Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise

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

Sewoong Oh
Allen School of Computer Science & Engineering
University of Washington, Seattle, WA, 98195
sewoong@cs.washington.edu

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

June 28, 2021

Abstract

The empirical success of deep learning is often attributed to SGD's mysterious ability to avoid sharp local minima in the loss landscape, as sharp minima are known to lead to poor generalization. Recently, empirical evidence of *heavy-tailed* gradient noise was reported in many deep learning tasks, and it was shown in [28, 29] that SGD can *escape* sharp local minima under the presence of such heavy-tailed gradient noise, providing a partial solution to the mystery. In this work, we analyze a popular variant of SGD where gradients are truncated above a fixed threshold. We show that it achieves a stronger notion of avoiding sharp minima: it can effectively *eliminate* sharp local minima entirely from its training trajectory. We characterize the dynamics of truncated SGD driven by heavy-tailed noises. First, we show that the truncation threshold and width of the attraction field dictate the order of the first exit time from the associated local minimum. Moreover, when the objective function satisfies appropriate structural conditions, we prove that as the learning rate decreases, the dynamics of the heavy-tailed truncated SGD closely resemble those of a continuous-time Markov chain that never visits any sharp minima. Real data experiments on deep learning confirm our theoretical prediction that heavy-tailed SGD with gradient clipping finds a *flatter* local minima and achieves better generalization.

1 Introduction

Stochastic gradient descent (SGD) and its variants have seen unprecedented empirical successes in training deep neural networks. The training of deep neural networks is typically posed as a non-convex optimization problem without explicit regularization, but 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 is well known to lead to poor generalization [7, 11, 14]. Despite significant efforts to explain such phenomena theoretically, understanding how SGD avoids sharp local minima 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, [28] and [29] report the empirical evidence of heavy-tails in stochastic gradient noise in popular deep learning architectures (see also [30, 3]) and show that SGD can escape sharp local minima in polynomial time under the presence of the heavy-tailed gradient noise. More specifically, they view heavy-tailed SGDs as discrete approximations of Lévy driven Langevin equations and argue that the amount of time SGD spends in each local minimum is proportional to the width of the associated minimum according to the metastability theory [23, 9, 10] for such simplified 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 a popular 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 [2, 16, 5, 21, 34, 4]. We uncover a rich mathematical structure in the dynamics of SGD under this scheme and prove that the long-run 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, we prove that gradient clipping *completely eliminates sharp minima from the trajectory of SGDs*. Moreover, we rigorously establish that, under small learning rates, the dynamics of clipped heavy-tailed SGD closely resemble a continuous-time Markov chain (CTMC) that never visits sharp local minima in the loss landscape. Our theoretical results provide critical insights into how heavy-tailed dynamics of SGD can be utilized to find a local minimum that generalizes better.

Figure 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 (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 (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 (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 (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 if we stop training SGD at an arbitrary time point, it is almost guaranteed that it won't be at a sharp minimum, effectively eliminating sharp minima from its training trajectories. Behind these phenomena is the different scaling of first exit times from different attraction fields with different widths. Figure 1 (Right) demonstrates that Theorem 1 (the dashed lines) accurately predicts the first exit times.

We also propose a novel training strategy that takes advantage of the newly discovered global dynamics of truncated heavy-tailed SGDs. Despite the evidence of heavy tails reported in numerous statistical learning tasks [29, 28, 3, 6, 8, 19, 15, 30, 34], there seem to be deep learning contexts where the heavy-tails are lacking; see, for instance, [20, 33]. Given the critical role of the heavy-tails in our theory, the natural idea is to inflate the tail distribution by carefully injecting heavy-tailed noises. At first glance, the benefits of introducing heavy-tails may seem unclear, given that we truncate the stochastic gradient in the end. Nevertheless, in light of the global dynamics of SGD revealed in our theory (demonstrated in Figure 1), truncation of heavy-tailed noises produces a strong regularization effect. As detailed in Section 4, we investigate various image classification tasks and deep neural network architectures and compare the performance of different optimization methods with or without the tail inflation and gradient clipping mechanism. Results reported in Tables 2 and 3 illustrate that the tail-inflation strategy we propose here indeed consistently improves the generalization performance of the SGD and drives the SGD iterates to reach local minima with "flatter" geometry. In contrast, the pure form of (unclipped) SGD exhibits drastically deteriorated performance when heavy-tails are

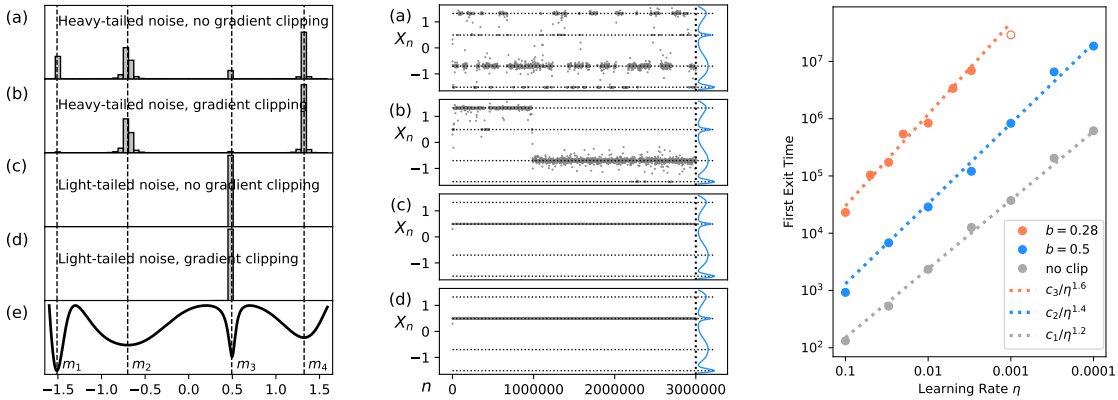

Figure 1: Elimination of Sharp Minima under Truncated Heavy-tailed Noises in $\mathbb{R}^1$. **(Left)** Histograms of the locations visited by SGD. With truncated heavy-tailed noises, SGD hardly ever visits the two sharp minima m_1 and m_3. The objective function f is plotted at the bottom, and dashed 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 1 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 $\eta = 0.001$.

injected. This clearly shows that injecting heavy-tailed noises alone may not be sufficient in real-world deep learning problems, and gradient clipping is indeed a key ingredient. To the best of our knowledge the idea of adding heavy-tailed noises has not been explored successfully in statistical learning context, let alone for the purpose of achieving better generalization performance for the test data. For analyses of injecting light-tailed (Gaussian) noises in deep learning tasks, see [31, 35].

The rest of the paper is organized as follows. Section 2 formulates the problem setting and characterizes the global dynamics of the SGD driven by heavy-tailed noises. Section 3 presents numerical experiments that confirm our theory. Section 4 proposes a new algorithm that artificially injects heavy tailed gradient noise in actual deep learning tasks and demonstrates the improved performance. Section 5 concludes with several potential future directions.

Technical Contributions: 1) We rigorously characterize the global behavior—Eyring-Kramer type formula and Markov chain model reduction for metastability—of the heavy-tailed SGD with gradient clipping. We focus on the case where the loss function is in $\mathbb{R}^1$ with some (standard) simplifying assumptions on its geometry. Even with such assumptions, the proofs our theorems involve substantial technical challenges since the traditional tools for analyzing SGD fail in our context due to the adaptive nature of its dynamics and non-Gaussian distributional assumptions. Moreover, we believe that the technical ideas we developed in this paper lays the foundation for a fully general theory—Freidlin-Wentzell type sample path large deviations and metastability of general heavy-tailed stochastic processes—that can address the high-dimensional loss landscapes. 2) We propose a novel computational strategy for improving the generalization performance of SGD by carefully injecting heavy-tailed noise. We test the proposed algorithm with deep learning tasks and confirm that it improves the generalization performance of SGD. This also suggests that the key phenomenon we characterize in our theory— elimination of sharp local minima—manifests itself in real deep learning

problems.

2 Global Dynamics of Truncated Heavy-tailed SGD

This section characterizes the global dynamics of SGD iterates with gradient clipping when applied to a non-convex objective function f. Specifically, we provide a sharp characterization of the time it takes for SGDs to exit from an attraction field: the first exit time is of order $O(1/\eta^{1+l(\alpha-1)})$ as the learning rate η approaches 0, where $\alpha > 1$ is the heavy-tailed index of noise distribution and the integer l indicates how "wide" the attraction field is when compared to the gradient clipping threshold. This characterization immediately reveals a *first-exit-time strata* on all attraction fields in the optimization landscape, where the sojourn times at wide attraction fields clearly dominate the sojourn times at other locations. Moreover, in terms of the global dynamics of clipped heavy-tailed SGD, we rigorously establish that the time-scaled version of the sample path of clipped heavy-tailed SGD converges in distribution to a continuous-time Markov chain (CTMC) as η apporaches 0. In particular, this CTMC only visits flat minima in the optimization landscape and completely avoids any sharp local minima in narrow attraction fields. As a result, the truncated heavy-tailed noises exhibit strong regularization effect and eliminate all sharp local minima from SGD trajectories.

2.1 Problem setting

We make the following assumptions for the sake of the simplicity of analysis. However, as illustrated in Section 3 and 4, we believe that the gist of the phenomena we analyze—elimination of sharp local minima from the global dynamics of SGD—persists in general contexts where the domain of f is multi-dimensional, and the stationary points are not necessarily strict local optima separated from one another.

Assumption 1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^2$ function. There exist a positive integer n_{min}, a real number $L \in (0, \infty)$, and an ordered sequence of real numbers $m_1, s_1, m_2, s_2, \cdots, s_{n_{min}-1}, m_{n_{min}}$ such that (1) $-L < m_1 < s_1 < m_2 < s_2 < \cdots < s_{n_{min}-1} < m_{n_{min}} < L$; (2) $f'(x) = 0$ iff $x \in \{m_1, s_1, \cdots, s_{n_{min}-1}, m_{n_{min}}\}$; (3) For any $x \in \{m_1, m_2, \cdots, m_{n_{min}}\}$, $f''(x) > 0$; (4) For any $x \in \{s_1, s_2, \cdots, s_{n_{min}-1}\}$, $f''(x) < 0$.*

As illustrated in Figure 2 (Left), the assumption above requires that f has finitely many local minima (to be specific, the count is $n_{\min}$) and they all lie in the interval $[-L, L]$. Moreover, the points $s_1, \cdots, s_{n_{\min}-1}$ naturally partition the entire real line into different regions $\Omega_i = (s_{i-1}, s_i)$ (here we adopt the convention that $s_0 = -\infty, s_{n_{\min}} = +\infty$). We call each region Ω_i the **attraction field** of the local minimum m_i, as the gradient flow in Ω_i always points to m_i.

Throughout the optimization procedure, given any location $x \in \mathbb{R}$ we assume that we have access to the noisy estimator $f'(x) - Z_n$ of the true gradient $f'(x)$, and $f'(x)$ itself is difficult to evaluate. Specifically, in this work we are interested in the case where the iid sequence of noises $(Z_n)_{n\geq 1}$ are heavy-tailed. Typically, the heavy-tailed phenomena are captured by the concept of regular variation: for a measurable function $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$, we say that ϕ is regularly varying at $+\infty$ 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 [25]. In this paper, we work with the following distributional assumption on the gradient noise. Let

$$H_+(x) \triangleq \mathbb{P}(Z_1 > x), \quad H_-(x) \triangleq \mathbb{P}(Z_1 < -x), \quad H(x) \triangleq H_+(x) + H_-(x) = \mathbb{P}(|Z_1| > x).$$

Assumption 2. $\mathbb{E}Z_1 = 0$. *Furthermore, there exists some $\alpha \in (1, \infty)$ such that function $H(x)$ is regularly varying (at $+\infty$) with index $-\alpha$. Besides, regarding the positive and negative tail for distribution of the noises, we have*

$$\lim_{x\to\infty} \frac{H_+(x)}{H(x)} = p_+, \ \lim_{x\to\infty} \frac{H_-(x)}{H(x)} = p_- = 1 - p_+$$

where p_+ and p_- are constants in interval $(0,1)$.

Roughly speaking, Assumption 2 means that the shape of the tail for the distribution of noises Z_n resembles a polynomial function $x^{-\alpha}$, which is much heavier than the exponential tail of Gaussian distributions. Therefore, large values of Z_n are much more likely to be observed under the Assumption 2 compared to the typical Gaussian assumption. Note that the index α of regular variation encodes the heaviness of the tail—the smaller the heavier—and we are assuming that the left and right tails share the same index α. We make this simplifying assumption for the purpose of clear presentation, but it is straightforward to extend this to the case where the left and right tails have different regular variation indices.

Our work concerns a popular variant of SGD where the stochastic gradient is truncated. Specifically, when updating the SGD iterates with a learning rate $\eta > 0$, rather than using the original noisy gradient descent step $\eta(f'(X_n) - Z_n)$, we will truncate it at a threshold $b > 0$ and use $\varphi_b\big(\eta(f'(X_n) - Z_n)\big)$ instead. Here the truncation operator $\varphi_\cdot(\cdot)$ is defined as

$$\varphi_c(w) \triangleq \varphi(w,c) \triangleq (w \wedge c) \vee (-c) \quad \forall w \in \mathbb{R}, c > 0 \tag{1}$$

where $u \wedge v = \min\{u,v\}, u \vee v = \max\{u,v\}$. Besides truncating the stochastic gradient, we also project the SGD into $[-L, L]$ at each iteration; recall that L is the constant in Assumption 1. That is, the main object of our study is the stochastic process $\{X_n^\eta\}_{n\geq 1}$ driven by the following recursion:

$$X_n^\eta \triangleq \varphi_L\Big(X_n^\eta - \varphi_b\big(\eta(f'(X_n^\eta) - Z_n)\big)\Big). \tag{2}$$

The projection φ_L and truncation φ_b here are common practices in many statistical learning contexts as well as other optimization tasks for the purpose of ensuring that the SGD does not explode and drift to infinity. Besides, the projection also allows us to drop the sophisticated assumptions on the tail behaviors of f that are commonly seen in previous works; see, for instance, the dissipativity conditions in [19].

For technical reasons, we make the following assumption about the truncation threshold $b > 0$. Note that this assumption is a very mild one, as it is obviously satisfied by (Lebesgue) almost every $b > 0$.

Assumption 3. *For each $i = 1, 2, \cdots, n_{min}$, $\min\{|s_i - m_i|, |s_{i-1} - m_i|\}/b$ is* not *an integer.*

2.2 First exit times

Denote the SGD's first exit time from the attraction field Ω_i with

$$\sigma_i(\eta) \triangleq \min\{n \geq 0 : X_n^\eta \notin \Omega_i\}.$$

In this section, we prove that $\sigma_i(\eta)$ converges to an exponential distribution when scaled properly. To characterize such a scaling, we first introduce a few concepts. For each attraction field Ω_i, define (note that $\lceil x \rceil = \min\{n \in \mathbb{Z} : n \geq x\}$, $\lfloor x \rfloor = \max\{n \in \mathbb{Z} : n \leq x\}$)

$$r_i \triangleq \min\{|m_i - s_{i-1}|, |s_i - m_i|\}, \quad l_i^* \triangleq \lceil r_i/b \rceil. \tag{3}$$

Note that l_i^*'s in fact depend on the the value of gradient clipping threshold b even though this dependency is not highlighted by the notation. Here r_i can be interpreted as the radius or the effective *width* of the attraction field, and l_i^* is the *minimum number of jumps* required to escape Ω_i when starting from m_i. Indeed, the gradient clipping threshold b dictates that no single SGD update step can travel more than b, and to exit Ω_i when starting from m_i (which requires the length of travel to be at least r_i) we can see that at least $\lceil r_i/b \rceil$ steps are required. We can interpret l_i^* as the minimum effort required to exit Ω_i. Note that, (for a fixed b) the minimum number of jumps l_i^*

is an indicator of the width of the attraction field Ω_i. Theorem 1 states that l_i^* dictates the order of magnitude of the first exit time as well as where the iterates X_n^η land on at the first exit time.

To stress the initial condition, we write $\mathbb{P}_x$ for the probability law when conditioning on $X_0^\eta = x$, or simply write $X_n^\eta(x)$. For each Ω_i, define a scaling function

$$\lambda_i(\eta) \triangleq H(1/\eta)\Big(\frac{H(1/\eta)}{\eta}\Big)^{l_i^*-1}.$$

The distribution of the first exit time under this scaling is characterized in Theorem 1.

Theorem 1. *Under Assumptions 1-3, there exist constants $q_i > 0$ $\forall i \in \{1, 2, \cdots, n_{min}\}$ and $q_{i,j} \geq 0$ $\forall j \in \{1, 2, \cdots, n_{min}\} \setminus \{i\}$ such that*

(i) Suppose that $x \in \Omega_k$ for some $k \in \{1, 2, \cdots, n_{min}\}$. Under $\mathbb{P}_x$, $q_k \lambda_k(\eta)\sigma_k(\eta)$ converges in distribution to an exponential random variable with rate 1 as $\eta \to 0$.

(ii) For $k, l \in \{1, 2, \cdots, n_{min}\}$ such that $k \neq l$, we have $\lim_{\eta\to 0} \mathbb{P}_x(X^\eta_{\sigma_k(\eta)} \in \Omega_l) = q_{k,l}/q_k$.

See Section 2.5 for the proof of Theorem 1. We note here that Theorem 1 implies (i) for X_n^η to escape the current attraction field, say Ω_i, it takes $O\big(1/\lambda_i(\eta)\big)$ time, and (ii) the destination is most likely to be reachable within l_i^* jumps from m_i. Before concluding this section, we state the definition of the constants $q_i, q_{i,j}$ in Theorem 1 to show how these constants vary with the loss landscape f and the shape of gradient noise distributions (in particular α, p_+, p_-). Let $\mathbf{Leb}_+$ denote the Lebesgue measure restricted on $[0, \infty)$, and define a (Borel) measure ν_α on $\mathbb{R}\backslash\{0\}$ as

$$\nu_\alpha(dx) = \mathbb{1}\{x > 0\}\frac{\alpha p_+}{x^{\alpha+1}} + \mathbb{1}\{x < 0\}\frac{\alpha p_-}{|x|^{\alpha+1}} \tag{4}$$

where α, p_-, and p_+ are constants in Assumption 2. Define a Borel measure μ_i on $\mathbb{R}^{l_i^*} \times \Big(\mathbb{R}_+\Big)^{l_i^*-1}$ as the product measure $\mu_i = (\nu_\alpha)^{l_i^*} \times (\mathbf{Leb}_+)^{l_i^*-1}$. We also define mappings h_i as follows. For a real sequence $\mathbf{w} = (w_1, w_2, \cdots, w_{l_i^*})$ and a positive real number sequence $\mathbf{t} = (t'_j)_{j=2}^{l_i^*}$, define $t_1 = t'_1 = 0$ and $t_j = t'_1 + t'_2 + \cdots + t'_j$ for $j = 2, \cdots, l_i^*$. Now we define a path $\hat{\mathbf{x}} : [0, \infty) \mapsto \mathbb{R}$ as the solution to the following ODE with jumps:

$$\hat{\mathbf{x}}(0) = \varphi_L\big(m_i + \varphi_b(w_1)\big); \tag{5}$$

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = -f'(\hat{\mathbf{x}}(t)), \quad \forall t \in [t_{j-1}, t_j), \quad \forall j = 2, \cdots, l_i^*; \tag{6}$$

$$\hat{\mathbf{x}}(t_j) = \varphi_L\big(\hat{\mathbf{x}}(t_j-) + \varphi_b(w_j)\big), \quad \forall j = 2, \cdots, l_i^*. \tag{7}$$

Now we define the mappings $h_i : \mathbb{R}^{l_i^*} \times \Big(\mathbb{R}_+\Big)^{l_i^*-1} \mapsto \mathbb{R}$ as $h_i(\mathbf{w}, \mathbf{t}) = \hat{\mathbf{x}}(t_{l_i^*})$. It is easy to see that h_i's are continuous mappings. With these mappings, we define the following sets:

$$E_i \triangleq \{(\mathbf{w}, \mathbf{t}) \subseteq \mathbb{R}^{l_i^*} \times \mathbb{R}_+^{l_i^*-1} : h_i(\mathbf{w}, \mathbf{t}) \notin \Omega_i\}; \tag{8}$$

$$E_{i,j} \triangleq \{(\mathbf{w}, \mathbf{t}) \subseteq \mathbb{R}^{l_i^*} \times \mathbb{R}_+^{l_i^*-1} : h_i(\mathbf{w}, \mathbf{t}) \in \Omega_j\}. \tag{9}$$

Now we can define the constant q_i and $q_{i,j}$ as follows:

$$q_i = \mu_i(E_i), \quad q_{i,j} = \mu_i(E_{i,j}) \quad \forall i \neq j. \tag{10}$$

We add a few concluding remarks regarding the intuition behind Theorem 1.

