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(I^{\complement})$ *and* $I_\epsilon = (s_{left} + \epsilon, s_{right} - \epsilon)$ *is the* ϵ*-shrinkage of* I.

2. *Suppose that* $b = \infty$. *Then for any* $t > 0$ *and measurable set* $B \subseteq I^{\complement}$,

$$\limsup_{\eta\downarrow 0} \sup_{x\in I_\epsilon} \mathbf{P}\bigg(C^*\eta \cdot \lambda(\eta)\tau^{\eta}(x) > t;\ X^{\eta|\infty}_{\tau^{\eta|\infty}(x)}(x) \in B\bigg) \leq \frac{\check{\mathbf{C}}^{(1)|\infty}(B^-)}{C^*} \cdot \exp(-t),$$

$$\liminf_{\eta\downarrow 0}\inf_{x\in I_\epsilon}\mathbf{P}\left(C^*\eta\cdot\lambda(\eta)\tau^\eta(x)>t;\ X^{\eta|\infty}_{\tau^{\eta|\infty}(x)}(x)\in B\right)\geq\frac{\check{\mathbf{C}}^{(1)|\infty}(B^\circ)}{C^*}\cdot\exp(-t)$$

where $C^*\triangleq\check{\mathbf{C}}^{(1)|\infty}(I^{\complement})$.

We conclude this section with a few remarks. First, instead of focusing on a certain fixed $x\in I$, the asymptotics established above hold uniformly for initial values over any compact set within I. Next, in Theorem 2 we can further relax Assumption 1: that is, if $b=\infty$, then the claims still hold even if we remove the boundedness condition in Assumption 1. Lastly, note that our results characterize an intricate phase transition in the asymptotics of $\tau^{\eta|b}(x)$. To be specific, the first exit time $\tau^{\eta|b}(x)$ is roughly of order $1/\eta^{1+J_b^*\cdot(\alpha-1)}$ for small η. In other words, the order of the first exit time $\tau^{\eta|b}(x)$ does not vary continuously with b; rather, it exhibits a discretized dependency on b through the minimum number of jumps $J_b^*=\lceil|s_{\text{left}}|\vee s_{\text{right}}/b\rceil$ required for the exit. This phase transition phenomenon further exemplifies the catastrophe principle under regularly varying noises, as the quantity J_b^* dictates not only the most likely cause (i.e., through J_b^* large noises) but also the rarity of the exit (i.e., occuring roughly once every $1/\eta^{1+J_b^*\cdot(\alpha-1)}$ steps).

3.1 General Framework

Underlying Theorem 2 is a general framework of first exit analysis for general Markov processes. Consider a general metric space $\mathbb{S}$ and a family of $\mathbb{S}$-valued Markov chains $\big\{\{V_j^\eta(x):j\geq 0\}:\eta>0\big\}$ parameterized by η, where $x\in\mathbb{S}$ denotes the initial state and j denotes the time index. We use the notation $\boldsymbol{V}^\eta_{[0,T]}(x)\triangleq\{V^\eta_{\lfloor t/\eta\rfloor}(x):t\in[0,T]\}$ to denote the scaled sample path. For a given set E, let $\tau^\eta_E(x)\triangleq\min\{j\geq 0:V_j^\eta(x)\in E\}$ be $\{V_j^\eta(s):j\geq 0\}$'s first hitting time of E. We consider an asymptotic domain of attraction $I\subseteq\mathbb{S}$, within which $\boldsymbol{V}^\eta_{[0,T]}(x)$ typically (i.e., as $\eta\downarrow 0$) stays throughout any given time horizon $[0,T]$ as far as $x\in I$. We will make these informal descriptions precise in Condition 1. In many cases, however, $V^\eta_\cdot(x)$ is bound to escape I eventually due to the stochasticity if we do not constrain the time horizon. The goal of this section is to establish an asymptotic limit of the joint distribution of the exit time $\tau^\eta_{I^{\complement}}(x)$ and the exit location $V^\eta_{\tau^\eta_{I^{\complement}}(x)}(x)$. Throughout this section, we will denote $V^\eta_{\tau^\eta_{I(\epsilon)^{\complement}}(x)}(x)$ and $V^\eta_{\tau^\eta_{I^{\complement}}(x)}(x)$ with $V^\eta_{\tau_\epsilon}(x)$ and $V^\eta_{\tau}(x)$, respectively, for notation simplicity.