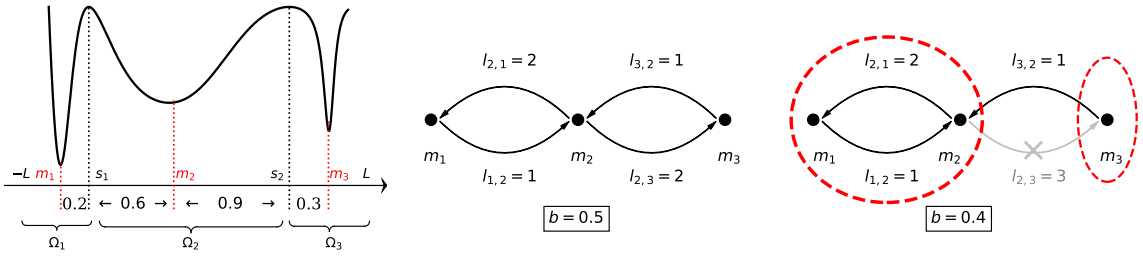

Figure 2: Typical transition graphs $\mathcal{G}$ under different gradient clipping thresholds b. (Left) The function f illustrated here has 3 attraction fields. For the second one $\Omega_2 = (s_1, s_2)$, we have $s_2 - m_2 = 0.9, m_2 - s_1 = 0.6$. (Middle) The typical transition graph induced by $b = 0.5$. The entire graph $\mathcal{G}$ is irreducible since all nodes communicate with each other. (Right) The typical transition graph induced by $b = 0.4$. When $b = 0.4$, since $0.6 < 2b$ and $0.9 > 2b$, the SGD can only exit Ω_2 from the left with only 2 jumps if started from m_2. Therefore, on the graph $\mathcal{G}$ there are two communication classes: $G_1 = \{m_1, m_2\}, G_2 = \{m_3\}$; G_1 is absorbing while G_2 is transient.

- Suppose that X_n^η is started at the i^{th} local minimum m_i of f, and consider the behavior of X_n^η over the time period $H_1 \triangleq \{1, \dots, \lceil t/\eta \rceil\}$ for a sufficiently large t. The heavy-tailed large deviations theory [26] and a heuristic application of the contraction principle implies that the path of $X_{\lceil n/\eta \rceil}^\eta$ over this period will converge to the gradient flow of f, and the event that $X_{\lceil n/\eta \rceil}^\eta$ escapes Ω_i within this period is a (heavy-tailed) rare event. This means that the probability of such an event is of order $(1/\eta)^{(\alpha-1)l_i^*}$. Moreover, whenever it happens, it is almost always because X_n^η is shaken by *exactly* l_i^* *large gradient noises of size* $\mathcal{O}(1/\eta)$, which translates to l_i^* jumps in $X_{\lceil n/\eta \rceil}^\eta$'s path, while the rest of its path closely resemble the deterministic gradient flow. Moreover, conditional on the event that X_n^η fails to escape from Ω_i within this period, the endpoint of the path is most likely to be close to the local minima, i.e., $X_{\lceil t/\eta \rceil}^\eta \approx m_i$. This suggests that over the next time period $H_2 \triangleq \{\lceil t/\eta \rceil + 1, \lceil t/\eta \rceil + 2, \dots, 2\lceil t/\eta \rceil\}$ of length $\lceil t/\eta \rceil$, X_n^η will behave similarly to its behavior over the first period H_1. The same argument applies to the subsequent periods $H_3, H_4, \dots$ as well. Therefore, over each time period of length $\lceil t/\eta \rceil$, there is $(1/\eta)^{(\alpha-1)l_i^*}$ probability of exit. In view of this, the exit time should be of order $(1/\eta)^{1+(\alpha-1)l_i^*}$ and resemble an exponential distribution when scaled properly.
- Part (i) of Theorem 1 builds on this intuition and rigorously prove that the first exit time is indeed roughly of order $1/\lambda_i(\eta) \approx (1/\eta)^{1+(\alpha-1)l_i^*}$ and resembles an exponential distribution.
- Given this, one would expect that $X_{\sigma_i(\eta)}^\eta$, the location of SGD right at the time of exit, will hardly ever be farther than $l_i^* b$ away from m_i: the length of each update is clipped by b, and there will most likely be only l_i^* large SGD steps during this successful attempt. Indeed, from the definition of $q_{i,j}$'s, one can see that $q_{i,j} > 0$ if and only if $\inf_{y \in \Omega_j} |y - m_i| < l_i^* b$.

Summarizing the three bullet points here, we see that the minimum number of jumps l_i^* dictates *how* heavy-tailed SGD escapes an attraction field, *where* the SGD lands on upon its exit, and *when* the exit occurs.

2.3 Elimination of Small Attraction Fields

In this section, we show that, under proper structural assumptions on the geometry of f, the "sharp minima" of f can be effectively eliminated from the trajectory of heavy-tailed SGD. Before we state the result, we first introduce a few new concepts. Similar to the the minimum number of jumps l_i^*

defined in 3, we can define the following quantities as the *minimum number of jumps to reach* Ω_j *from* m_i for any $j \neq i$:

$$l_{i,j} = \begin{cases} \lceil (s_{j-1} - m_i)/b \rceil & \text{if } \ j > i, \\ \lceil (m_i - s_j)/b \rceil & \text{if } \ j < i. \end{cases} \tag{11}$$

Recall that Theorem 1 dictates that X_n^η is most likely to move out of the current attraction field, say Ω_i, to somewhere else after $O\big(1/\lambda_i(\eta)\big)$ time steps, and the destination is most likely to be reachable within l_i^* jumps from m_i. Therefore, the transitions from Ω_i to Ω_j can be considered typical if Ω_j can be reached from m_i with l_i^* jumps—that is, $l_{i,j} = l_i^*$. Now we define the following directed graph that only includes these typical transitions.

Definition 1 (Typical Transition Graph)**.** *Given a function f satisfying Assumption 1 and gradient clipping threshold $b > 0$ satisfying Assumption 3, a directed graph $\mathcal{G} = (V, E)$ is the corresponding* typical transition graph *if (1) $V = \{m_1, \cdots, m_{n_{min}}\}$; (2) An edge $(m_i \to m_j)$ is in E iff $l_{i,j} = l_i^*$.*

Naturally, the typical transition graph $\mathcal{G}$ can be decomposed into different communication classes $G_1, \cdots, G_K$ that are mutually exclusive by considering the equivalence relation associated with the existence of the (two-way) paths between i and j. More specifically, for $i \neq j$, we say that i and j communicate if and only if there exists a path $(m_i, m_{k_1}, \cdots, m_{k_n}, m_j)$ as well as a path $(m_j, m_{k'_1}, \cdots, m_{k'_{n'}}, m_i)$ in $\mathcal{G}$; in other words, by travelling through edges on $\mathcal{G}$, m_i can be reached from m_j and m_j can be reached from m_i.

We say that a communication class G is *absorbing* if there does not exist any edge $(m_i \to m_j) \in E$ such that $m_i \in G$ and $m_j \notin G$. Otherwise, we say that G is *transient*. In the case that all m_i's communicate with each other on graph $\mathcal{G}$, we say $\mathcal{G}$ is *irreducible*. See Figure 2 (Middle) for the illustration of an irreducible case. When $\mathcal{G}$ is irreducible, we define the set of *largest* attraction fields $M^{\text{large}} \triangleq \{m_i : \ i = 1, 2, \cdots, n_{\min}, \ l_i^* = l^{\text{large}}\}$ where $l^{\text{large}} = \max_j l_j^*$; recall that l_i^* characterizes the width of Ω_i. Also, we define the longest time scale $\lambda^{\text{large}}(\eta) = H(1/\eta)\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^{\text{large}}-1}$. Note that this corresponds exactly to the order of the first exit time of the largest attraction fields; see Theorem 1. The following theorem is the main result of this paper.

Theorem 2. *Let Assumptions 1-3 hold and assume that the graph $\mathcal{G}$ is* ***irreducible****. Given $t > 0$, $\beta > 1 + (\alpha - 1)l^{large}$, and $x \in [-L, L]$,*

$$\frac{1}{\lfloor t/\eta^\beta \rfloor} \int_0^{\lfloor t/\eta^\beta \rfloor} \mathbb{1}\Big\{X_{\lfloor u \rfloor}^\eta(x) \in \bigcup_{j: m_j \notin M^{large}} \Omega_j\Big\} du \to 0 \tag{12}$$

in probability as $\eta \to 0$.

We refer the readers to Section 2.6 for the proof. Here we briefly discuss the implication of the result. Suppose that we terminate the training after a reasonably long time, say, $\lfloor t/\eta^\beta \rfloor$ iterations. Then the random variable that converges to zero in (12) is exactly the proportion of time that X_n^η spent in the attraction fields that are not wide. Therefore, by truncating the gradient noise of the heavy-tailed SGD, we can effectively eliminate small attraction fields from its training trajectory. In other words, it is almost guaranteed that SGD is in one of the widest attraction fields after sufficiently long training iterations.

Interestingly enough, Theorem 2 is merely a manifestation of the global dynamics of heavy-tailed SGD, and a lot more can be said even when $\mathcal{G}$ is not irreducible. The main message can be summarized as follows: (a) SGD with truncated and heavy-tailed noise naturally partitions the entire training landscape into different regions; (b) In each region, the dynamics of X_n^η for small η closely resemble that of a continuous-time Markov chain that *only* visits local minima; (3) In particular, any sharp minima within each region is almost completely avoided by SGD.

When the typical transition graph (see Definition 1 in the main paper) is not irreducible, there will be multiple communication classes on the graph. Suppose that there are K communication classes $G_1, \cdots, G_K$. From now on, we zoom in on a specific communication class $G \in \{G_1, \cdots, G_K\}$. For this communication class G, define $l_G^* \triangleq \max\{l_i^* : \ i = 1, 2, \cdots, n_{\min};\ m_i \in G\}$. For each local minimum $m_i \in G$, we call its attraction field Ω_i a *large* attraction field if $l_i^* = l_G^*$, and a *small* attraction field if $l_i^* < l_G^*$. We have thus classified all m_i in G into two groups: the ones in large attraction fields $m_1^{\text{large}}, \cdots, m_{i_G}^{\text{large}}$ and the ones in small attraction fields $m_1^{\text{small}}, \cdots, m_{i'_G}^{\text{small}}$. Also, define a scaling function λ_G associated with G as $\lambda_G(\eta) \triangleq H(1/\eta)\Big(\frac{H(1/\eta)}{\eta}\Big)^{l_G^*-1}$.

Theorem 3. *Under Assumptions 1-3, if G is **absorbing**, then there exists a continuous-time Markov chain Y on $\{m_1^{large}, \cdots, m_{i_G}^{large}\}$ such that for any $x \in \Omega_i, |x| \le L$ (where $i \in \{1, 2, \cdots, n_{min}\}$) with $m_i \in G$, and*

$$X^\eta_{\lfloor t/\lambda_G(\eta)\rfloor}(x) \to Y_t(\pi_G(m_i)) \quad \text{as } \eta \downarrow 0$$

in the follwing sense: for any positive integer k and any $0 < t_1 < \cdots < t_k$, the random vector $\Big(X^\eta_{\lfloor \frac{t_k}{\lambda_G(\eta)}\rfloor}(x), \cdots, X^\eta_{\lfloor \frac{t_k}{\lambda_G(\eta)}\rfloor}(x)\Big)$ converges in distribution to $\Big(Y_{t_1}(\pi_G(m_i)), \cdots, Y_{t_k}(\pi_G(m_i))\Big)$ as $\eta \downarrow 0$.Here π_G is a random mapping satisfying (1) $\pi_G(m) \equiv m$ if $m \in \{m_1^{large}, \cdots, m_{i_G}^{large}\}$; (2) $\pi_G(m)$ is a random variable that only takes value in $\{m_1^{large}, \cdots, m_{i_G}^{large}\}$ if $m \in \{m_1^{small}, \cdots, m_{i'_G}^{small}\}$.

The proof is provided in Section 2.6 and we add a few remarks here. Theorem 3 tells us that, when initialized on a absorbing class G, the dynamics of the clipped heavy-tailed SGD converge to a continuous-time Markov chain avoiding any local minima that is not in the largest attraction fields in G. Second, under small learning rate $\eta > 0$, if $X_n^\eta(x)$ is initialized at $x \in \Omega_i$ where Ω_i is NOT a largest attraction field in G, then SGD will quickly escape Ω_i and arrive at some Ω_j that is indeed a largest one—i.e., $m_j \in M^{\text{large}}$; such a transition is so quick that, under time scaling $\lambda^{\text{large}}(\eta)$, it is almost instantaneous as if $X_n^\eta(x)$ is actually initialized randomly at some of the largest attraction fields in G. This randomness is compressed in the random mapping π_G.

Next, to state the corresponding result for the *transient* case, we introduce a couple of extra definitions. We consider a version of X_n^η that is killed when X_n^η leaves G. Define stopping time

$$\tau_G(\eta) \triangleq \min\{n \ge 0 : X_n^\eta \notin \bigcup_{i: m_i \in G} \Omega_i\} \tag{13}$$

as the first time the SGD iterates leave all attraction fields in G, and we use a cemetery state $\dagger$ to construct the following process $X_n^{\dagger,\eta}$ as a version of X_n^η with killing at τ_G:

$$X_n^{\dagger,\eta} = \begin{cases} X_n^\eta & \text{if } \ n < \tau_G(\eta), \\ \dagger & \text{if } \ n \ge \tau_G(\eta). \end{cases} \tag{14}$$

Theorem 4. *Under Assumptions 1-3, if G is **transient**, then there exists a continuous-time Markov chain Y **with killing** that has state space $\{m_1^{large}, \cdots, m_{i_G}^{large}, \dagger\}$ (we say the Markov chain Y is killed when it enters the absorbing cemetery state $\dagger$) such that for any $x \in \Omega_i, |x| \le L$ (where $i \in \{1, 2, \cdots, n_{min}\}$) with $m_i \in G$, and $X^{\dagger,\eta}_{\lfloor t/\lambda_G(\eta)\rfloor}(x) \to Y_t(\pi_G(m_i))$ as $\eta \downarrow 0$ in the sense of finite-dimensional distributions, where π_G is a random mapping satisfying (1) $\pi_G(m) \equiv m$ if $m \in \{m_1^{large}, \cdots, m_{i_G}^{large}\}$; (2) $\pi_G(m)$ is a random variable that only takes value in $\{m_1^{large}, \cdots, m_{i_G}^{large}, \dagger\}$ if $m \in \{m_1^{small}, \cdots, m_{i'_G}^{small}\}$.*

See Section 2.6 for the proof. Theorem 3 and 4 tell us that, regardless of the initial location of SGD iterates, the elimination of small attraction fields can be observed on each communication class in the graph $\mathcal{G}$. In the case that $\mathcal{G}$ is indeed irreducible, the following result follows immediately from Theorem 3 and guarantees the elimination of small attraction fields of the entire loss landscape.

Theorem 5. *Let Assumptions 1-2 hold. Let $x \in \Omega_i \cap [-L, L]$ for some $i = 1, 2, \cdots, n_{min}$. If Assumptions 1-3 hold and $\mathcal{G}$ is irreducible, then there exist a continuous-time Markov chain Y on M^{large} as well as a random mapping π such that the scaled process $\{X^{\eta}_{\lfloor t/\lambda^{large}(\eta)\rfloor}(x) : t \geq 0\}$ converges to process $\{Y_t(\pi(m_i)) : t \geq 0\}$ in the sense of finite-dimensional distributions.*

2.4 Implications of the theoretical results

Systematic control of the exit times from attraction fields: In light of the wide minima folklore, one may want to find techniques to modify the sojourn time of SGD at each attraction field. Theorem 1 suggests that the order of the first exit time (w.r.t. learning rate η) is directly controlled by the gradient clipping threshold b. Recall that for an attraction field with minimum jump number l^*, Theorem 1 tells us the exit time from this attraction field is roughly of order $(1/\eta)^{1+(\alpha-1)l^*}$. Given the width of the attraction field, its minimum jump number l^* is dictated by gradient clipping threshold b. Therefore, gradient clipping provides us with a very systematic method to control the exit time of each attraction field. For instance, given clipping threshold b, the exit time from an attraction field with width less than b is of order $(1/\eta)^{\alpha}$, while the exit time from one larger than b is at least $(1/\eta)^{2\alpha-1}$, which dominates the exit time from smaller ones.

The role of structural properties of $\mathcal{G}$ and f: Recall that in order for Theorem 5 to apply, the irreducibility of $\mathcal{G}$ is required. Along with the choice of b, the geometry of function f is a deciding factor of the irreducibility. For instance, we say that $\mathcal{G}$ is *symmetric* if for any attraction field Ω_i such that $i = 2, 3, \cdots, n_{\min} - 1$ (so that Ω_i is not the leftmost or rightmost one at the boundary), we have $q_{i,i-1} > 0, q_{i,i+1} > 0$. One can see that $\mathcal{G}$ is symmetric if and only if, for any $i = 2, 3, \cdots, n_{\min} - 1$, $|s_i - m_i| \vee |m_i - s_{i-1}| < l_i^* b$, and symmetry is a sufficient condition for the irreducibility of $\mathcal{G}$. The graph illustrated in Figure 2(Middle) is symmetric, while the one in Figure 2(Right) is not. As the name suggests, in the $\mathbb{R}^1$ case the symmetry of $\mathcal{G}$ is more likely to hold if the shape of attraction fields in f is also nearly symmetric around its local minimum. If not, the symmetry (as well as irreducibility) of $\mathcal{G}$ can be violated as illustrated in Figure 2, especially when a small gradient clipping threshold b is used.

Generally speaking, our results imply that, even with the truncated heavy-tailed noises, the function f needs to satisfy certain regularity conditions to ensure that SGD iterates avoid undesirable minima. This is consistent with the observations in [14]: the deep neural nets that are more trainable with SGD tend to have a much more regular structure in terms of the number and shape of local minima.

Heavy-tailed SGD without gradient clipping: It is worth mentioning that our results also characterize the dynamics of heavy-tailed SGDs without gradient clipping. For instance, recall that we restrict the iterates on the compact set $[-L, L]$. If we set truncation threshold $b > 2L$, then the gradient clipping at norm b becomes superfluous given the weight clipping at $\pm L$ and the update recursion degenerates to

$$X_n^{\eta,\text{unclipped}} = \varphi_L\Big(X_{n-1}^{\eta,\text{unclipped}} - \eta f'(X_{n-1}^{\eta,\text{unclipped}}) + \eta Z_n\Big). \tag{15}$$

The next result follows immediately from Theorem 1 and 3.

Corollary 6. *There exist constants $q_i > 0\ \forall i$, $q_{i,j} > 0\ \forall j \neq i$ such that the following claims hold for any i and any $x \in \Omega_i, |x| \leq L$.*

1) *Under $\mathbb{P}_x$, $q_i H(1/\eta)\sigma_i(\eta)$ converges weakly to an Exponential random variable with rate 1 as $\eta \downarrow 0$;*
2) *For any $j = 1, 2, \cdots, n_{min}$ such that $j \neq i$, $\lim_{\eta\downarrow 0} \mathbb{P}_x(X^{\eta}_{\sigma_i(\eta)} \in \Omega_j) = q_{i,j}/q_i$.*
3) *Let Y be a continuous-time Markov chain on $\{m_1, \cdots, m_{n_{min}}\}$ with generator matrix Q parametrized by $Q_{i,i} = -q_i, Q_{i,j} = q_{i,j}$. Then the scaled process $\{X^{\eta,unclipped}_{\lfloor t/H(1/\eta)\rfloor}(x) : t > 0\}$ converges to $\{Y_t(m_i) : t > 0\}$ in the sense of finite-dimensional distributions.*

At first glance, Corollary 6 may seem similar to the results in [28] and [23]. However, the object studied in [28, 23] is different: they study the following Langevin-type stochastic differential equation (SDE) driven by an α-stable Lévy process L_t: $dY_t^\eta = -f'(Y_{t-}^\eta)dt + \eta dL_t$. In particular, [23] studies the metastability of Y_t^η and concludes that as the scaling factor $\eta \downarrow 0$, the first exit time and global dynamics of Y_t^η admit a similar characterization as described in our Theorem 6. Then Theorem 4 in [28] argues that when the learning rate η is sufficiently small, the distribution of the first exit time of the SGD X_n^η and that of the Lévy-driven Langevin SDE Y_t^η are similar. However, the analysis of [28] hinges critically on the assumption that L_t^α is symmetric and α-stable. While such an assumption is convenient for their analysis, it is a strong assumption. It implies that the gradient noise distribution belongs to a very specific parametric family and excludes all the other heavy-tailed distributions. In particular, the assumption precludes analysis of any heavy tails with finite variance. On the contrary, our work directly analyzes the SGD X_n^η and reveals the heavy-tailed SGD dynamics at a much greater level of generality. Specifically, we allow the noise to have general regularly varying distributions with arbitrary tail index—which includes α-stable distributions as a (very) special case—and extend the characterization of global dynamics of heavy-tailed SGD to the adaptive versions of SGD where gradient clipping is applied.

2.5 Proof of Theorem 1

This section is organized as follows. We first introduce some key lemmas that analyze the dynamics of truncated heavy-tailed SGD before the first exit from an attraction field. Building upon these lemmas, we then present the proof of Theorem 1. We detail the proof of the technical lemmas in Appendix B.

As stated above, the proof of Theorem 1 hinges on the following two lemmas that characterize the behavior of X_n^η in two different phases respectively. Let $k \in [n_{\min}]$ and $x \in \Omega_k$. We consider the SGD iterates initialized at $X_0^\eta = x$. In the first phase, the SGD iterates return to $[m_k - 2\epsilon, m_k + 2\epsilon]$, a small neighborhood of the local minimizer in attraction field Ω_k; in other words, it ends at

$$T_{\text{return}}^{(k)}(\eta, \epsilon) \triangleq \min\{n \geq 0 : \ X_n^\eta \in [m_k - 2\epsilon, m_k + 2\epsilon]\}. \tag{16}$$

During this phase, we show that for all learning rate η that is sufficiently small, it is almost always the case that X_n^η would quickly return to $[m_k - 2\epsilon, m_k + 2\epsilon]$, and it never leaves Ω_k before $T_{\text{return}}^{(k)}$.

Lemma 7. *Under Assumptions 1-3, there exists some $c \in (0, \infty)$ such that for any $k \in [n_{min}]$, the following claim holds for all $\epsilon > 0$ small enough:*

$$\lim_{\eta \downarrow 0} \inf_{y \in [-L,L]:\ y \in (s_{k-1}+\epsilon, s_k - \epsilon)} \mathbb{P}_y\Big(X_n^\eta \in \Omega_k \ \forall n \in \big[T_{return}^{(k)}(\eta,\epsilon)\big],\ T_{return}^{(k)}(\eta,\epsilon) \leq \frac{c\log(1/\epsilon)}{\eta}\Big) = 1.$$

During the second phase, X_n^η starts from somewhere in $[m_k - 2\epsilon, m_k + 2\epsilon]$ and tries to escape from Ω_k, meaning that the phase ends at $\sigma_k(\eta)$. During this phase, we show that the distributions of the first exit time $\sigma_k(\eta)$ and the location $X_{\sigma_k(\eta)}^\eta$ do converge to the ones described in Theorem 1 as learning rate η tends to 0.

Lemma 8. *Let Assumptions 1-3 hold. Given $C > 0, u > 0$ and $k, l \in [n_{min}]$ with $k \neq l$, the following claims*

$$\limsup_{\eta \downarrow 0} \sup_{x \in [-L,L],\ x \in (m_k - 2\epsilon, m_k + 2\epsilon)} \mathbb{P}_x\Big(q_k \lambda_k(\eta)\sigma_k(\eta) > u\Big) \leq C + \exp\big(-(1-C)u\big) \tag{17}$$

$$\liminf_{\eta \downarrow 0} \inf_{x \in [-L,L],\ x \in (m_k - 2\epsilon, m_k + 2\epsilon)} \mathbb{P}_x\Big(q_k \lambda_k(\eta)\sigma_k(\eta) > u\Big) \geq -C + \exp\big(-(1+C)u\big) \tag{18}$$

$$\limsup_{\eta \downarrow 0} \sup_{x \in [-L,L],\ x \in (m_k - 2\epsilon, m_k + 2\epsilon)} \mathbb{P}_x\Big(X_{\sigma_k(\eta)}^\eta \in \Omega_l\Big) \leq \frac{q_{k,l} + C}{q_k} \tag{19}$$

$$\liminf_{\eta \downarrow 0} \inf_{x \in [-L,L],\ x \in (m_k - 2\epsilon, m_k + 2\epsilon)} \mathbb{P}_x\Big(X_{\sigma_k(\eta)}^\eta \in \Omega_l\Big) \geq \frac{q_{k,l} - C}{q_k} \tag{20}$$

hold for all $\epsilon > 0$ that are sufficiently small.

Now we are ready to show Theorem 1.

Proof of Theorem 1. Fix some $k \in [n_{\min}]$ and $x \in \Omega_k \cap [-L, L]$. Let q_k and $q_{k,l}$ be the constants in Lemma 8.

We first prove the weak convergence claim in Theorem 1(i). Arbitrarily choose some $u > 0$ and $C \in (0, 1)$. It suffices to show that

$$\limsup_{\eta \downarrow 0} \mathbb{P}_x(q_k \lambda_k(\eta)\sigma_k(\eta) > u) \le 2C + \exp\big(-(1-C)u\big),$$
$$\liminf_{\eta \downarrow 0} \mathbb{P}_x(q_k \lambda_k(\eta)\sigma_k(\eta) > u) \ge (1-C)\Big(-C + \exp\big(-(1+C)u\big)\Big).$$

Recall the definition of the stopping time $T^{(k)}_{\text{return}}$ in (16). Define event

$$A_k(\eta, \epsilon) \triangleq \Big\{ X^\eta_n \in \Omega_k \ \forall n \in \big[T^{(k)}_{\text{return}}(\eta, \epsilon)\big],\ T^{(k)}_{\text{return}}(\eta, \epsilon) \le \frac{c \log(1/\epsilon)}{\eta} \Big\}$$

where $c < \infty$ is the constant in Lemma 7. First, since $x \in \Omega_k = (s_{k-1}, s_k)$, it holds for all $\epsilon > 0$ small enough that $x \in (s_{k-1} + \epsilon, s_k - \epsilon)$. Next, one can find some $\epsilon > 0$ such that

- (Due to Lemma 7)
$$\mathbb{P}_x\big((A_k(\eta, \epsilon))^c\big) \le C \quad \forall \eta \text{ sufficiently small;}$$
- (Due to (17)(18) and strong Markov property) For all η sufficiently small,
$$\mathbb{P}_x\Big(q_k \lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta, \epsilon)\big) > (1-C)u \ \Big|\ A_k(\eta, \epsilon)\Big) \le C + \exp\big(-(1-C)u\big),$$
$$\mathbb{P}_x\Big(q_k \lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta, \epsilon)\big) > u \ \Big|\ A_k(\eta, \epsilon)\Big) \ge -C + \exp\big(-(1+C)u\big).$$

Fix such ϵ. Lastly, for this fixed ϵ, due to $\lambda_k \in \mathcal{RV}_{-1-l^*_k(\alpha-1)}(\eta)$ and $\alpha > 1$, we have $q_k \lambda_k(\eta) \cdot \frac{c \log(1/\epsilon)}{\eta} < Cu$ for all η sufficiently small. In summary, for all η sufficiently small, we have

$$\begin{aligned}
&\mathbb{P}_x(q_k \lambda_k(\eta)\sigma_k(\eta) > u) \\
\le&\mathbb{P}_x\big((A_k(\eta, \epsilon))^c\big) + \mathbb{P}_x\Big(\big\{q_k \lambda_k(\eta)\sigma_k(\eta) > u\big\} \cap A_k(\eta, \epsilon)\Big) \\
\le&C + \mathbb{P}_x\Big(\big\{q_k \lambda_k(\eta)\sigma_k(\eta) > u\big\} \cap A_k(\eta, \epsilon)\Big) \\
=&C + \mathbb{P}_x\Big(q_k \lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta, \epsilon)\big) > u - q_k \lambda_k(\eta) T^{(k)}_{\text{return}}(\eta, \epsilon) \ \Big|\ A_k(\eta, \epsilon)\Big) \cdot \mathbb{P}_x(A_k(\eta, \epsilon)) \\
\le&C + \mathbb{P}_x\Big(q_k \lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta, \epsilon)\big) > (1-C)u \ \Big|\ A_k(\eta, \epsilon)\Big) \\
\le&2C + \exp\big(-(1-C)u\big)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_x(q_k \lambda_k(\eta)\sigma_k(\eta) > u) \\
\ge&\mathbb{P}_x\Big(\big\{q_k \lambda_k(\eta)\sigma_k(\eta) > u\big\} \cap A_k(\eta, \epsilon)\Big) \\
=&\mathbb{P}_x\Big(q_k \lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta, \epsilon)\big) > u - q_k \lambda_k(\eta) T^{(k)}_{\text{return}}(\eta, \epsilon) \ \Big|\ A_k(\eta, \epsilon)\Big) \cdot \mathbb{P}_x(A_k(\eta, \epsilon)) \\
\ge&\mathbb{P}_x\Big(q_k \lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta, \epsilon)\big) > u - q_k \lambda_k(\eta) T^{(k)}_{\text{return}}(\eta, \epsilon) \ \Big|\ A_k(\eta, \epsilon)\Big) \cdot (1-C)
\end{aligned}$$

$$\geq \mathbb{P}_x\Big(q_k\lambda_k(\eta)\big(\sigma(\eta) - T^{(k)}_{\text{return}}(\eta,\epsilon)\big) > u \;\Big|\; A_k(\eta,\epsilon)\Big)\cdot(1-C)$$
$$\geq (1-C)\Big(-C + \exp\big(-(1+C)u\big)\Big)$$

so this concludes the proof for Theorem 1(i).

In order to prove claims in Theorem 1(ii), we first observe that on event $A_k(\eta,\epsilon)$ we must have $\sigma(\eta) > T^{(k)}_{\text{return}}(\eta,\epsilon)$. Next, arbitrarily choose some $C \in (0,1)$, and note that it suffices to show that $\mathbb{P}_x(X^\eta_{\sigma(\eta)} \in \Omega_l) \in \Big((1-C)\frac{q_{k,l}-C}{q_k}, C + \frac{q_{k,l}+C}{q_k}\Big)$ holds for all η sufficiently small. Again, we can find $\epsilon > 0$ such that

- (Due to Lemma 7)
$$\mathbb{P}_x\big((A_k(\eta,\epsilon))^c\big) \leq C \quad \forall \eta \text{ sufficiently small;}$$
- (Due to (19)(20) and strong Markov property) For all η sufficiently small,
$$\frac{q_{k,l}-C}{q_k} \leq \mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l \;\Big|\; A_k(\eta,\epsilon)\Big) \leq \frac{q_{k,l}+C}{q_k}.$$

In summary, for all η sufficiently small, we have

$$\begin{aligned}
\mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big) &\leq \mathbb{P}_x\big((A_k(\eta,\epsilon))^c\big) + \mathbb{P}_x\Big(\big\{X^\eta_{\sigma_k(\eta)} \in \Omega_l\big\} \cap A_k(\eta,\epsilon)\Big)\\
&\leq C + \mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l \;\Big|\; A_k(\eta,\epsilon)\Big)\mathbb{P}_x(A_k(\eta,\epsilon))\\
&\leq C + \frac{q_{k,l}+C}{q_k},\\
\mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big) &\geq \mathbb{P}_x\Big(\big\{X^\eta_{\sigma_k(\eta)} \in \Omega_l\big\} \cap A_k(\eta,\epsilon)\Big)\\
&= \mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l \;\Big|\; A_k(\eta,\epsilon)\Big)\mathbb{P}_x(A_k(\eta,\epsilon))\\
&\geq (1-C)\frac{q_{k,l}-C}{q_k}
\end{aligned}$$

and this concludes the proof. □

2.6 Proofs of Section 2.3

In this section, we first present some key lemmas and then use them to prove the Theorem 2-5. The proofs of the technical lemmas are deferred to Appendix C.

In order to prove Theorem 2, we will make use of the following lemma, where we show that the type of claim in Theorem 2 is indeed valid if we look at a much shorter time interval. Then when we move onto the proof of Theorem 2, it suffices to partition the entire horizon into pieces of these short time intervals, on each of which we analyze the dynamics of SGD respectively.

Lemma 9. *Let Assumptions 1-3 hold. Assume the graph $\mathcal{G}$ is* irreducible*, and let $\epsilon > 0, \delta > 0$ be any positive real numbers. For the following random variables (indexed by η)*

$$V^{small}(\eta,\epsilon,t) \triangleq \frac{1}{\lfloor t/\lambda^{large}(\eta)\rfloor}\int_0^{\lfloor t/\lambda^{large}(\eta)\rfloor} \mathbb{1}\Big\{X^\eta_{\lfloor u\rfloor} \in \bigcup_{j:m_j\notin M^{large}} \Omega_j\Big\}du, \tag{21}$$

the following claim holds for any sufficiently small $t > 0$:

$$\limsup_{\eta\downarrow 0}\sup_{x\in[-L,L]} \mathbb{P}_x\Big(V^{small}(\eta,\epsilon,t) > \epsilon\Big) \leq 5\delta$$

Proof of Theorem 2. It suffices to show that for any $\epsilon > 0, \delta \in (0, \epsilon)$, we have

$$\limsup_{\eta \downarrow 0} \mathbb{P}_x\Big(V^*(\eta, t, \kappa) > 3\epsilon\Big) < \delta.$$

Let us fix some $\epsilon > 0, \delta \in (0, \epsilon)$. First, let

$$N(\eta) = \lceil \frac{\lfloor t/\eta^\kappa \rfloor}{\lfloor t/\lambda^{\text{large}}(\eta) \rfloor} \rceil.$$

The regularly varying nature of H implies that $\lambda^{\text{large}} \in \mathcal{RV}_{-(1+l^{\text{large}}(\alpha-1))}$. Since $\kappa > 1 + l^{\text{large}}(\alpha - 1)$, we know that $\lim_{\eta \downarrow 0} N(\eta) = \infty$. Next, due to Lemma 9, we can find $t_0 > 0$ and $\bar{\eta} > 0$ such that for any $\eta \in (0, \bar{\eta})$

$$\sup_{y \in [-L, L]} \mathbb{P}_y(V^{\text{small}}(\eta, \epsilon, t_0) > \epsilon) < \delta. \tag{22}$$

For any $k \geq 1$, define

$$V_k(\eta) \triangleq \frac{1}{\lfloor t_0/\lambda^{\text{large}}(\eta) \rfloor} \int_{(k-1)\lfloor t_0/\lambda^{\text{large}}(\eta) \rfloor}^{k\lfloor t_0/\lambda^{\text{large}}(\eta) \rfloor} \mathbb{1}\Big\{X^\eta_{\lfloor u \rfloor} \in \bigcup_{j: m_j \notin M^{\text{large}}} \Omega_j\Big\} du. \tag{23}$$

It is clear from its definition that V_k stands for the proportion of time that the SGD iterates are outside of *large* attraction fields on the interval $[(k-1)\lfloor \frac{t_0}{\lambda^{\text{large}}(\eta)} \rfloor, k\lfloor \frac{t_0}{\lambda^{\text{large}}(\eta)} \rfloor]$. From (22) and Markov property, one can see that for any $\eta \in (0, \bar{\eta})$

$$\sup_{x \in [-L, L]} \mathbb{P}_x(V_k(\eta) > \epsilon \mid X^\eta_0, \cdots, X^\eta_{(k-1)\lfloor t_0/\lambda^{\text{large}}(\eta) \rfloor}) \leq \delta$$

uniformly for all $k \geq 1$. Now define $K(\eta) \triangleq \#\{n = 1, 2, \cdots, N(\eta) : \ V_k(\eta) > \epsilon\}$. By a simple stochastic dominance argument, we have

$$\sup_{x \in [-L, L]} \mathbb{P}_x(K(\eta) \geq j) \leq \mathbb{P}(\text{Binomial}(N(\eta), \delta) \geq j) \ \ \forall j = 1, 2, \cdots.$$

Meanwhile, strong law of large numbers implies the existence of some $\bar{\eta}_1 > 0$ such that $\mathbb{P}(\frac{\text{Binomial}(N(\eta), \delta)}{N(\eta)} > 2\delta) < \delta$ for all $\eta \in (0, \bar{\eta}_1)$, thus

$$\sup_{|x| \leq L} \mathbb{P}_x(K(\eta)/N(\eta) > 2\delta) \leq \delta \ \ \forall \eta \in (0, \bar{\eta}_1 \wedge \bar{\eta}).$$

Lastly, from the definition of $K(\eta)$ and $N(\eta)$, we know that for all the $N(\eta)$ intervals $[(k-1)\lfloor \frac{t_0}{\lambda^{\text{large}}(\eta)} \rfloor, k\lfloor \frac{t_0}{\lambda^{\text{large}}(\eta)} \rfloor]$ with $k \in [N(\eta)]$, only on $K(\eta)$ of them did the SGD iterates spent more then ϵ proportion of time outside of the *large* attraction fields, hence

$$V^*(\eta, t, \kappa) \leq \epsilon + \frac{K(\eta)}{N(\eta)}.$$

In summary, we now have

$$\mathbb{P}_x(V^*(\eta, t, \kappa) > 3\epsilon) < \delta$$

for all $\eta \in (0, \bar{\eta}_1 \wedge \bar{\eta})$. This concludes the proof. □

To show Theorem 3 and 4, we introduce the following concepts. First, we consider the case where the SGD iterates X_n^η is initialized on the communication class G and G is absorbing. For some $\Delta > 0, \eta > 0$, define (let $B(u,v) \triangleq [u-v, u+v]$)

$$\sigma_0^G(\eta,\Delta) \triangleq \min\{n \geq 0 : X_n^\eta \in \bigcup_{i:\ m_i \in G} B(m_i, 2\Delta)\} \tag{24}$$

$$\tau_0^G(\eta,\Delta) \triangleq \min\{n \geq \sigma_0^G(\eta,\Delta) :\ X_n^\eta \in \bigcup_{i: m_i \notin G^{\text{small}}} B(m_i, 2\Delta)\} \tag{25}$$

$$I_0^G(\eta,\Delta) = j \iff X^\eta_{\tau_0^G} \in B(m_j, 2\Delta), \quad \widetilde{I}_0^G(\eta,\Delta) = j \iff X^\eta_{\sigma_0^G} \in B(m_j, 2\Delta) \tag{26}$$

$$\sigma_k^G(\eta,\Delta) \triangleq \min\{n > \tau_{k-1}^G(\eta,\Delta) :\ X_n^\eta \in \bigcup_{i: m_i \in G,\ i \neq I_{k-1}^G} B(m_i, 2\Delta)\}\ \forall k \geq 1 \tag{27}$$

$$\tau_k^G(\eta,\Delta) \triangleq \min\{n \geq \sigma_{k-1}^G(\eta,\Delta) :\ X_n^\eta \in \bigcup_{i: m_i \notin G^{\text{small}}} B(m_i, 2\Delta)\}\ \forall k \geq 1 \tag{28}$$

$$I_k^G(\eta,\Delta) = j \iff X^\eta_{\tau_k^G} \in B(m_j, 2\Delta), \quad \widetilde{I}_k^G(\eta,\Delta) = j \iff X^\eta_{\sigma_k^G} \in B(m_j, 2\Delta) \forall k \geq 1. \tag{29}$$

Intuitively speaking, at each τ_k^G the SGD iterates visits a minimizer that is not in a *small* attraction field on G, and we use I_k^G to mark the label of that large attraction field. Stopping time σ_k^G is the first time that SGD visits a minimizer that is *different* from the one visited at τ_k^G, and τ_{k+1}^G is the first time that a minimizer not in a small attraction field of G is visited again since σ_k^G (and including σ_k^G). It is worth mentioning that, under this definition, we could have $I_k^G = I_{k+1}^G$ for any $k \geq 0$. Meanwhile, define the following process that only keeps track of the updates on the labels $(I_k^G)_{k\geq 0}$ instead of the information of the entire trajectory of $(X_n^\eta)_{n\geq 0}$:

$$\hat{X}_n^{\eta,\Delta} = \begin{cases} m_{I_k^G} & \text{if } \exists k \geq 0 \text{ such that } \tau_k^G \leq n < \tau_{k+1}^G \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

In other words, when $n < \tau_0^G$ we simply let $\hat{X}_n^{\eta,\Delta} = 0$, otherwise it is equal to the latest "marker" for the last visited wide minimum up until step n. This marker process $\hat{X}$ jumps between the different minimizers of the large attractions in G. In particular, if for some n we have $X_n^\eta \in B(m_j, 2\Delta)$ for some j with $m_j \in G^{\text{large}}$, then we must have $\hat{X}_n^{\eta,\Delta} = m_j$, which implies that, in this case, $\hat{X}_n^{\eta,\Delta}$ indeed indicates the location of X_n^η.