To facilitate the analysis, we introduce the notion of asymptotic atom, where the process recurrently enters and (almost) regenerates. Let $\{I(\epsilon)\subseteq I:\epsilon>0\}$ and $\{A(\epsilon)\subseteq\mathbb{S}:\epsilon>0\}$ be collections of subsets of I such that $\bigcup_{\epsilon>0}I(\epsilon)=I$ and $\bigcap_{\epsilon>0}A(\epsilon)\neq\emptyset$. Let $C(\cdot)$ is a probability measure on $\mathbb{S}\setminus I$ satisfying $C(\partial I)=0$. In Wang and Rhee [2023], we introduced the following concept.

Definition 1 (Asymptotic Atoms)**.** $\big\{\{V_j^\eta(x):j\geq 0\}:\eta>0\big\}$ *possesses an asymptotic atom* $\{A(\epsilon)\subseteq\mathbb{S}:\epsilon>0\}$ *associated with the domain* I*, location measure* $C(\cdot)$*, scale* $\gamma:(0,\infty)\to(0,\infty)$*, and covering* $\{I(\epsilon)\subseteq I:\epsilon>0\}$ *if the following holds: For each measurable set* $B\subseteq\mathbb{S}$*, there exist* $\delta_B:(0,\infty)\times(0,\infty)\to(0,\infty)$*,* $\epsilon_B>0$*, and* $T_B>0$ *such that for any* $\epsilon\leq\epsilon_B$ *and* $T\geq T_B$*,*

$$\begin{aligned}
C(B^\circ)-\delta_B(\epsilon,T)&\leq\liminf_{\eta\downarrow 0}\frac{\inf_{x\in A(\epsilon)}\mathbf{P}\big(\tau^\eta_{I(\epsilon)^{\complement}}(x)\leq T/\eta;\ V^\eta_{\tau_\epsilon}(x)\in B\big)}{\gamma(\eta)T/\eta}\\
&\leq\limsup_{\eta\downarrow 0}\frac{\sup_{x\in A(\epsilon)}\mathbf{P}\big(\tau^\eta_{I(\epsilon)^{\complement}}(x)\leq T/\eta;\ V^\eta_{\tau_\epsilon}(x)\in B\big)}{\gamma(\eta)T/\eta}\leq C(B^-)+\delta_B(\epsilon,T)\\
&\limsup_{\eta\downarrow 0}\frac{\sup_{x\in I(\epsilon)}\mathbf{P}\big(\tau^\eta_{(I(\epsilon)\setminus A(\epsilon))^{\complement}}(x)>T/\eta\big)}{\gamma(\eta)T/\eta}=0\\
&\liminf_{\eta\downarrow 0}\inf_{x\in I(\epsilon)}\mathbf{P}\big(\tau^\eta_{A(\epsilon)}(x)\leq T/\eta\big)=1
\end{aligned}$$

where $\gamma(\eta)/\eta\to 0$ *as* $\eta\downarrow 0$ *and* δ_B*'s are such that* $\lim_{\epsilon\to 0}\lim_{T\to\infty}\delta_B(\epsilon,T)=0$.

In short, when studying the first exit of $V^\eta_\cdot$ from the domain I, asymptotic atoms identify regions at which the process recurrently visits and (almost) regenerates upon each visit, thus providing tight controls over the behavior of the process afterwards regardless of the exact initial values. In particular,

it leads to a sharp characterization of the first exit times and locations for V^η. The following theorem is the key result of the general framework. See Wang and Rhee [2023] for the detailed statement and the proof of the theorem.