Note that results in Theorem 3 and 4 concern a *scaled* version of X^η. Here we also define the corresponding *scaled* version of the processes

$$X_t^{*,\eta} \triangleq X^\eta_{\lfloor t/\lambda_G(\eta) \rfloor} \tag{31}$$

$$\hat{X}_t^{*,\eta,\Delta} \triangleq \hat{X}^{\eta,\Delta}_{\lfloor t/\lambda_G(\eta) \rfloor}, \tag{32}$$

a mapping $\mathbf{T}^*(n,\eta) \triangleq n\lambda_G(\eta)$ that translates a step n to the corresponding timestamp for the scaled processes, and the following series of scaled stopping times

$$\tau_k^*(\eta,\Delta) = \mathbf{T}^*\big(\tau_k^G(\eta,\Delta),\eta\big),\ \sigma_k^*(\eta,\Delta) = \mathbf{T}^*\big(\sigma_k^G(\eta,\Delta),\eta\big). \tag{33}$$

Before presenting the proof of Theorem 3 and 4, we make several preparations. First, our proof is inspired by ideas in [22] and here we provide a briefing. At any time $t > 0$, if we can show that $X_t^{*,\eta}$ is almost always in set $\bigcup_{i:m_i \in G^{\text{large}}} B(m_i, 2\Delta)$ (so the SGD iterates is almost always close to a minimizer in a large attraction field), then the marker process $\hat{X}_t^{*,\eta,\Delta}$ is a pretty accurate indicator of the location of $X_t^{*,\eta}$, so it suffices to show that the marker process $\hat{X}_t^{*,\eta,\Delta}$ converges to a continuous-time Markov chain Y.

Second, we construct the limiting process Y and the random mapping π_G before utilizing them in Theorem 3 and 4. As an important building block for this purpose, we start by considering the following discrete time Markov chain (DTMC) on the entire graph $\mathcal{G} = (V, E)$. Let $\mathbf{P}^{DTMC}$ be a transition matrix with $\mathbf{P}^{DTMC}(m_i, m_j) = \mu_i(E_{i,j})/\mu_i(E_i)$ for all $j \neq i$, and $Y^{DTMC} = (Y_j^{DTMC})_{j\geq 0}$ be the DTMC induced by the said transition matrix. Let

$$T_G^{DTMC} \triangleq \min\{j \geq 0 :\ Y_j^{DTMC} \notin G^{\text{small}}\} \tag{34}$$

be the first time this DTMC visits a large attraction field on the communication class G, or escapes from G. Lastly, define (for any j such that $m_j \notin G^{\text{small}}$)

$$p_{i,j} \triangleq \mathbb{P}\big(Y^{DTMC}_{T_G^{DTMC}}(m_i) = m_j\big) \tag{35}$$

as the probability that the first large attraction field on G visited by Y^{DTMC} is m_j when initialized at m_i.

We add a comment regarding the stopping times T_G^{DTMC} and probabilities $p_{i,j}$ defined above. In the case that G is absorbing, we have $Y_j^{DTMC}(m_i) \in G$ for all $j \geq 0$ if $m_i \in G$. Therefore, in this case, given any i with $m_i \in G$, we must have

$$T_G^{DTMC} = \min\{j \geq 0 :\ Y_j^{DTMC}(m_i) \in G^{\text{large}}\}, \qquad \sum_{j:\ m_j \in G^{\text{large}}} p_{i,j} = 1.$$

On the contrary, when G is transient we may have $\sum_{j:\ m_j\in G^{\text{large}}} p_{i,j} < 1$ and $\sum_{j:\ m_j\notin G} p_{i,j} > 0$. Lastly, whether G is absorbing or transient, we always have $p_{i,j} = \mathbb{1}\{i = j\}$ if $m_i \in G^{\text{large}}$.

Next, consider the following definition of (continuous-time) jump processes.

Definition 2. *A continuous-time process Y_t on $\mathbb{R}$ is a $\Big((U_j)_{j\geq 0}, (V_j)_{j\geq 0}\Big)$ **jump process** if*

$$Y_t = \begin{cases} 0 & \text{if } t < U_0 \\ \sum_{j\geq 0} V_j \mathbb{1}_{[U_0+U_1+\cdots+U_j,\ U_0+U_1+\cdots+U_{j+1})}(t) & \text{otherwise} \end{cases},$$

where $(U_j)_{j\geq 0}$ is a sequence of non-negative random variables such that $U_j > 0\ \forall j \geq 1$ almost surely, and $(V_j)_{j\geq 0}$ is a sequence of random variables in $\mathbb{R}$.

Obviously, the definition above implies that $Y_t = V_j$ for any $t \in [U_j, U_{j+1})$.

Now we are ready to construct the limiting continuous-time Markov chain Y. To begin with, we address the case where G is absorbing. For any $m' \in G^{\text{large}}$, let $Y(m')$ be a $\big((S_k)_{k\geq 0}, (W_k)_{k\geq 0}\big)$-jump process where $S_0 = 0, W_0 = m'$ and (for all $k \geq 0$ and i, j with $m_i \in G^{\text{large}}, m_j \notin G^{\text{small}}$)

$$\mathbb{P}\Big(W_{k+1} = m_j,\ S_{k+1} > t \ \Big|\ W_k = m_i,\ (W_l)_{l=0}^{k-1},\ (S_l)_{l=0}^{k}\Big) \tag{36}$$

$$= \mathbb{P}\Big(W_{k+1} = m_j,\ S_{k+1} > t \ \Big|\ W_k = m_i\Big) = \exp(-q_i t)\frac{q_{i,j}}{q_i}\ \forall t > 0 \tag{37}$$

where

$$q_i = \mu_i(E_i) \tag{38}$$

$$q_{i,j} = \mathbb{1}\{i \neq j\}\mu_i(E_{i,j}) + \sum_{k:\ m_k\in G^{\text{small}}} \mu_i(E_{i,k})p_{k,j} \tag{39}$$

and $p_{k,j}$ is defined in (35). In other words, conditioning on $W_k = m_i$, the time until next jump S_{k+1} and the jump location W_{k+1} are independent, where S_{k+1} is $\text{Exp}(q_i)$ and $W_{k+1} = m_j$ with probability $q_{i,j}/q_i$. First, it is easy to see that Y is a continuous-time Markov chain. Second, under this definition

Y is allowed to have some *dummy* jumps where $W_k = W_{k+1}$: in this case the process Y_t does not move to a different minimizer after the $k+1$-th jump, and by inspecting the path of Y we cannot tell that this dummy jump has occurred. As a result, that generator Q of this Markov chain admits the form (for all $i \neq j$ with $m_i, m_j \in G^{\text{large}}$)

$$Q_{i,i} = -\sum_{k:k\neq i,\ m_k \in G^{\text{large}}} q_{i,k},\ Q_{i,j} = q_{i,j}.$$

Moreover, define the following random function $\pi_G(\cdot)$ such that for any $m_i \in G$,

$$\pi_G(m_i) = \begin{cases} m_j & \text{with probability } q_{i,j}/q_i \text{ if } m_i \in G^{\text{small}} \\ m_i & \text{if } m_i \in G^{\text{large}} \end{cases} \tag{40}$$

By $Y(\pi_G(m_i))$ we refer to the version of the Markov chain Y where we randomly initialize $W_0 = \pi_G(m_i)$. The following lemma is the key tool for proving Theorem 3.

Lemma 10. *Let Assumptions 1-3 hold. Assume that the communication class G is absorbing. Given any $m_i \in G$, $x \in \Omega_i$, finitely many real numbers $(t_l)_{l=1}^{k'}$ such that $0 < t_1 < t_2 < \cdots < t_{k'}$, and a sequence of strictly positive real numbers $(\eta_n)_{n\geq 1}$ with $\lim_{n\to 0} \eta_n = 0$, there exists a sequence of strictly positive real numbers $(\Delta_n)_{n\geq 1}$ with $\lim_n \Delta_n = 0$ such that*

- *As n tends to ∞,*

$$\big(\hat{X}^{*,\eta_n,\Delta_n}_{t_1}(x), \cdots, \hat{X}^{*,\eta_n,\Delta_n}_{t_{k'}}(x)\big) \Rightarrow \big(Y_{t_1}(\pi_G(m_i)), \cdots, Y_{t_{k'}}(\pi_G(m_i))\big) \tag{41}$$

- *For all $k \in [k']$,*

$$\lim_{n\to\infty} \mathbb{P}_x\Big(X^{*,\eta_n}_{t_k} \notin \bigcup_{j:\ m_j \in G^{large}} B(m_j, \Delta_n)\Big) = 0. \tag{42}$$

Now we address the case where G is transient, let $\dagger$ be a real number such that $\dagger \notin [-L, L]$, and we use $\dagger$ as the cemetery state since the processes X^{η}_n or $X^{*,\eta}_t$ are restricted on $[-L, L]$. Recall the definition of τ_G defined in (13). Analogous to the process $X^{\dagger}$ in (14), we can also define

$$X^{\dagger,*,\eta}_t = \begin{cases} X^{*,\eta}_t & \text{if } \ t < \mathbf{T}^*(\tau_G(\eta), \eta) \\ \dagger & \text{otherwise} \end{cases},\quad \hat{X}^{\dagger,*,\eta,\Delta}_t = \begin{cases} \hat{X}^{*,\eta,\Delta}_t & \text{if } \ t < \mathbf{T}^*(\tau_G(\eta), \eta) \\ \dagger & \text{otherwise,} \end{cases}. \tag{43}$$

Next, analogous to τ_G, consider the stopping time

$$\tau^Y_G \triangleq \min\{t > 0 : Y_t \notin G\}.$$

When G is transient, due to the construction of Y we know that $\tau^Y_G < \infty$ almost surely. The introduction of τ^Y_G allows us to define

$$Y^{\dagger}_t = \begin{cases} Y_t & \text{if } \ t < \tau^Y_G \\ \dagger & \text{otherwise.} \end{cases} \tag{44}$$

The following Lemma will be used to prove Theorem 4.

Lemma 11. *Let Assumptions 1-3 hold. Assume that the communication class G is transient. Given any $m_i \in G$, $x \in \Omega_i$, finitely many real numbers $(t_l)_{l=1}^{k'}$ such that $0 < t_1 < t_2 < \cdots < t_{k'}$, and a sequence of strictly positive real numbers $(\eta_n)_{n\geq 1}$ with $\lim_{n\to 0} \eta_n = 0$, there exists a sequence of strictly positive real numbers $(\Delta_n)_{n\geq 1}$ with $\lim_n \Delta_n = 0$ such that*

- *As n tends to ∞,*

$$\big(\hat{X}^{\dagger,*,\eta_n,\Delta_n}_{t_1}(x),\cdots,\hat{X}^{\dagger,*,\eta_n,\Delta_n}_{t_{k'}}(x)\big)\Rightarrow\big(Y^{\dagger}_{t_1}(\pi_G(m_i)),\cdots,Y^{\dagger}_{t_{k'}}(\pi_G(m_i))\big) \tag{45}$$

- *For all $k\in[k']$,*

$$\lim_{n\to\infty}\mathbb{P}_x\Big(X^{\dagger,*,\eta_n}_{t_k}\notin\bigcup_{j:\ m_j\in G^{large}}B(m_j,\Delta_n)\ \textit{and}\ X^{\dagger,*,\eta_n}_{t_k}\neq\dagger\Big)=0. \tag{46}$$

Proof of Theorem 3 and 4. We first address the case where G is absorbing. Arbitrarily choose some $\Delta>0$, a sequence of strictly positive real numbers $(\eta_n)_{n\geq1}$ with $\lim_n\eta_n=0$, a positive integer k', a series of real numbers $(t_j)_{j=1}^{k'}$ with $0<t_1<\cdots<t_{k'}$, and a sequence $(w_j)_{j=1}^{k'}$ with $w_j\in G^{\text{large}}$ for all $j\in[k']$. It suffices to show that

$$\lim_n\mathbb{P}_x\Big(X^{*,\eta_n}_{t_k}\in B(w_k,\Delta)\ \forall k\in[k']\Big)=\mathbb{P}\Big(Y_{t_k}(\pi_G(m_i))=w_k\ \forall k\in[k']\Big).$$

Using Lemma 10, we can find a sequence of strictly positive real numbers $(\Delta_n)_{n\geq1}$ with $\lim_n\Delta_n=0$ such that (41) and (42) hold. From the weak convergence in (41), we only need to show

$$\lim_n\mathbb{P}_x\Big(X^{*,\eta_n}_{t_k}\notin B(\hat{X}^{*,\eta_n,\Delta_n}_{t_k},\Delta)\Big)=0\ \forall k\in[k'].$$

For all n large enough, we have $2\Delta_n<\Delta$. For such large n, observe that

$$\begin{aligned}
&\mathbb{P}_x\Big(X^{*,\eta_n}_{t_k}\notin B(\hat{X}^{*,\eta_n,\Delta_n}_{t_k},\Delta)\Big)\\
\leq&\mathbb{P}_x\Big(X^{*,\eta_n}_{t_k}\notin\bigcup_{j:m_j\in G^{\text{large}}}B(m_j,2\Delta_n)\Big)\ \text{due to definition of marker process }\hat{X}\text{ in (24)-(32)},\\
\leq&\mathbb{P}_x\Big(X^{*,\eta_n}_{t_k}\notin\bigcup_{j:m_j\in G^{\text{large}}}B(m_j,\Delta_n)\Big)
\end{aligned}$$

and by applying (42) we conclude the proof for Theorem 3.

The proof of Theorem 4 is almost identical, with the only modification being that we apply Lemma 11 instead of Lemma 10. In doing so, we are able to find a sequence of $(\Delta_n)_{n\geq1}$ with $\lim_n\Delta_n=0$ such that (45) and (46) hold. Given the weak convergence claim in (45), it suffices to show that

$$\lim_n\mathbb{P}_x\Big(X^{\dagger*,\eta_n}_{t_k}\notin B(\hat{X}^{\dagger,*,\eta_n,\Delta_n}_{t_k},\Delta)\Big)=0\ \forall k\in[k'].$$

If $X^{\dagger,*,\eta_n}_{t_k}=\dagger$, we must have $\hat{X}^{\dagger,*,\eta_n,\Delta_n}_{t_k}=\dagger$ as well. Therefore, for all n large enough so that $3\Delta_n<\Delta$,

$$\begin{aligned}
&\mathbb{P}_x\Big(X^{\dagger*,\eta_n}_{t_k}\notin B(\hat{X}^{\dagger,*,\eta_n,\Delta_n}_{t_k},\Delta)\Big)\\
=&\mathbb{P}_x\Big(X^{\dagger*,\eta_n}_{t_k}\notin B(\hat{X}^{\dagger,*,\eta_n,\Delta_n}_{t_k},\Delta),\ X^{\dagger*,\eta_n}_{t_k}\neq\dagger\Big)\\
\leq&\mathbb{P}_x\Big(X^{\dagger*,\eta_n}_{t_k}\notin\bigcup_{j:\ m_j\in G^{\text{large}}}B(m_j,2\Delta_n),\ X^{\dagger*,\eta_n}_{t_k}\neq\dagger\Big)\\
\leq&\mathbb{P}_x\Big(X^{\dagger*,\eta_n}_{t_k}\notin\bigcup_{j:\ m_j\in G^{\text{large}}}B(m_j,\Delta_n),\ X^{\dagger*,\eta_n}_{t_k}\neq\dagger\Big).
\end{aligned}$$

Apply (46) and we conclude the proof. □

3 Simulation Experiments

3.1 $\mathbb{R}^1$ experiment

Our numerical experiments in this section demonstrate that, (a) as indicated by Theorem 1, the minimum jump number defined in (3) accurately characterizes the first exit times of the SGDs with clipped heavy-tailed gradient noises; (b) with clipped heavy-tailed noises, sharp minima can be effectively eliminated from SGD; (c) properties studied in this paper are exclusive to heavy-tailed noises; under light-tailed noises SGDs are trapped in sharp minima for extremely long time. The test function $f \in \mathcal{C}^2(\mathbb{R})$ is the same one depicted in Figure 1 (Left,e). Note that 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. For all experiments on this f, heavy-tailed noises are $Z_n = 0.1U_nW_n$ where W_n are sampled from Pareto distributions with shape parameter $\alpha = 1.2$ and the signs U_n are iid RVs such that $\mathbb{P}(U_n = 1) = \mathbb{P}(U_n = -1) = 1/2$; All *light-tailed* noises are $\mathcal{N}(0, 1)$. See Appendix A for exact expression for test function f and experiment details.

In the first experiment, 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. To prevent excessively long running time, each simulation run is forced to terminate after 5×10^7 SGD iterations. According to Theorem 1, 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. As demonstrated in Figure 1 (Right), our results accurately predict how first exit time varies with learning rate and gradient clipping scheme.

Next, we investigate the global dynamics of heavy-tailed SGD. We compared the clipped case (with $b = 0.5$) against the case without gradient clipping. For each scenario we simulate 10 SGD paths, each of length $10,000,000$ iterations and initialized at $X_0 = 0.3$. Figure 1 (Left, a, b) show the histograms of the empirical distributions of SGD, and Figure 1(Middle, a,b) plots the SGD trajectories. We can clearly see that, 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, both of which are in larger attraction fields. This is exactly the dynamics predicted by Theorem 2-5: 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(Left,c,d) and Figure 1(Middle, c,d), light-tailed SGD are easily trapped at sharp minima for extremely long time.

3.2 $\mathbb{R}^d$ experiment

Figure 3 illustrates that the same phenomena are observed in $\mathbb{R}^2$. The details of the experiment are provided in Appendix A, but here we point out that the objective function f in this experiment 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. Besides, under gradient 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 proved in Section 2 holds under more general contexts than Assumptions 1-3.