Condition 1. *A family $\big\{\{V_j^\eta(x) : j \geq 0\} : \eta > 0\big\}$ of Markov chains possesses an asymptotic atom $\{A(\epsilon) \subseteq \mathbb{S} : \epsilon > 0\}$ associated with the domain I, location measure $C(\cdot)$, scale $\gamma : (0,\infty) \to (0,\infty)$, and covering $\{I(\epsilon) \subseteq I : \epsilon > 0\}$.*

Theorem 3. *If Condition 1 holds, then the first exit time $\tau^\eta_{I^{\complement}}(x)$ scales as $1/\gamma(\eta)$, and the distribution of the location $V^\eta_\tau(x)$ at the first exit time converges to $C(\cdot)$. Moreover, the convergence is uniform over $I(\epsilon)$ for any $\epsilon > 0$. That is, for each $\epsilon > 0$, measurable $B \subseteq I^{\complement}$, and $t \geq 0$,*

$$\begin{aligned} C(B^\circ)\cdot e^{-t} &\leq \liminf_{\eta\downarrow 0} \inf_{x\in I(\epsilon)} \mathbf{P}\big(\gamma(\eta)\tau^\eta_{I^{\complement}}(x) > t,\, V^\eta_\tau(x) \in B\big) \\ &\leq \limsup_{\eta\downarrow 0} \sup_{x\in I(\epsilon)} \mathbf{P}\big(\gamma(\eta)\tau^\eta_{I^{\complement}}(x) > t,\, V^\eta_\tau(x) \in B\big) \leq C(B^-)\cdot e^{-t}. \end{aligned}$$

In light of Theorem 3, the proof of Theorem 2 boils down to verifying Condition 1. Specifically, we set $V_j^\eta(x) = X_j^{\eta|b}(x)$, use covering $I(\epsilon) = (s_{\text{left}} + \epsilon, s_{\text{right}} - \epsilon)$, and let

$$C(\,\cdot\,) \triangleq \frac{\check{\mathbf{C}}^{(\mathcal{J}_b^*)|b}(\,\cdot\,\setminus I)}{C_b^*}, \qquad \gamma(\eta) \triangleq C_b^* \cdot \eta \cdot \big(\lambda(\eta)\big)^{\mathcal{J}_b^*}.$$

This allows us to verify that $A(\epsilon) = (-\epsilon, \epsilon)$, i.e., the neighborhoods of the local minimum, are asymptotic atoms. Theorem 2 then follows immediately from the results in Theorem 3. See Section 4 of Wang and Rhee [2023] for the detailed proofs.

3.2 Implications in Deep Learning

The success of modern machine learning algorithms, especially deep learning, is often attributed to the algorithms' ability to avoid sharp local minima in the loss landscape, as sharp minima are often associated with poor generalization performance during test phase; see, e.g., Keskar et al. [2017], Jiang et al. [2020]. A direct implication of Theorem 2 is that, under truncated heavy-tailed noises, SGD spends much longer time around the wider minima, especially when η is small. In fact, a much stronger result can be stated about the global dynamics of truncated heavy-tailed SGD. As η approaches 0, the truncated heavy-tailed SGD almost always stays around the widest minima of a multi-well potential. For clarity of the article, we present an informal version of the theorem below. See Section 2.4 of Wang and Rhee [2023] for the formal statement of the theorem.

Theorem 4 (Informal)**.** *Under proper structural conditions, there exists some (deterministic) scale function $\lambda_b^* : (0,\infty) \to (0,\infty)$ such that*

$$\frac{1}{T}\int_0^T \mathbb{I}\Big\{ X^{\eta|b}_{\lfloor t/\lambda_b^*(\eta)\rfloor}(x) \in \textit{widest minima} \Big\} dt \xrightarrow{\mathbf{P}} 1 \qquad \textit{as } \eta \downarrow 0.$$