4 Improving Generalization Performance in Deep Learning with Injected Heavy-Tailed Noises

In this section, we verify our theoretical results and demonstrate the effectiveness of clipped heavy-tailed noise in the deep learning context. Theorems in Section 2 suggest the possibility of modifying gradient noise distributions to drive SGD to "flatter" local minima and achieve better generalization performance. Meanwhile, recent experiments (such as [20, 33]) show that SGD noises in image classification tasks might be light-tailed in several cases. (As displayed in Appendix A, for tasks considered

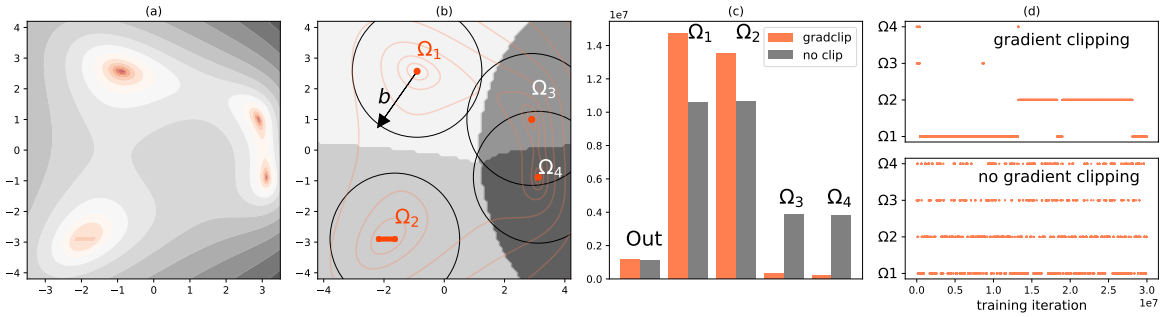

Figure 3: Experiment result of heavy-tailed SGD when optimizing the modified Himmelblau function. **(a)** Contour plot of the test function f. **(b)** Different shades of gray are used to indicate the area of the four different attraction fields $\Omega_1, \Omega_2, \Omega_3, \Omega_4$ of f. We say that 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 attraction field each iteration lies in.

in this section, the gradient noise distributions are not heavy-tailed either when models are randomly initialized.) Inspired by these facts, we conduct experiments to show that SGD with gradient clipping exhibits improved performance and prefers solutions with a flatter geometry if we *make the SGD noise heavy-tailed*. Let θ be the current model weight of a neural net during training, $g_{SB}(\theta)$ be the typical small-batch gradient direction, and $g_{GD}(\theta)$ be the true, deterministic gradient direction evaluated on the entire training dataset. Then by evaluating $g_{SB}(\theta) - g_{GD}(\theta)$ we obtain a sample of gradient noise in SGD. Due to the prohibitive cost of evaluating the true gradient in real-world deep learning tasks, we consider using $g_{SB}(\theta) - g_{LB}(\theta)$ as its approximation where g_{LB} denotes the gradient evaluated on a large batch of samples. This approximation is justified by the unbiasedness in $\mathbb{E}g_{LB}(\theta) = g_{GD}(\theta)$. For some heavy-tailed random variable Z, by multiplying Z with SGD noise, we obtain the following perturbed gradient direction

$$g_{heavy}(\theta) = g_{SB}(\theta) + Z\big(g_{SB*}(\theta) - g_{LB}(\theta)\big) \tag{47}$$

where SB and $SB*$ are two mini batches that may or may not be identical. In other words, we consider 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. In particular, we consider two different implementations: in *our method 1* (labeled as "our 1" in Table 2), SB and $SB*$ are chosen independently, while in *our method 2* (labeled as "our 2" in Table 2), we use the same batch of samples 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.

To concretely demonstrate the effect of clipped heavy-tailed noise, we 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 mostly adapted from [35]. We consider three different tasks: (1) LeNet [13] on corrupted FashionMNIST [32]; we use a 1200-sample subset of the original training dataset, and 200 samples points thereof are randomly assigned a label; (2)

Table 1: Hyperparameters for Training in Different Tasks

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 2: Test accuracy and expected sharpness of different methods across different tasks. The reported numbers are averaged over 5 replications. All the data points are provided in the appendix.

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

VGG11 [27] on SVHN [17], where we use a 25000-sample subset of the training dataset; (3) VGG11 on CIFAR10 [12] using the entire training dataset. For all tasks we use the entire test dataset when evaluating test accuracy. Whenever heavy-tailed noise is needed, the heavy-tailed multipliers used in the experiment are $Z_n = cW_n$ where W_n are iid Pareto(α) RVs. For details of the hyperparameters for training, see Table 1. 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 multiplier Z_n 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 Appendix A 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 2 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 report the *expected sharpness* metric $\mathbb{E}_{\nu\sim\mathcal{N}(\mathbf{0},\delta^2\mathbf{I})}|L(\theta^*+\nu)-L(\theta^*)|$ used in [35, 18]: 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,

[1]Another interpretation of our proposed heavy-tailed methods is that it is a simplified version of GD + annealed heavy-tailed perturbation, where a detailed annealing is substituted by a two-phase training schedule.

Table 3: Our method's gain on test accuracy persists even when applied with standard techniques for high-quality predictions: data augmentation and scheduled learning rates.

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 method	90.76%	90.57%	90.49%	90.85%	90.79%	**90.67%**
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 method	67.43%	65.12%	65.14%	65.96%	63.57%	**65.44%**

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 2. Smaller sharpness of clipped heavy-tailed SGD confirms its preference for minimizers that have a "flatter" geometry. In particular, the comparisons to *SB* confirm that, even in the deep learning context, clipped heavy-tailed noise can drive SGD to explore "wider" attraction fields, thus attaining better test performances.

Results in Table 2 shows that both heavy-tailed noise and gradient clipping are necessary to achieve better generalization, which is expected from our theoretical analyses. *SB* and *SB + Clip* achieve similar inferior performances, confirming that clipping does not help when noise is light-tailed. In *SB + Noise*, we inject heavy-tailed noise without gradient clipping, which still achieves an inferior performance. The poor performance of this method—even after extensive parameter tuning and engineering (see Appendix A for more details)—demonstrates the difficulty on the optimization front 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 [34, 4] where adaptive gradient clipping methods are proposed to improve convergence of SGD under heavy-tailed noises. This confirms that gradient clipping is crucial for heavy-tailed SGD.

Lastly, Table 3 shows that the gain of truncated heavy-tailed noise persists even when trained together with more sophisticated techniques—in particular, data augmentation and scheduled learning rates—to achieve higher performances. For experiment details, see Appendix A.

5 Conclusion

We characterized the global dynamics of SGD with truncated heavy-tailed gradient noise and illustrated our theoretical results with numerical experiments. Our characterizations provide key insights into the global dynamics of SGD and reveal the strong regularization effects of truncated heavy-tailed noises.

For the sake of analysis and presentation, we made some simplifying assumptions. However, as indicated by our numerical experiments, it is very likely that the same overall principle—elimination of sharp minima—persists in high-dimensional optimization landscapes that arise in deep learning contexts. We aim to develop the theories of heavy-tailed SGD at the full level of generality.

Also, the conditions—irreducibility and symmetry of graph $\mathcal{G}$—we impose on the geometry of f in Theorem 2 are likely to be stronger than necessary for eliminating sharp minima. In view of these, a natural next step is to investigate how the structures of state-of-the-art deep neural nets lend themselves to the SGD's global dynamics under general conditions.

Lastly, the results of our deep learning experiments suggest the possibility of a full-fledged, rigorously justified large-batch SGD algorithm with heavy-tailed perturbations that can overcome the generalization gap between small-batch and large-batch optimization methods. For instance, it is worth exploring whether the training would benefit from a more delicate approach for inducing heavy-tailed noises or a detailed annealing of the noise magnitude.

References

[1] P. Billingsley. *Convergence of probability measures.* John Wiley & Sons, 2013.

[2] 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.

[3] 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.

[4] 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.

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

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

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

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

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

[21] 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.

[22] I. Pavlyukevich. Metastable behaviour of small noise lévy-driven diffusion. *arXiv preprint math/0601771*, 2005.

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

[24] P. E. Protter. *Stochastic integration and differential equations*. Springer, 2005.

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

[26] C.-H. Rhee, J. Blanchet, B. 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.

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

[28] 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.

[29] 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.

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

[31] Y. Wen, K. Luk, M. Gazeau, G. Zhang, H. Chan, and J. Ba. An empirical study of stochastic gradient descent with structured covariance noise. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3621–3631. PMLR, 26–28 Aug 2020.

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

[33] Z. Xie, I. Sato, and M. Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

[34] 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.

[35] 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.

A Details of Numerical Experiments

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

The function f used in the experiments is

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

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 $L = 1.6$.

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

As illustrated in the contour plot in Figure 3 (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 3 (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 3 (a) and the red line segment in in Figure 3 (b).

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:

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

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 at $X_0 = (2.9, 1.0)$, which is very close to the local minimizer in 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 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.

A.3 Details of the deep learning experiments comparing different methods

We first mention that the all experiments using neural networks are conducted on Nvidia GeForce GTX 1080 T, and the scripts are adapted from the ones in [35].[2]. In Figure A.1 (Top), we display the gradient noise distribution in the three tasks after the model is randomly initialized.

The hyperparameters and training procedures for "*SB + Noise*" method is different from the other methods: we use learning rate $\eta = 0.005$; for FashionMNIST task we train for 100,000 iterations and

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

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 Figure A.1 (Bottom), we plot the test accuracy of our method against that of the SGD for all 5 replications and 3 tasks.

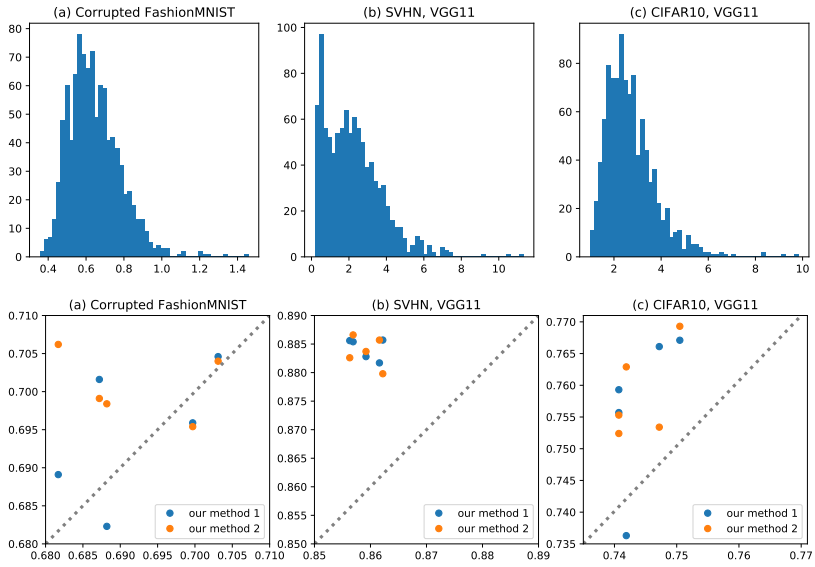

Figure A.1: **(Top)** Distribution of gradient noise in different tasks. **(Bottom)** Test accuracy of the proposed clipped heavy-tailed methods vs. test accuracy of vanilla SGD in different tasks.

A.4 Details of CIFAR10 experiments with data augmentation

For both methods, we train the model for 300 epochs. The initial learning rate is set at 0.1, and 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 (47), 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.

B Proofs of Lemma 7, 8

First, note that Assumption 1 implies the following:

- There exist $c_0 > 0, \epsilon_0 \in (0,1)$ such that for any $x \in \{m_1, s_1, \cdots, s_{n_{\min}-1}, m_{n_{\min}}\}, |y-x| < \epsilon_0$,

$$|f'(y)| > c_0|y-x|, \tag{B.1}$$

  and for any $y \in [-L, L]$ such that $|y-x| \geq \epsilon_0$ for all $x \in \{m_1, s_1, \cdots, s_{n_{\min}-1}, m_{n_{\min}}\}$, we have

$$|f'(y)| > c_0; \tag{B.2}$$

- There exist constants $L \in (0,\infty), M \in (0,\infty)$ such that $|m_0| < L, |m_{n_{\min}}| < L$, and (for any $x \in [-L, L]$)

$$|f'(x)| \leq M,\ |f''(x)| \leq M. \tag{B.3}$$

Recall that in (1) we define the truncation operator as $\varphi_c(w) \triangleq \varphi(w,c) \triangleq (w \wedge c) \vee (-c)$. Also, in (2) we defined the update recursion for SGD iterates under clipping threshold b as

$$X_n^\eta = \varphi_L\Big(X_n^\eta - \varphi_b\big(\eta(f'(X_n^\eta) - Z_{n+1})\big)\Big). \tag{B.4}$$

Here $\eta > 0$ is the learning rate (step length) and $b > 0$ is the gradient clipping threshold. Also, recall that we let $\sigma_k(\eta) \triangleq \min\{n \geq 0 :\ X_n^\eta \notin \Omega_k\}$ to be the first exit time of X_n^η from the $k-$th attraction field Ω_k. Meanwhile, recall that in (3) we have defined

$$r_i \triangleq \min\{m_i - s_{i-1}, s_i - m_i\}, \tag{B.5}$$

$$l_i^* \triangleq \lceil r_i/b \rceil. \tag{B.6}$$

Intuitively speaking, l_i^* tells us the minimum number of jumps with size no larger than b required in order to escape the attraction field if we start from the local minimum of this attraction field Ω_i. Lastly, recall the definition of $H(\cdot) = \mathbb{P}(|Z_1| > \cdot)$ and the scaling function for the $i-$th attraction field

$$\lambda_i(\eta) \triangleq H(1/\eta)\Big(\frac{H(1/\eta)}{\eta}\Big)^{l_i^*-1}.$$

The rest of this section is devoted to the proofs of Lemma 7 and Lemma 8. Specifically, Lemma 7 is an immediate Corollary of Lemma B.11, the proof of which will be provided below. The proof of Lemma 8 can be found at the end of this section.

The following three lemmas will be applied repeatedly throughout this section. The proofs are straightforward but provided in Appendix D for the sake of completeness.

Lemma B.1. *Given two real functions $a : \mathbb{R}_+ \mapsto \mathbb{R}_+$, $b : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $a(\epsilon) \downarrow 0, b(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$, and a family of geometric RVs $\{U(\epsilon) : \epsilon > 0\}$ with success rate $a(\epsilon)$ (namely, $\mathbb{P}(U(\epsilon) > k) = (1-a(\epsilon))^k$ for $k \in \mathbb{N}$), for any $c > 1$, there exists $\epsilon_0 > 0$ such that for any $\epsilon \in (0, \epsilon_0)$,*

$$\exp\Big(-\frac{c \cdot a(\epsilon)}{b(\epsilon)}\Big) \leq \mathbb{P}\Big(U(\epsilon) > \frac{1}{b(\epsilon)}\Big) \leq \exp\Big(-\frac{a(\epsilon)}{c \cdot b(\epsilon)}\Big).$$

Lemma B.2. *Given two real functions $a : \mathbb{R}_+ \mapsto \mathbb{R}_+$, $b : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $a(\epsilon) \downarrow 0, b(\epsilon) \downarrow 0$ and*

$$a(\epsilon)/b(\epsilon) \to 0$$

as $\epsilon \downarrow 0$, and a family of geometric RVs $\{U(\epsilon) : \epsilon > 0\}$ with success rate $a(\epsilon)$ (namely, $\mathbb{P}(U(\epsilon) > k) = (1-a(\epsilon))^k$ for $k \in \mathbb{N}$), for any $c > 1$ there exists some $\epsilon_0 > 0$ such that for any $\epsilon \in (0, \epsilon_0)$,

$$a(\epsilon)/(c \cdot b(\epsilon)) \leq \mathbb{P}(U(\epsilon) \leq 1/b(\epsilon)) \leq c \cdot a(\epsilon)/b(\epsilon)$$

Lemma B.3. *Suppose that a function $g : I_g \mapsto \mathbb{R}$ (where I_g is an open interval of $\mathbb{R}$) is $g \in \mathcal{C}^2(I_g)$ and $|g''(\cdot)| \leq C$ on its domain I_g for some constant $C < \infty$. For a finite integer n, a sequence of real numbers $\{z_1, \cdots, z_n\}$, and real numbers $x, \widetilde{x} \in I_g, \eta > 0$, consider two sequences $\{x_k\}_{k=0,\ldots,n}$ and $\{\widetilde{x}_k\}_{k=0,\ldots,n}$ constructed as follows:*

$$\begin{aligned}
&x_0 = x \\
&x_k = x_{k-1} - \eta g'(x_{k-1}) + \eta z_k \quad \forall k = 1, 2, \cdots, n \\
&\widetilde{x}_0 = \widetilde{x} \\
&\widetilde{x}_k = \widetilde{x}_{k-1} - \eta g'(\widetilde{x}_{k-1}) \quad \forall k = 1, 2, \cdots, n
\end{aligned}$$

If we have $x_k \in I_g, \widetilde{x}_k \in I_g$, and there exists $\widetilde{c} \in (0, \infty)$ such that $\eta|z_1 + \cdots + z_k| + |x - \widetilde{x}| \leq \widetilde{c}$ for $k = 1, 2, \cdots, n$, then

$$|x_k - \widetilde{x}_k| \leq \widetilde{c} \cdot \exp(\eta C k) \quad \forall k = 1, 2, \cdots, n.$$

To facilitate the analysis below, we introduce some additional notations. First, we will group the noises Z_n based on a threshold level $\delta > 0$: let us define

$$Z_n^{\leq \delta, \eta} \triangleq Z_n \mathbb{1}\{\eta|Z_n| \leq \delta\}, \tag{B.7}$$

$$Z_n^{> \delta, \eta} \triangleq Z_n \mathbb{1}\{\eta|Z_n| > \delta\}. \tag{B.8}$$

The former are viewed as *small noises* while the latter will be referred to as *large noises* or *large jumps*. Furthermore, for any $j \geq 1$, define the j^{th} arrival time and size of large jumps as

$$T_j^\eta(\delta) \triangleq \min\{n > T_{j-1}^\eta(\delta) : \ \eta|Z_n| > \delta\}, \quad T_0^\eta(\delta) = 0 \tag{B.9}$$

$$W_j^\eta(\delta) \triangleq Z_{T_j^\eta(\delta)}. \tag{B.10}$$

Next, for any $\epsilon > 0$, let $\Omega_i(\epsilon) = [m_i - \epsilon, m_i + \epsilon]$ be an $\epsilon-$neighborhood of the local minimum m_i, and $S_i(\epsilon) = [s_i - \epsilon, s_i + \epsilon]$ be an $\epsilon-$neighborhood of the local maximum s_i.

For most part of this section, we will zoom in on one of the local minima m_i and its attraction field $\Omega_i = (s_{i-1}, s_i)$. Without loss of generality, we assume $m_i = 0$, and denote the attraction field as $\Omega = (s_-, s_+)$. (If m_i happens to be the local minimum at the left or right boundary, then the attraction field is $[-L, s_+)$ or $(s_-, L]$ where the SGD iterates will be reflected at $\pm L$.) Henceforth we will drop the dependency on notation i when referring to this specific attraction field until the very end of this section. Throughout the proof, the following (deterministic) dynamic systems will be used frequently as benchmark processes to indicate the most likely location of the SGD iterates. Specifically, given any $x \in \Omega$, we use $X_n(x)$ to indicate that the starting point is x, namely $X_0(x) = x$. Similarly, consider the following ODE $\mathbf{x}^\eta(t; x)$ as

$$\mathbf{x}^\eta(0; x) = x; \tag{B.11}$$

$$\frac{d\mathbf{x}^\eta(t; x)}{dt} = -\eta f'\Big(\mathbf{x}^\eta(t; x)\Big). \tag{B.12}$$

When we use update rate $\eta = 1$, we will drop the dependency of η and simply use $\mathbf{x}(t; x)$ to denote the process.

Based on Assumption 3, we know the existence of some constant $\bar{\epsilon} \in (0, \epsilon_0)$ (note that ϵ_0 is the constant in (B.1)) such that

$$r \triangleq \min\{-s_-, s_+\}, \tag{B.13}$$

$$l^* \triangleq \lceil r/b \rceil, \tag{B.14}$$

$$(l^* - 1)b + 100l^*\bar{\epsilon} < r - 100l^*\bar{\epsilon} \tag{B.15}$$

$$r + 100l^*\bar{\epsilon} < l^*b - 100l^*\bar{\epsilon}. \tag{B.16}$$

Here r can be understood as the effective *radius* of the said attraction field. Also, we fix such $\bar{\epsilon}$ small enough so that (let $c_-^L = -f'(-L), c_+^L = -f'(-L)$), we have

$$0.9c_-^L \leq -f'(x) \leq 1.1c_-^L \quad \forall x \in [-L, -L + 100\bar{\epsilon}], \quad 0.9c_+^L \geq -f'(x) \geq 1.1c_+^L \quad \forall x \in [L - 100\bar{\epsilon}, L]. \tag{B.17}$$

Similar to the definition of ODE $\mathbf{x}^\eta$, let us consider the following construction of ODE $\tilde{\mathbf{x}}^\eta$ that can be understood as $\mathbf{x}^\eta$ perturbed by l^* shocks. Specifically, consider a sequence or real numbers $0 = t_1 < t_2 < t_3 < \cdots < t_{l^*}$ and real numbers $w_1, \cdots, w_{l^*}$ where $|w_j| \leq b$ for each j. Let $\mathbf{t} = (t_1, \cdots, t_{l^*}), \mathbf{w} = (w_1, \cdots, w_{l^*})$. Based on these two sequences and rate $\eta > 0$, define $\tilde{\mathbf{x}}^\eta(t; x)$ as

$$\tilde{\mathbf{x}}^\eta(0, x; \mathbf{t}, \mathbf{w}) = \varphi_L(x + \varphi_b(w_1)); \tag{B.18}$$

$$\frac{d\tilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w})}{dt} = -\eta f'\big(\tilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w})\big) \quad \forall t \notin \{t_1, t_2, \cdots, t_{l^*}\} \tag{B.19}$$

$$\tilde{\mathbf{x}}^\eta(t_j, x; \mathbf{t}, \mathbf{w}) = \varphi_L\big(\tilde{\mathbf{x}}^\eta(t_j-, x; \mathbf{t}, \mathbf{w}) + \varphi_b(w_j)\big) \quad \forall j = 2, \cdots, l^* \tag{B.20}$$

Again, when $\eta = 1$ we drop the notational dependency on η and use $\tilde{\mathbf{x}}$ to denote the process. Now from (B.15)(B.16) one can easily see the following fact: there exist constants $\bar{t}, \bar{\delta} > 0$ such that $\tilde{\mathbf{x}}(t_{l^*}, 0; \mathbf{t}, \mathbf{w}) \notin \Omega$ (note that the starting point is 0, the local minimum) only if (under the condition that $|w_j| \leq b \ \ \forall j$)

$$t_j - t_{j-1} \leq \bar{t} \ \ \forall j = 2, 3, \cdots, l^* \tag{B.21}$$

$$|w_j| > \bar{\delta} \ \ \forall j = 1, 2, \cdots, l^*. \tag{B.22}$$

The intuition is as follows: if the inter-arrival time between any of the l^* jumps is too long, then the path of $\tilde{\mathbf{x}}^\eta(t; x)$ will drift back to the local minimum m_i so that the remaining $l^* - 1$ shocks (whose sizes are bounded by b) cannot overcome the radius r which is strictly larger than $(l^* - 1)b$; similarly, if size of any of the shocks is too small, then since all other jumps have sizes bounded by b, the shock created by the l_i^* jumps will be smaller than $(l^* - 1)b + 100\bar{\epsilon}$, which is strictly less than r. We fix these constants $\bar{t}, \bar{\delta}$ throughout the analysis, and stress again that their values are dictated by the geometry of the function f, thus *do not* vary with the accuracy parameters ϵ and δ mentioned earlier. In particular, choose $\bar{\delta}$ such that $\bar{\delta} < \bar{\epsilon}$.

In our analysis below, $\epsilon > 0$ will be a variable representing the level of *accuracy* in our analysis. For instance, for small ϵ, the chance that SGD iterates will visit somewhere that is $\epsilon-$close to s_- or s_+ (namely, the boundary of the attraction filed) should be small. Consider some $\epsilon \in (0, \epsilon_0)$ where ϵ_0 is the constant in Assumption 1. Due to (B.1)(B.2), one can see the existence of some $g_0 > 0, c_1 < \infty$ such that

- $|f'(x)| \geq g_0$ for any $x \in \Omega$ such that $|x - s_-| > \epsilon_0, ||x - s_+| > \epsilon_0$;
- Let $\hat{t}_{\text{ODE}}(x, \eta) \triangleq \min\{t \geq 0 : \ \mathbf{x}^\eta(t, x) \in [-\epsilon, \epsilon]\}$ be the time that the ODE returns to a $\epsilon-$neighborhood of local minimum of Ω when starting from x. As proved in Lemma 3.5 of [22], for any $x \in \Omega$ such that $|x - s_-| > \epsilon, |x - s_+| > \epsilon$, we have

$$\hat{t}_{\text{ODE}}(x, \eta) \leq c_1 \frac{\log(1/\epsilon)}{\eta} \tag{B.23}$$

  and we define the function

$$\hat{t}(\epsilon) \triangleq c_1 \log(1/\epsilon). \tag{B.24}$$

In short, given any accuracy level ϵ, the results above give us an upper bound for how fast the ODE would return to a neighborhood of the local minimum, if the starting point is not too close to the boundary of this attraction field Ω.

For the first few technical results established below, we show that, without large jumps, the SGD iterates $X_n^\eta(x)$ are unlikely to show significant deviation from the deterministic gradient descent process $\mathbf{y}_n^\eta(x)$ defined as

$$\mathbf{y}_0^\eta(x) = x, \tag{B.25}$$

$$\mathbf{y}_n^\eta(x) = \mathbf{y}_{n-1}^\eta(x) - \eta f'\Big(\mathbf{y}_{n-1}^\eta(x)\Big). \tag{B.26}$$

We are ready to state the first lemma, where we bound the distance between the gradient descent iterates $\mathbf{y}_n^\eta(y)$ and the ODE $\mathbf{x}^\eta(t,x)$ when the initial conditions x, y are close enough.

Lemma B.4. *The following claim holds for all $\eta > 0$: for any $t > 0$, we have*

$$\sup_{s\in[0,t]} |\boldsymbol{x}^\eta(s,x) - \boldsymbol{y}_{\lfloor s\rfloor}^\eta(y)| \le (2\eta M + |x-y|)\exp(\eta M t)$$

where $M \in (0,\infty)$ is the constant in (B.3).

Proof. Define a continuous-time process $\mathbf{y}^\eta(s;y) \triangleq \mathbf{y}_{\lfloor s\rfloor}^\eta(y)$, and note that

$$\mathbf{x}^\eta(s,x) = \mathbf{x}^\eta(\lfloor s\rfloor, x) - \eta\int_{\lfloor s\rfloor}^{s} f'(\mathbf{x}^\eta(u,x))du$$

$$\mathbf{x}^\eta(\lfloor s\rfloor, x) = x - \eta\int_0^{\lfloor s\rfloor} f'(\mathbf{x}^\eta(u,x))du$$

$$\mathbf{y}_{\lfloor s\rfloor}^\eta(y) = \mathbf{y}^\eta(\lfloor s\rfloor, y) = y - \eta\int_0^{\lfloor s\rfloor} f'(\mathbf{y}^\eta(u,y))du.$$

Therefore, if we define function

$$b(u) = \mathbf{x}^\eta(u,x) - \mathbf{y}^\eta(u,y),$$

from the fact $|f'(\cdot)| \le M$, one can see that $|b(u)| \le \eta M + |x-y|$ for any $u \in [0,1)$ and $|b(1)| \le 2\eta M + |x-y|$. In case that $s > 1$, from the display above and the fact $|f''(\cdot)| \le M$, we now have

$$|\mathbf{y}_{\lfloor s\rfloor}^\eta(x) - \mathbf{x}^\eta(s,x)| \le |b(\lfloor s\rfloor)| + \eta M;$$

$$|b(\lfloor s\rfloor)| \le \eta M\int_1^{\lfloor s\rfloor} |b(u)|du.$$

From Gronwall's inequality (see Theorem 68, Chapter V of [24], where we let function $\alpha(u)$ be $\alpha(u) = |b(u+1)|$), we have

$$|\mathbf{y}_{\lfloor s\rfloor}^\eta(x) - \mathbf{x}^\eta(s,x)| \le (2\eta M + |x-y|)\exp(\eta M t).$$

This concludes the proof. □

Now we consider an extension of the previous Lemma in the following sense: we add perturbations to the gradient descent process and ODE, and show that, when both perturbed by l^* similar perturbations, the ODE and gradient descent process should still stay close enough. Analogous to the definition of the perturbed ODE $\widetilde{x}^\eta$ in (B.18)-(B.20), we can construct a process $\widetilde{Y}^\eta$ as a perturbed gradient descent process as follows. For a sequence of integers $0 = t_1 < t_2 < \cdots < t_{l^*}$ (let $\mathbf{t} = (t_j)_{j\ge1}$) and a sequence of real numbers $\widetilde{w}_1, \cdots, \widetilde{w}_{l^*}$ (let $\widetilde{\mathbf{w}} = (\widetilde{w}_j)_{j\ge1}$) and $y \in \mathbb{R}$, define (for all $n = 1, 2, \cdots, t_{l^*}$) the perturbed gradient descent iterates with gradient clipping at b and reflection at $\pm L$ as

$$\widetilde{\mathbf{y}}_n^\eta(y;\mathbf{t},\widetilde{\mathbf{w}}) = \varphi_L\Big(\widetilde{\mathbf{y}}_{n-1}^\eta(y;\mathbf{t},\widetilde{\mathbf{w}}) + \varphi_b\big(-\eta f'(\widetilde{\mathbf{y}}_{n-1}^\eta(y;\mathbf{t},\widetilde{\mathbf{w}})) + \sum_{j=2}^{l^*}\mathbb{1}\{n=t_j\}\widetilde{w}_j\big)\Big) \tag{B.27}$$

with initial condition $\widetilde{\mathbf{y}}_0^\eta(y;\mathbf{t},\widetilde{\mathbf{w}}) = \varphi_L\big(y + \varphi_b(\widetilde{w}_1)\big)$.

Corollary B.5. *Given any $\epsilon > 0$, the following claim holds for all sufficiently small $\eta > 0$: for any $x, y \in \Omega$, and sequence of integers $\boldsymbol{t} = (t_j)_{j=1}^{l^*}$ and any two sequence of real numbers $\boldsymbol{w} = (w_j)_{j=1}^{l^*}, \widetilde{\boldsymbol{w}} = (\widetilde{w}_j)_{j\geq 1}^{l^*}$ such that*

- *$|x - y| < \epsilon$;*
- *$t_1 = 0$, and $t_j - t_{j-1} \leq 2\bar{t}/\eta$ for all $j \geq 1$ where $\bar{t}$ is the constant in* (B.21)*;*
- *$|w_j - \widetilde{w}_j| < \epsilon$ for all $j \geq 1$;*

then we have

$$\sup_{t\in[0,t_{l^*}]} |\widetilde{\boldsymbol{x}}^\eta(t, x; \boldsymbol{t}, \boldsymbol{w}) - \widetilde{\boldsymbol{y}}^\eta_{\lfloor t \rfloor}(y; \boldsymbol{t}, \widetilde{\boldsymbol{w}})| \leq \bar{\rho}\epsilon$$

where the constant $\bar{\rho} = (3\exp(\eta M\bar{t}) + 3)^{l^*}$.

Proof. Throughout this proof, fix some $\eta \in (0, \epsilon/2M)$. We will show that for any η in the range the claim would hold.

First, on interval $[0, t_2)$, from Lemma B.4, one can see that (since $2M\eta < \epsilon$)

$$\sup_{t\in[0,t_2)} |\widetilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w}) - \widetilde{\mathbf{y}}^\eta_{\lfloor t \rfloor}(y; \mathbf{t}, \widetilde{\mathbf{w}})| \leq 3\exp(\eta M\bar{t}) \cdot \epsilon.$$

The at $t = t_2$, by considering the difference between w_2 and $\widetilde{w}_2$, and the possible change due to one more gradient descent step (which is bounded by $\eta M < \epsilon$), we have

$$\sup_{t\in[0,t_2]} |\widetilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w}) - \widetilde{\mathbf{y}}^\eta_{\lfloor t \rfloor}(y; \mathbf{t}, \widetilde{\mathbf{w}})| \leq (3\exp(\eta M\bar{t}) + 2) \cdot \epsilon.$$

Now we proceed inductively. For any $j = 2, 3, \cdots, l^* - 1$, assume that

$$\sup_{t\in[0,t_j]} |\widetilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w}) - \widetilde{\mathbf{y}}^\eta_{\lfloor t \rfloor}(y; \mathbf{t}, \widetilde{\mathbf{w}})| \leq (3\exp(\eta M\bar{t}) + 3)^{j-1} \cdot \epsilon.$$

Then by focusing on interval $[t_j, t_{j+1}]$ and using Lemma B.4 again, one can show that

$$\begin{aligned}\sup_{t\in[t_j,t_{j-1}]} |\widetilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w}) - \widetilde{\mathbf{y}}^\eta_{\lfloor t \rfloor}(y; \mathbf{t}, \widetilde{\mathbf{w}})| &\leq 2\epsilon + \left((3\exp(\eta M\bar{t}) + 3)^{j-1} + 1\right)\exp(\eta M\bar{t})\epsilon \\ &\leq (3\exp(\eta M\bar{t}) + 3)^j \cdot \epsilon.\end{aligned}$$

This concludes the proof. □

In the next few results, we show that the same can be said for gradient descent iterates $\widetilde{\mathbf{y}}_n$ and the SGD iterates X_n. Specifically, our first goal is to show that before any *large* jump (see the definition in B.8), it is unlikely that the gradient descent process $\mathbf{y}_n^\eta$ would deviate too far from X_n^η. Define the event

$$A(n, \eta, \epsilon, \delta) = \left\{ \max_{k=1,2,\cdots,n\wedge(T_1^\eta(\delta)-1)} \eta|Z_1 + \cdots + Z_k| \leq \epsilon \right\} \tag{B.28}$$

and recall that arrival times $T_j^\eta(\delta)$ are defined in (B.9).

As a building block, we first study the case when the starting point x is close to the reflection boundary $-L$. The takeaway from the next result is that the reflection operator hardly comes into play, since the SGD iterates would most likely quickly move to somewhere far enough from $\pm L$; besides, throughout this procedure the SGD iterates would most likely stay pretty close to the corresponding deterministic gradient descent process.

Lemma B.6. *Given $\epsilon \in (0, \bar{\epsilon}/9)$, it holds for any sufficiently small $\epsilon, \delta, \eta > 0$ that, if $x \in [-L, -L+\bar{\epsilon}]$ and $\rho_0(|x-y|+9\epsilon) < \bar{\epsilon}$, then on event $A(n, \eta, \epsilon, \delta)$ we have*

$$|X_k^\eta(x) - \boldsymbol{y}_k^\eta(y)| \le \rho_0 \cdot (|x-y|+9\epsilon) \quad \forall k = 1, 2, \cdots, n \wedge (T_1^\eta(\delta) - 1) \wedge \widetilde{T}_{escape}^\eta(x)$$

where $\widetilde{T}_{escape}^\eta(x) \triangleq \min\{n \ge 0 : \ X_n^\eta(x) > -L + \bar{\epsilon}\}$ and $\rho_0 \triangleq \exp\left(\frac{2M\bar{\epsilon}}{0.9c_-^L}\right)$ is a constant that does not vary with our choice of ϵ, δ, η.

Proof. For any $k < T_1^\eta(\delta)$, we know that $Z_k = Z_k^{\le\delta}$ (thus $\eta|Z_k| < \delta$). Also, recall that $|f'(x)| \le M$ for any $x \in \Omega_i$. Therefore, as long as η and δ are small enough, we will have that

$$|\eta(-f'(X_n^\eta(x)) + Z_k^{\le\delta})| \le b \tag{B.29}$$

so the gradient clipping operator in (B.4) has no effect when $k < T_1^\eta(\delta)$, and in fact the only possible time for the gradient clipping trick to work is at $T_j^\eta(\delta)$. Therefore, we can safely rewrite the SGD update as

$$X_k^\eta(x) = X_{k-1}^\eta(x) - \eta f'(X_{k-1}^\eta(x)) + \eta Z_k + R_k \quad \forall k < T_1^\eta(\delta)$$

where each $R_k \ge 0$ and it represents the push caused by reflection at $-L$.

First, choose ϵ small enough so that $9\epsilon < \bar{\epsilon}$. Next, based on (B.17) we have the following lower bound:

$$X_k^\eta(x) \ge x + 0.9c_-^L\eta k - \epsilon \quad \forall k < T_1^\eta(\delta).$$

Let $\widetilde{t}_0(x, \epsilon) \triangleq \min\{n \ge 0 : X_n^\eta(x) \ge -L + 2\epsilon\}$. Due to the inequality above, we know that

$$\widetilde{t}_0(x, \epsilon) \le \frac{3\epsilon}{0.9c_-^L\eta}. \tag{B.30}$$

One the other hand, given the current choice of ϵ, if we choose η and δ small enough, then using the same argument leading to (B.29), we will have

$$X_{\widetilde{t}_0(x,\epsilon)}^\eta(x) \le -L + 2.1\epsilon \le x + 1.1c_-^L\eta k + 2.1\epsilon$$

if $\widetilde{t}_0(x, \epsilon) \ge 1$ (namely $x < -L + 2\epsilon$).

Let us inspect the two scenarios separately. First, assume $\widetilde{t}_0(x, \epsilon) \ge 1$. For the deterministic gradient descent process $\mathbf{y}_n^\eta(y)$, we have the following bounds:

$$y + 0.9c_-^L\eta k \le \mathbf{y}_k^\eta(y) \le y + 1.1c_-^L\eta k \ \ \forall k \le \widetilde{t}_0(x, \epsilon) \wedge (T_1^\eta(\delta) - 1).$$

This gives us

$$|X_k^\eta(x) - \mathbf{y}_k^\eta(y)| \le |x-y| + 0.2c_-^L\eta k + 2.1\epsilon \ \ \forall k \le \widetilde{t}_0(x, \epsilon) \wedge (T_1^\eta(\delta) - 1).$$

At time $k = \widetilde{t}_0(x, \epsilon)$, due to previous bound on $\widetilde{t}_0(x, \epsilon)$, we know that $|X_{\widetilde{t}_0(x,\epsilon)}^\eta(x) - \mathbf{y}_{\widetilde{t}_0(x,\epsilon)}^\eta(y)| \le |x-y| + 7\epsilon$. If $n \wedge (T_1^\eta(\delta) - 1) \le \widetilde{t}_0(x, \epsilon)$ then we have already shown the desired claim. Otherwise, starting from time $\widetilde{t}_0(x, \epsilon)$, due to the definition of event $A(n, \eta, \epsilon, \delta)$ in (B.28), we know that the SGD iterates $X_n^\eta(x)$ will not touch the boundary $-L$ afterwards. Therefore, by directly applying Lemma B.3, and notice that $|f''(x)| \le M$ for any $x \in [-L, L]$ and, we have

$$|X_k^\eta(x) - \mathbf{y}_k^\eta(y)| \le (|x-y| + 9\epsilon) \cdot \exp\left(\frac{2M\bar{\epsilon}}{0.9c_-^L}\right) \quad \forall k \le n \wedge (T_1^\eta(\delta) - 1) \wedge \widetilde{T}_{\text{escape}}^\eta(x).$$

Indeed, it suffices to use Lemma B.3 for the next $\lceil 2\bar{\epsilon}/(0.9\eta c_-^L)\rceil$ steps to show that $|X_k^\eta(x) - \mathbf{y}_k^\eta(y)|$ should be smaller than ϵ for the next $\lceil 2\bar{\epsilon}/(0.9\eta c_-^L)\rceil$ steps, while $\mathbf{y}_k^\eta(y)$ will reach some where in $(-L+2\bar{\epsilon}, -L+3\bar{\epsilon})$ within $\lceil 2\bar{\epsilon}/(0.9\eta c_-^L)\rceil$ steps so we must have

$$n \wedge (T_1^\eta(\delta) - 1) \wedge \widetilde{T}_{\text{escape}}^\eta(x) \wedge \widetilde{t}_0(x,\epsilon) - \widetilde{t}_0(x,\epsilon) \le 2\bar{\epsilon}/(0.9\eta c_-^L) \tag{B.31}$$

Lastly, in the case $\widetilde{t}_0(x,\epsilon) = 0$ (which means $x \ge -L+\epsilon$), we can use Lemma B.3 directly as we did above and establish the same bound. This concludes the proof. □

Obviously, a similar result can be shown if x is in the rightmost attraction field $(s_{n_{\min-1}}, L]$ and the approach is identical. We omit the details here. In the next Lemma, we consider the scenario where the starting point x is far enough from the boundaries.