In light of Theorem 4, it is natural to consider implementing training algorithms for deep learning tasks by capitalizing the elimination of sharp minima under truncated heavy-tailed dynamics. In particular, Theorem 4 implies that truncated heavy-tailed SGDs are almost guaranteed to stay at a flat and wide basin. As a result, there is no need to devise a specific mechanism for evaluating the sharpness of the solution obtained at the end of the training. Indeed, as has been confirmed in numerical experiments, by injecting and then truncating heavy-tailed noises during the training phase, better test performances can be achieved (see Wang et al. [2022]). It is also worth mentioning that the observed improvements in test performance are achievable only when we combine the injection of heavy-tailed noises with the truncation mechanism (i.e., gradient clipping). In fact, it is equally important to emphasize that, without truncation, the presence of heavy-tailed noises leads to notably inferior performance, which is aligned with observations in existing works such as Gorbunov et al. [2020] and Zhang et al. [2020]. In other words, the truncation mechanism is integral to the performance of SGD in practice under heavy-tailed noises.

References

A D Ventsel' and M I Freidlin. On small random perturbations of dynamical systems. *Russian Mathematical Surveys*, 25(1):1, feb 1970. doi: 10.1070/RM1970v025n01ABEH001254. URL `https://dx.doi.org/10.1070/RM1970v025n01ABEH001254`.

Yuri Kifer. A Discrete-Time Version of the Wentzell-Friedlin Theory. *The Annals of Probability*, 18 (4):1676 – 1692, 1990. doi: 10.1214/aop/1176990641. URL https://doi.org/10.1214/aop/1176990641.

C Donati-Martin. Large deviations for wishart processes. *Probability and Mathematical Statistics*, 28, 2008.

Paul Dupuis and Richard S Ellis. *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, 2011.

Amarjit Budhiraja and Paul Dupuis. A variational representation for positive functionals of infinite dimensional brownian motion. *Probability and mathematical statistics-Wroclaw University*, 20(1): 39–61, 2000.

Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2): 107–115, 1935. doi: 10.1063/1.1749604. URL https://doi.org/10.1063/1.1749604.

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

A. D. Wentzell M. I. Freidlin. *Random Perturbations of Dynamical Systems*. Springer New York, NY, 1998. doi: https://doi.org/10.1007/978-1-4612-0611-8.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/simsekli19a.html.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3964–3975. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/gurbuzbalaban21a.html.

Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=yxHPRAqCqn.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgnXpVYwS.

Liam Hodgkinson and Michael W Mahoney. Multiplicative noise and heavy tails in stochastic optimization. *arXiv preprint arXiv:2006.06293*, 2020.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. *arXiv preprint arXiv:2006.04740*, 2020.

P. Imkeller and I. Pavlyukevich. First exit times of sdes driven by stable lévy processes. *Stochastic Processes and their Applications*, 116(4):611–642, 2006. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2005.11.006. URL https://www.sciencedirect.com/science/article/pii/S0304414905001614.

Peter Imkeller, Ilya Pavlyukevich, and Michael 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.

Peter Imkeller, Ilya Pavlyukevich, and Torsten Wetzel. First exit times for Lévy-driven diffusions with exponentially light jumps. *The Annals of Probability*, 37(2):530 – 564, 2009. doi: 10.1214/08-AOP412. URL https://doi.org/10.1214/08-AOP412.

Imkeller, Peter and Pavlyukevich, Ilya. Metastable behaviour of small noise lévy-driven diffusions. *ESAIM: PS*, 12:412–437, 2008. doi: 10.1051/ps:2007051. URL https://doi.org/10.1051/ps:2007051.

Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël RICHARD. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a97da629b098b75c294dffdc3e463904-Paper.pdf.

Xingyu Wang, Sewoong Oh, and Chang-Han Rhee. Eliminating sharp minima from SGD with truncated heavy-tailed noise. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=B3Nde6lvab.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1etN1rtPB.

Eduard Gorbunov, Marina Danilova, and Alexander 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. URL https://proceedings.neurips.cc/paper/2020/file/abd1c782880cc59759f4112fda0b8f98-Paper.pdf.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Xingyu Wang and Chang-Han Rhee. Large deviations and metastability analysis for heavy-tailed dynamical systems, 2023.

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

Chang-Han Rhee, Jose Blanchet, Bert Zwart, et al. Sample path large deviations for lévy processes and random walks with regularly varying increments. *The Annals of Probability*, 47(6):3551–3605, 2019.