Lemma B.7. *Given any $\epsilon > 0$, the following holds for all sufficiently small $\eta > 0$: for any $x, y \in \Omega$ and positive integer n such that $|x-L| > 2\epsilon, |x+L| > 2\epsilon, |x-s_-| > 2\epsilon, |x-s_+| > 2\epsilon$ and $|x-y| < \frac{\epsilon}{2\exp(\eta M n)}$, on event*

$$A(n,\eta,\frac{\epsilon}{2\exp(\eta M n)},\delta) \cap \Big\{ |\boldsymbol{y}_j^\eta(y)| \in \Omega, |X_j^\eta(x)| \in \Omega \ \ \forall j = 1,2,\cdots,n \wedge (T_1^\eta(\delta)-1) \Big\}$$

we have

$$|\widetilde{X}_m^\eta(y) - X_m^\eta(x)| \le \epsilon \ \ \forall m = 1,2,\cdots,n \wedge (T_1^\eta(\delta)-1).$$

Proof. For sufficiently small η, we will have that the (deterministic) gradient descent iterates $|\mathbf{y}_n^\eta|$ is monotonically decreasing in n, which ensures that $\mathbf{y}_n^\eta$ always stays in the range that are at least $\epsilon-$away from $\pm L$ or s_-, s_+. We now show that the claim holds for any such η.

On event $\Big\{ |\mathbf{y}_j^\eta(y)| \in \Omega, |X_j^\eta(x)| \in \Omega \ \ \forall j = 1,2,\cdots,n \wedge (T_1^\eta(\delta)-1) \Big\}$, we are able to apply Lemma B.3 inductively for any $m \in [n]$ and obtain that

$$|\mathbf{y}_j^\eta(y) - X_j^\eta(x)| \le (|x-y| + \frac{\epsilon}{2\exp(\eta M n)})\exp(\eta M j) < \epsilon \quad \forall j = 1,2,\cdots,m$$

and conclude the proof. The reason to apply the Lemma inductively for $m = 1,2,\cdots,n$, instead of directly at step n, is to ensure that SGD iterates X_n^η would not hit the boundary $\pm L$ (so the reflection operator would not come into play on the time interval we are currently interested in), thus ensuring that Lemma B.3 is applicable. □

Similar to the extension from Lemma B.4 to Corollary B.5, we can extend Lemma B.7 to show that, if we consider the a gradient descent process that is only perturbed by large noises, then it should stay pretty close to the SGD iterates X_n^η. To be specific, let

$$Y_0^\eta(x) = x \tag{B.32}$$

$$Y_n^\eta(x) = \varphi_L\Big(Y_{n-1}^\eta(x) - \varphi_b\big(-\eta f'\big(Y_{n-1}^\eta(x)\big) + \sum_{j\ge 1} \mathbb{1}\{n = T_j^\eta(\delta)\}\eta Z_n\big)\Big). \tag{B.33}$$

be a gradient descent process (with gradient clipping at threshold b) that is only shocked by large noises in $(Z_n)_{n\ge 1}$. The next corollary can be shown by an approach that is identical to Corollary B.5 (namely, inductively repeating Lemma B.7 at each jump time) so we omit the details here.

Corollary B.8. *Given any $\epsilon > 0$, the following holds for any sufficiently small $\eta > 0$: For any $|x| < 2\epsilon$, on event $A_0(\epsilon,\eta,\delta) \cap B_0(\epsilon,\eta,\delta)$, we have*

$$|Y_n^\eta(x) - X_n^\eta(x)| < \widetilde{\rho}\epsilon \ \ \forall n = 1,2,\cdots,T_{l^*}^\eta(\delta)$$

where

$$A_0(\epsilon,\eta,\delta) \triangleq \Big\{\forall i=1,\cdots,l^*,\ \max_{j=T^\eta_{i-1}(\delta)+1,\cdots,T^\eta_i(\delta)-1} \eta|Z_{T^\eta_{i-1}(\delta)+1}+\cdots+Z_j| \le \frac{\epsilon}{2\exp(2\bar{t}M)}\Big\};$$
$$B_0(\epsilon,\eta,\delta) \triangleq \Big\{\forall j=2,\cdots,l^*, T^\eta_j(\delta)-T^\eta_{j-1}(\delta)\le 2\bar{t}/\eta\Big\}$$

and $\widetilde{\rho}\in(0,\infty)$ *is a constant that does not vary with* η,δ,ϵ.

The next two results shows that the type of events $A(n,\eta,\epsilon,\delta)$ defined in (B.28) is indeed very likely to occur, especially for small ϵ. For clarity of the presentation, we introduce the following definitions that are slightly more general than the *small* and *large* jumps defined in (B.7)(B.8) (for any $c>0$)

$$Z_n^{\le c} \triangleq Z_n \mathbb{1}\{|Z_n|\le c\},$$
$$Z_n^{>c} \triangleq Z_n \mathbb{1}\{|Z_n|> c\}.$$

Lemma B.9. *Define functions* $u(\eta)=\delta/\eta^{1-\Delta},\ v(\eta)=\epsilon\eta^{\widetilde{\Delta}}$ *with* $\epsilon,\delta>0$. *If real numbers* $\Delta,\widetilde{\Delta},\beta,\epsilon,\delta$ *and positive integers* j,N *are such that the following conditions hold:*

$$\Delta\in\big[0,(1-\frac{1}{\alpha})\wedge\frac{1}{2}\big), \tag{B.34}$$
$$\beta\in\big(1,(2-2\Delta)\wedge\alpha(1-\Delta)\big), \tag{B.35}$$
$$\widetilde{\Delta}\in[0,\frac{\Delta}{2}],\ \widetilde{\Delta}<\alpha(1-\Delta)-\beta, \tag{B.36}$$
$$N<\big(\alpha(1-\Delta)-\beta\big)j, \tag{B.37}$$
$$v(\eta)-j\eta u(\eta)\ge v(\eta)/2 \quad \textit{for all } \eta>0 \textit{ sufficiently small}, \tag{B.38}$$

then

$$\mathbb{P}\Big(\max_{k=1,2,\cdots,\lceil 1/\eta^\beta\rceil}\eta\big|Z_1^{\le u(\eta)}+\cdots+Z_k^{\le u(\eta)}\big|>3v(\eta)\Big)=o(\eta^N)$$

as $\eta\downarrow 0$.

Proof. From the stated range of the parameters, we know that

$$\alpha(1-\Delta)>\beta,$$
$$\big(\alpha(1-\Delta)-\beta\big)j>N,$$

so we are able to find $\gamma\in(0,1)$ small enough such that

$$\alpha(1-\Delta)(1-2\gamma)>\beta, \tag{B.39}$$
$$\big(\alpha(1-\Delta)(1-2\gamma)-\beta\big)j>N. \tag{B.40}$$

Fix such $\gamma\in(0,1)$ for the rest of the proof, and let $n(\eta)\triangleq\lceil(1/\eta)^\beta\rceil$, $I\triangleq\#\Big\{i\in[n(\eta)]:|Z_i^{\le u(\eta)}|>u(\eta)^{1-\gamma}\Big\}$. Then

$$\mathbb{P}\Big(\big|Z_1^{\le u(\eta)}+\cdots+Z_{n(\eta)}^{\le u(\eta)}\big|>v(\eta)\Big)$$
$$=\sum_{i=0}^{j-1}\underbrace{\mathbb{P}\Big(\big|Z_1^{\le u(\eta)}+\cdots+Z_{n(\eta)}^{\le u(\eta)}\big|>v(\eta),\ I=i\Big)}_{\triangleq(\mathrm{I})}+\underbrace{\mathbb{P}\Big(\big|Z_1^{\le u(\eta)}+\cdots+Z_{n(\eta)}^{\le u(\eta)}\big|>v(\eta),\ I\ge j\Big)}_{\triangleq(\mathrm{II})}$$

Note that since $|Z_i^{\le u(\eta)}| < u(\eta)$,

$$\begin{aligned}
\text{(I)} &\le \binom{n(\eta)}{i} \cdot \mathbb{P}\Big(\big|Z_1^{\le u(\eta)} + \cdots + Z_{n(\eta)-i}^{\le u(\eta)}\big| > \frac{v(\eta) - i\eta u(\eta)}{\eta}, |Z_i^{\le u(\eta)}| \le u(\eta)^{1-\gamma} \ \forall i \in [n(\eta) - i]\Big) \\
&\le n(\eta)^i \cdot \mathbb{P}\Big(\big|Z_1^{\le u(\eta)^{1-\gamma}} + \cdots + Z_{n(\eta)-i}^{\le u(\eta)^{1-\gamma}}\big| > \frac{v(\eta) - i\eta u(\eta)}{\eta}\Big) \\
&\le n(\eta)^i \cdot \mathbb{P}\Big(\big|Z_1^{\le u(\eta)^{1-\gamma}} + \cdots + Z_{n(\eta)-i}^{\le u(\eta)^{1-\gamma}}\big| > \frac{v(\eta)}{2\eta}\Big) \qquad \text{(B.41)}
\end{aligned}$$

where the last inequality follows from (B.38). First, since $\mathbb{E}Z_1 = 0$, we have

$$\begin{aligned}
\big|\mathbb{E}Z_1^{\le u(\eta)^{1-\gamma}}\big| &= \big|\mathbb{E}Z_1^{> u(\eta)^{1-\gamma}}\big| \\
&= \int_{u(\eta)^{1-\gamma}}^{\infty} \mathbb{P}(|Z_1| > x)dx \in \mathcal{RV}_{(\alpha-1)(1-\gamma)(1-\Delta)}(\eta).
\end{aligned}$$

Therefore, for all $\eta > 0$ that are sufficiently small,

$$\begin{aligned}
&|\mathbb{E}Z_1^{\le u(\eta)^{1-\gamma}} + \cdots + \mathbb{E}Z_{n(\eta)-i}^{\le u(\eta)^{1-\gamma}}| \\
\le& n(\eta) \cdot \eta^{(\alpha-1)(1-\Delta)(1-2\gamma)} \le 2\eta^{(\alpha-1)(1-\Delta)(1-2\gamma)-\beta} \\
\le& (1/\eta)^{(1-\Delta)(1-2\gamma)} \quad \text{due to (B.39)} \\
\le& \frac{v(\eta)}{4\eta} \quad \text{due to } \widetilde{\Delta}/2 \le \Delta \text{ in (B.36) and } 1-\gamma < 1.
\end{aligned}$$

If we let $Y_n = Z_n^{\le u(\eta)^{1-\gamma}} - \mathbb{E}Z_n^{\le u(\eta)^{1-\gamma}}$ and plug the bound above back into (B.41), then (for all $\eta > 0$ that are sufficiently small)

$$\begin{aligned}
\text{(I)} &\le n(\eta)^i \cdot \mathbb{P}(|Y_1 + \cdots + Y_{n(\eta)-i}| > \frac{v(\eta)}{4\eta}) \\
&\le n(\eta)^i \exp\left(-\frac{\frac{\epsilon^2}{16} \cdot 1/\eta^{2-2\widetilde{\Delta}}}{2\big(n(\eta)-i\big)\mathbb{E}|Y_1|^2 + \frac{2}{3}\delta^{1-\gamma} \cdot (1/\eta)^{(1-\Delta)(1-\gamma)} \cdot \frac{\epsilon}{4}/\eta^{1-\widetilde{\Delta}}}\right) \qquad \text{(B.42)}
\end{aligned}$$

where the last inequality is obtained from Bernstein's inequality. Note that from Karamata's theorem,

$$\begin{aligned}
\mathbb{E}|Y_1|^2 = var(Z_1^{\le u(\eta)^{1-\gamma}}) &\le \mathbb{E}|Z_1^{\le u(\eta)^{1-\gamma}}|^2 \\
&\le \int_0^{u(\eta)^{1-\gamma}} 2x\mathbb{P}(|Z_1| > x)dx \in \mathcal{RV}_{-(1-\Delta)(1-\gamma)(2-\alpha)}(\eta).
\end{aligned}$$

Now note that

- In case that $\alpha < 2$, for all $\eta > 0$ that are sufficiently small, we have (using (B.36))

$$\begin{aligned}
&2\big(n(\eta)-i\big)\mathbb{E}|Y_1|^2 \le (1/\eta)^{\beta+(2-\alpha)(1-\Delta)} < (1/\eta)^{2(1-\Delta)} \\
\Rightarrow &\frac{1/\eta^{2-2\widetilde{\Delta}}}{2\big(n(\eta)-i\big)\mathbb{E}|Y_1|^2} \ge 1/\eta^{\Delta};
\end{aligned}$$

- In case that $\alpha \ge 2$, for all $\eta > 0$ that are sufficiently small,

$$2\big(n(\eta)-i\big)\mathbb{E}|Y_1|^2 < 1/\eta^{\beta+\frac{\Delta}{2}}$$

  and we know that $\beta + \frac{\Delta}{2} < 2 - 2\widetilde{\Delta}$ due to $2 - \beta > 2\Delta$ and $2\widetilde{\Delta} \le \Delta$ (see (B.34)-(B.36));

- Since $\gamma > 0$ and $2\widetilde{\Delta} \le \Delta$, we know that

$$(1-\Delta)(1-\gamma) + (1-\widetilde{\Delta}) < 2 - 2\widetilde{\Delta}.$$

Therefore, it is easy to see that the R.H.S. of (B.42) decays at a geometric rate as η tends to zero, hence $o(\eta^N)$. On the other hand,

$$\text{(II)} \le \mathbb{P}(I \ge j) \le \binom{n(\eta)}{j} \cdot \mathbb{P}\Big(|Z_i^{\le u(\eta)}| > u(\eta)^{1-\gamma} \ \forall i = 1, \dots, j\Big) \le n(\eta)^j \cdot \mathbb{P}\Big(|Z_1^{\le u(\eta)}| > u(\eta)^{1-\gamma}\Big)^j,$$

which is regularly varying w.r.t. η with index $\big(\alpha(1-\gamma)(1-\Delta) - \beta\big)j$. Therefore, for all $\eta > 0$ sufficiently small,

$$\text{(II)} \le \eta^{\big(\alpha(1-2\gamma)(1-\Delta)-\beta\big)j} < \eta^N \quad \text{due to (B.40).}$$

Collecting results above, we have established that

$$\mathbb{P}\Big(\eta\big|Z_1^{\le u(\eta)} + \cdots + Z_{n(\eta)}^{\le u(\eta)}\big| > v(\eta)\Big) = o(\eta^N).$$

The conclusion of the lemma now follows from Etemadi's theorem. □

Now consider the following setting. Let us fix some positive integer N and $\beta \in (1, 2 \wedge \alpha)$. Then we can find some positive integer j such that $(\alpha - \beta)j > N$. Meanwhile, given any $\epsilon > 0$, we will have $\epsilon - j\delta \ge \epsilon/2$ for all $\delta > 0$ sufficiently small. Therefore, by applying Lemma B.9 with $\Delta = \widetilde{\Delta} = 0$ (hence $u(\eta) = \delta/\eta,\ v(\eta) = \epsilon$) and $\beta, j, N, \epsilon, \delta$ as described here, we immediately get the following result.

Lemma B.10. *Given any $\beta \in (1, \alpha \wedge 2)$, $\epsilon > 0$, and $N > 0$, the following holds for any sufficiently small $\delta > 0$:*

$$\mathbb{P}\Big(\max_{j=1,2,\cdots,\lceil(1/\eta)^\beta\rceil} \eta|Z_1^{\le \delta/\eta} + \cdots + Z_j^{\le \delta/\eta}| > \epsilon\Big) = o(\eta^N)$$

as $\eta \downarrow 0$.

Using results and arguments above, we are able to illustrate the typical behavior of the SGD iterates X_n^η in the following two scenarios. First, we show that, when starting from most parts in the attraction field Ω, the SGD iterates X_n^η will most likely return to the neighborhood of the local minimum within a short period of time without exiting Ω. Given that there are only finitely many attraction fields on f, it is easy to see that the key technical tool Lemma 7 follows immediately from the next result.

Lemma B.11. *For sufficiently small $\epsilon > 0$, the following claim holds:*

$$\lim_{\eta\downarrow 0} \sup_{x\in\Omega:|x-s_-|\wedge|x-s_+|>\epsilon} \mathbb{P}_x\Big(X_n^\eta \in \Omega \ \forall n \le T_{return}(\eta,\epsilon), \text{and } T_{\text{return}}(\eta,\epsilon) \le \rho(\epsilon)/\eta\Big) = 1$$

where the stopping time involved is defined as

$$T_{return}(\eta,\epsilon) \triangleq \min\{n \ge 0 : X_n^\eta(x) \in [-2\epsilon, 2\epsilon]\}$$

the function $\hat{t}(\epsilon)$ is defined in (B.24)*, and the function $\rho(\cdot)$ is defined as $\rho(\epsilon) = \frac{3\bar{\epsilon}}{0.9c_-^L \wedge c_+^L} + 2\hat{t}(\epsilon)$*

Proof. Throughout, we only consider ϵ small enough so that Lemma B.6 could hold. Also, fix some $N > 0, \Delta_\alpha \in (0, \alpha - 1)$ and $\beta \in (1, \alpha)$. Let $\sigma(x, \eta) \triangleq \min\{n \geq 0 : X_n^\eta \notin \Omega\}$.

Without loss of generality, we assume $\Omega = [-L, s_+)$ and $x < 0$ (so reflection at $-L$) is a possibility. Any other case can be addressed similarly as shown below.

From Lemma B.2 and the regular varying nature of $H(\cdot)$, we have, for any $\epsilon, \delta > 0$,

$$\mathbb{P}(T_1^\eta(\delta) \leq \rho(\epsilon)/\eta) \leq \eta^{\alpha - 1 - \Delta_\alpha} \tag{B.43}$$

for any sufficiently small η.

Let $\widetilde{T}^\eta_{\text{escape}}(x)$ be the stopping time defined in Lemma B.6. From (B.30),(B.31),(B.43) and Lemma B.10, we know that

$$\begin{aligned}\sup_{x \in [-L, -L+\bar{\epsilon}]} &\mathbb{P}\Big(\widetilde{T}^\eta_{\text{escape}}(x) < \sigma(x, \eta), \widetilde{T}^\eta_{\text{escape}}(x) \leq \frac{3\bar{\epsilon}}{0.9 c_-^L \eta} \text{ and } X^\eta_{\widetilde{T}^\eta_{\text{escape}}}(x) \in [-L + \bar{\epsilon}, -L + 2\bar{\epsilon}]\Big) \\ &\geq 1 - \eta^N - \eta^{\alpha - 1 - \Delta_\alpha}\end{aligned} \tag{B.44}$$

for all sufficiently small η.

Next, we focus on $x \in \Omega$ such that $|x - s_-| \wedge |x - s_+| > \epsilon$ and $x \geq -L + \bar{\epsilon}$. We start by considering the time it took for the (deterministic) gradient descent process $\mathbf{y}_n^\eta(x)$ to return to $[-1.5\epsilon, 1.5\epsilon]$. From the definition of $\hat{t}(\epsilon)$ in (B.24) and Lemma B.4, we know that for η small enough such that $\eta \exp(2M\hat{t}(\epsilon)) < 0.5\epsilon$, we have

$$\min\{n \geq 0 : \mathbf{y}_n^\eta(x) \in [-1.5\epsilon, 1.5\epsilon]\} \leq 2\hat{t}(\epsilon)/\eta.$$

Now consider event $A(\lceil (1/\eta)^\beta \rceil, \eta, \frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}, \delta)$ (see definition in (B.28)). From Lemma B.10, we know that for any sufficiently small δ, we have

$$\mathbb{P}\Big(\big(A(\lceil (1/\eta)^\beta \rceil, \eta, \frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}, \delta)\big)^c\Big) = o(\eta^N). \tag{B.45}$$

Combining this result with (B.43)(B.45) and Lemma B.7, we get

$$\sup_{x \in \Omega : |x - s_-| \wedge |x - s_+| > \epsilon, x \geq -L + \bar{\epsilon}} \mathbb{P}_x\Big(T_{\text{return}}(\eta, \epsilon) < \sigma(x, \eta), T_{\text{return}}(\eta, \epsilon) \leq 2\hat{t}(\epsilon)/\eta\Big) \geq 1 - \eta^N - \eta^{\alpha - 1 - \Delta_\alpha} \tag{B.46}$$

for any sufficiently small η. To conclude the proof, we only to combine strong Markov property (at $\widetilde{T}^\eta_{\text{escape}}$) with bounds in (B.44)(B.46). □

In the next result, we show that, once entering a $\epsilon-$small neighborhood of the local minimum, the SGD iterates will most likely stay there until the next large jump.

Lemma B.12. *Given $N_0 > 0$, the following claim holds for any $\epsilon, \delta > 0$ that are sufficiently small:*

$$\sup_{x \in [-2\epsilon, 2\epsilon]} \mathbb{P}\Big(\exists n < T_1^\eta(\delta) \text{ s.t. } |X_n^\eta(x)| > 3\epsilon\Big) = o(\eta^{N_0})$$

as $\eta \downarrow 0$.

Proof. Fix ϵ small enough such that $3\epsilon < \epsilon_0$ (see Assumption 1 for the constant ϵ_0). Also, fix some $\Delta\alpha \in (0, 1), \beta \in (1, \alpha)$, $N > \alpha + \Delta\alpha - \beta + N_0$. Due to Lemma B.10, for any δ sufficiently small, we will have

$$\mathbb{P}\Big(\max_{j = 1, 2, \cdots, \lceil (1/\eta)^\beta \rceil} \eta |Z_1^{\leq, \delta} + \cdots + Z_j^{\leq, \delta}| > \frac{\epsilon}{\exp(2M)}\Big) = o(\eta^N). \tag{B.47}$$

Fix such $\delta > 0$. We now show that the desired claim is true for the chosen ϵ, δ.

First of all, from Lemma B.1, we know the existence of some $\theta > 0$ such that

$$\mathbb{P}(T_1^\eta(\delta) > 1/\eta^{\alpha+\Delta\alpha}) = o(\exp(-1/\eta^\theta)). \tag{B.48}$$

Next, let us zoom in on the first $\lceil (1/\eta)^\beta \rceil$ SGD iterates. For any η small enough, we will have $\mathbf{y}_n^\eta(x) \in [-2\epsilon, 2\epsilon]$ for any $n \geq 1$ and $\mathbf{y}_{\lceil (1/\eta)^\beta \rceil}^\eta(x) \in [-\epsilon, \epsilon]$ given $x \in [-2\epsilon, 2\epsilon]$. From now on we only consider such η. Due to Lemma B.7, we know that on event $\big\{ \max_{j=1,2,\cdots,\lceil (1/\eta) \rceil} \eta | Z_1^{\leq,\delta} + \cdots + Z_j^{\leq,\delta}| > \frac{\epsilon}{\exp(2M)} \big\}$, we have

$$|X_n^\eta(x)| \leq 3\epsilon \ \ \forall n \leq \lceil 1/\eta^\beta \rceil \wedge (T_1^\eta(\delta) - 1)$$

and on event $\big\{ \max_{j=1,2,\cdots,\lceil (1/\eta) \rceil} \eta | Z_1^{\leq,\delta} + \cdots + Z_j^{\leq,\delta}| > \frac{\epsilon}{\exp(2M)} \big\} \cap \{ T_1^\eta(\delta) > \lceil (1/\eta)^\beta \rceil \}$, we have $X_{T_1^\eta(\delta)}^\eta(x) \in [-2\epsilon, 2\epsilon]$. Now by repeating the same argument inductively for $\lceil 1/\eta^{\alpha+\Delta\alpha-\beta} \rceil$ times, we can show that on event

$$\{ \forall i = 1, 2, \cdots, \lceil 1/\eta^{\alpha+\Delta\alpha-\beta} \rceil, \max_{j=1,2,\cdots,\lceil (1/\eta)^\beta \rceil} \eta | Z_{i\lceil (1/\eta)^\beta \rceil + 1}^{\leq,\delta} + \cdots + Z_{i\lceil (1/\eta)^\beta \rceil + j}^{\leq,\delta}| > \frac{\epsilon}{\exp(2M)} \}, \tag{B.49}$$

we have $|X_n^\eta(x)| \leq 3\epsilon \ \forall n \leq 1/\eta^{\alpha+\Delta\alpha} \wedge (T_1^\eta(\delta) - 1)$. To conclude the proof, we only need to combine this fact with (B.47). □

The following result is an immediate product from the geometric tail bound in (B.48) and the inductive argument in (B.49).

Corollary B.13. *Given $N > 0, \widetilde{\epsilon} > 0$, the following claim holds for all sufficiently small $\delta > 0$:*

$$\mathbb{P}\Big(\max_{j=1,2,\cdots,T_1^\eta(\delta)-1} \eta | Z_1 + \cdots + Z_j| > \widetilde{\epsilon} \Big) = o(\eta^N). \tag{B.50}$$

We introduce a few concepts that will be crucial in the analysis below. Recall the definition of perturbed ODE $\widetilde{\mathbf{x}}^\eta$ in (B.18)-(B.20) (note that we will drop the notational dependency on learning rate η when we choose $\eta = 1$). Consider the definition of the following two mappings from where $\mathbf{w} = (w_1, \cdots, w_{l^*})$ is a sequence of real numbers and $\mathbf{t} = (t_1, t_2, \cdots, t_{l^*})$ with $0 = t_1 < t_2 < t_3 < \cdots$ as

$$h(\mathbf{w}, \mathbf{t}) = \widetilde{\mathbf{x}}(t_{l^*}, 0; \mathbf{t}, \mathbf{w}).$$

Next, define sets (for any $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$)

$$E(\epsilon) = \{ (\mathbf{w}, \mathbf{t}) \subseteq \mathbb{R}^{l^*} \times \Big(\mathbb{R}_+ \Big)^{l^*-1} : h(\mathbf{w}, \mathbf{t}) \notin [s_- - \epsilon, s_+ + \epsilon] \}. \tag{B.51}$$

We add a few remarks about the two types of sets defined above.

- Intuitively speaking, $E(\epsilon)$ contains all the perturbations (with times and sizes) that can send the ODE out of the current attraction field (allowing for some error with size ϵ);
- From the definition of $\bar{t}, \bar{\delta}$ in (B.21)(B.22) and Corollary B.5, one can easily see that for a fixed $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$,

$$(\mathbf{w}, \mathbf{t}) \in E(\epsilon) \Rightarrow |w_j| > \bar{\delta}, t_j - t_{j-1} \leq \bar{t} \ \ \forall j;$$

- Lastly, $E(\epsilon)$ are open sets due to $f \in C^2$.

Use $\mathbf{Leb}_+$ to denote the Lebesgue measure restricted on $[0,\infty)$, and define (Borel) measure ν_α with density on $\mathbb{R}\backslash\{0\}$:

$$\nu_\alpha(dx) = \mathbb{1}\{x>0\}\frac{\alpha p_+}{x^{\alpha+1}} + \mathbb{1}\{x<0\}\frac{\alpha p_-}{|x|^{\alpha+1}}$$

where $\alpha>1$ is the regular variation index for the distribution of Z_1 and $p_-, p_+ \in (0,1)$ are constants in Assumption 2. Now we can define a Borel measure μ on $\mathbb{R}^{l^*}\times\Big(\mathbb{R}_+\Big)^{l^*-1}$ as product measure

$$\mu = (\nu_\alpha)^{l^*}\times(\mathbf{Leb}_+)^{l^*-1}. \tag{B.52}$$

Due to remarks above, one can see that for $\epsilon\in(-\bar\epsilon,\bar\epsilon)$, we have $\mu(E(\epsilon))<\infty$. We are now ready to analyze a specific type of noise Z_n.

Definition B.1. *For any $n\geq 1$ and any $\epsilon\in(-\bar\epsilon,\bar\epsilon), \delta\in(0,b\wedge\bar\delta), \eta>0$, we say that the jump Z_n has (ϵ,δ,η)-**overflow** if*

- *$\eta|Z_n|>\delta$;*
- *In the set $\{n+1,\cdots,n+2\lceil l^*\bar t/\eta\rceil\}$, there are at least (l^*-1) elements (ordered as $n<t_2<t_3<\cdots<t_{l^*}$) such that $\eta|Z_{t_i}|>\delta$ for any $i=2,\cdots,l^*$;*
- *Let $t_1=n$ and $t_i'=t_i-t_{i-1}$ for any $i=2,\cdots,l^*$, $w_i=\eta Z_i$ for any $i=1,\cdots,l^*$, for real sequence $\boldsymbol{w}=(w_1,w_2,\cdots,w_{l^*})$ and a sequence of positive number $\boldsymbol{t}=(\eta(t_i-n))_{i=2}^{l^*}$, we have*

$$(\boldsymbol{w},\boldsymbol{t})\in E(\epsilon).$$

*Moreover, if Z_n has $(\epsilon,\delta,\eta)-$overflow, then we call $h(\boldsymbol{w},\boldsymbol{t})$ as its $(\epsilon,\delta,\eta)-$**overflow endpoint**.*

Due to the iid nature of $(Z_j)_{j\geq1}$, let us consider an iid sequence $(V_j)_{j\geq0}$ where the sequence has the same law of Z_1. Note that for any fixed $n\geq1$, the probability that Z_n has (ϵ,δ,η)-overflow is equal to the probability that V_0 has(ϵ,δ,η)-overflow. More specifically, we know that $\mathbb{P}(\eta|V_0|>\delta)=H(\delta/\eta)$, and now we focus on conditional probability admitting the following form:

$$p(\epsilon,\delta,\eta)=\mathbb{P}\Big(V_0 \text{ has } (\epsilon,\delta,\eta)\text{-overflow}\ \Big|\ \eta|V_0|>\delta\Big). \tag{B.53}$$

For any open interval $A=(a_1,a_2)$ such that $A\cap[s_-+\bar\epsilon,s_+-\bar\epsilon]=\emptyset$, we also define

$$p(\epsilon,\delta,\eta;A)=\mathbb{P}\Big(V_0 \text{ has } (\epsilon,\delta,\eta)\text{-overflow and the endpoint is in } A\ \Big|\ \eta|V_0|>\delta\Big). \tag{B.54}$$

Lemma B.14. *For any $\epsilon\in(-\bar\epsilon,\bar\epsilon), \delta\in(0,b\wedge\bar\delta)$, and any open interval $A=(a_1,a_2)$ such that $|a_1|\wedge|a_2|>r-\bar\epsilon$ and $|a_1|\neq L, |a_2|\neq L$, we have*

$$\lim_{\eta\downarrow0}\frac{p(\epsilon,\delta,\eta;A)}{\delta^\alpha\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}}=\mu\big(E(\epsilon)\cap h^{-1}(A)\big)$$

where μ is the measure defined in (B.52)*, and $p(\cdot,\cdot,\cdot;A)$ is the conditional probability defined in* (B.54)*.*

Proof. Let us start by fixing some notations. Let $T_1=0$, and define stopping times $T_j=\min\{n>T_{j-1}:\eta|V_n|>\delta\}$ and inter-arrival times $T_j'=T_j-T_{j-1}$ for any $j\geq1$, and large jump $W_j=V_{T_j}$ for any $j\geq0$. Note that: first of all, the pair (T_i',W_i) is independent of (T_j',W_j) whenever $i\neq j$; besides, W_j and T_j' are independent for all $j\geq1$.

Define the following sequence (of random elements) $\mathbf{w}=(w_1,\cdots,w_{l^*})$ and $\mathbf{t}=(t_1,\cdots,t_{l^*})$ by

$$w_j=\eta W_j,\quad t_j=\eta T_j.$$

If V_0 has (ϵ,δ,η)-overflow, then the following two events must occur:

- $T_j' \le 2\bar{t}/\eta$ for any $j = 2, \cdots, l^*$;
- $\eta|W_j| > \bar{\delta}$ for any $j = 1, 2, \cdots, l^*$;
- $(\mathbf{w}, \mathbf{t}) \in E(\epsilon)$

Therefore, for sufficiently small η, we now have

$$\begin{aligned}
&p(\epsilon, \delta, \eta) \\
=&\Big(\mathbb{P}(T_1' \le 2\bar{t}/\eta)\Big)^{l^*-1} \cdot \int \mathbb{1}\Big\{(\mathbf{w}, \mathbf{t}) \in E(\epsilon)\Big\} \\
&\qquad \cdot \mathbb{P}(\eta W_1 = dw_1 | \eta|W_1| > \delta) \cdots \mathbb{P}(\eta W_{l^*} = dw_{l^*} |\ \eta|W_{l^*}| > \delta) \\
&\qquad \cdot \mathbb{P}(\eta T_2' = dt_2' | \eta T_2' \le 2\bar{t}) \cdots \mathbb{P}(\eta T_{l^*}' = dt_{l^*}' | \eta T_{l^*}' \le 2\bar{t}) \\
=&\Big(\mathbb{P}(T_1' \le 2\bar{t}/\eta)\Big)^{l^*-1} \cdot \mathbb{Q}_{\eta,\delta}\big(E(\epsilon) \cap h^{-1}(A)\big)
\end{aligned} \tag{B.55}$$

where $\mathbb{Q}_{\eta,\delta}$ is the Borel-measurable probability measure on $\mathbb{R}^{l^*} \times \Big(\mathbb{R}_+\Big)^{l^*-1}$ induced by a sequence of independent random variables $(W_1^{\uparrow}(\eta,\delta), \cdots, W_{l^*}^{\uparrow}(\eta,\delta), T_2^{\uparrow}(\eta,\delta), \cdots, T_{l^*}^{\uparrow}(\eta,\delta))$ such that

- For any $i = 1, \cdots, l^*$, the distribution of $W_i^{\uparrow}(\eta,\delta)$ follows from $\mathbb{P}\Big(\eta W_1 \in \cdot \ \Big|\ \eta|W_{l^*}| > \delta\Big)$;
- For any $i = 2, \cdots, l^*$, the distribution of $T_i^{\uparrow}(\eta,\delta)$ follows from $\mathbb{P}\Big(\eta T_1 \in \cdot \ \Big|\ \eta T_1 \le 2\bar{t}\Big)$;
- $\mathbb{Q}_{\eta,\delta}(\cdot) = \mathbb{P}\Big((\eta W_1^{\uparrow}(\eta,\delta), \cdots, \eta W_{l^*}^{\uparrow}(\eta,\delta), \eta T_2^{\uparrow}(\eta,\delta), \cdots, \eta \sum_{j=2}^{l^*} T_j^{\uparrow}(\eta,\delta)) \in \cdot\Big)$.

Now we study the weak convergence of $W_1^{\uparrow}, T_1^{\uparrow}$:

- Due to the regularly varying nature of distribution of Z_1 (hence for W_1), we know that: for any $x > \delta$,

  $$\lim_{\eta\downarrow 0} \mathbb{P}\Big(\eta W_1 > x \ \Big|\ \eta|W_{l^*}| > \delta\Big) = p_+ \frac{\delta^\alpha}{x^\alpha}, \quad \lim_{\eta\downarrow 0} \mathbb{P}\Big(\eta W_1 < -x \ \Big|\ \eta|W_{l^*}| > \delta\Big) = p_- \frac{\delta^\alpha}{x^\alpha};$$

  therefore, $W_1^{\uparrow}(\eta,\delta)$ weakly converges to a (randomly signed) Pareto RV that admits the density

  $$\nu_{\alpha,\delta}(dx) = \mathbb{1}\{x > 0\} p_+ \frac{\alpha\delta^\alpha}{x^{\alpha+1}} + \mathbb{1}\{x < 0\} p_- \frac{\alpha\delta^\alpha}{|x|^{\alpha+1}}$$

  as $\eta \downarrow 0$;
- For any $x \in [0, 2\bar{t}]$, since $\lim_{\eta\downarrow 0} \lfloor x/\eta \rfloor H(\delta/\eta) = 0$, it is easy to show that

  $$\lim_{\eta\downarrow 0} \frac{1 - (1 - H(\delta/\eta))^{\lfloor x/\eta \rfloor}}{\lfloor x/\eta \rfloor H(\delta/\eta)} = 1;$$

  therefore, we have (for any $x \in (0, 2\bar{t}]$)

  $$\mathbb{P}(\eta T_1 \le x \mid \eta T_1 \le 2\bar{t}) = \frac{1 - (1 - H(\delta/\eta))^{\lfloor x/\eta \rfloor}}{1 - (1 - H(\delta/\eta))^{\lfloor 2\bar{t}/\eta \rfloor}} \to \frac{x}{2\bar{t}}$$

  as $\eta \downarrow 0$, which implies that $T_1^{\uparrow}$ converges weakly to a uniform RV on $[0, 2\bar{t}]$.

Let us denote the weak limit of measure $\mathbb{Q}_{\eta,\delta}$ as $\mu_{\delta,2\bar{t}}$. In the discussion before the Lemma we have shown that, for any $(\mathbf{w},\mathbf{t}) \in E(\epsilon)$ (with $\delta \in (0,\bar{\delta})$), we have $|w_i| \geq \bar{\delta}$ and $|t_i'| \leq 2\bar{t}$; since we require $\delta < \bar{\delta}$, by definition of measures μ and $\mu_{\delta,2\bar{t}}$ we have

$$\mu_{\delta,2\bar{t}}\big(E(\epsilon)\cap h^{-1}(A)\big) = \frac{\delta^{\alpha l^*}}{(2\bar{t})^{l^*-1}} \cdot \mu\big(E(\epsilon)\cap h^{-1}(A)\big).$$

For simplicity of notations, we let $E(\epsilon,A) \triangleq E(\epsilon)\cap h^{-1}(A)$. By definition of the set $E(\epsilon)$, we have (recall that A is an open interval (a_1,a_2) that does not overlap with $[s_- + \bar{\epsilon}, s_+ - \bar{\epsilon}]$)

$$\begin{aligned} E(\epsilon,A) &= h^{-1}\big((-\infty,s_- - \epsilon)\cup(s_+ + \epsilon,\infty)\big)\cap h^{-1}\big((a_1,a_2)\big) \\ &= h^{-1}\Big(\big((-\infty,s_- - \epsilon)\cup(s_+ + \epsilon,\infty)\big)\cap(a_1,a_2)\Big) \\ &= h^{-1}\big(F(\epsilon,a_1,a_2)\big) \end{aligned}$$

where $F(\epsilon,a_1,a_2) \triangleq \big((-\infty,s_- - \epsilon)\cup(s_+ + \epsilon,\infty)\big)\cap(a_1,a_2)$. Meanwhile, it is easy to see that h is a continuous mapping, hence

$$(\mathbf{w},\mathbf{t}) \in \partial E(\epsilon,A) \Rightarrow h(\mathbf{w},\mathbf{t}) \in \{s_- + \epsilon, s_+ - \epsilon, a_1, a_2\}.$$

Fix some s with $s \neq \pm L, |s| > (l^*-1)b + \bar{\epsilon}$. For any fixed real numbers $t_2,\cdots,t_{l^*-1}$, $w_1,\cdots,w_{l^*}$, if $h(w_1,\cdots,w_{l^*},t_2,\cdots,t_{l^*-1},t) = s$, then since $\widetilde{\mathbf{x}}(t_{l^*}-1,0;w_1,\cdots,w_{l^*},t_2,\cdots,t_{l^*-1},t) \in [s-b,s+b]$, due to Assumption 1 (in particular, there is no point x on this interval with $|f'(x)| \leq c_0$), there exists at most one possible t that makes $h(w_1,\cdots,w_{l^*},t_2,\cdots,t_{l^*-1},t) = s$. Therefore, let W_j^* be iid RVs from law $\nu_{\alpha,\delta}$ defined above, and $(T_j^{*,\prime})_{j\geq 2}$ be iid RVs from Unif$[0,2\bar{t}]$, $T_0^* = 0, T_k^* = \sum_{j=2}^k T_j^{*,\prime}$. By conditioning on all W_j^* and all $T_2^{*,\prime},\cdots,T_{l^*-1}^{*,\prime}$, we must have

$$\mathbb{P}\Big(h(W_1^*,\cdots,W_j^*,T_2^*,\cdots,T_{l^*}^*) = s \;\Big|\; W_1^* = dw_1,\cdots,W_{l^*}^* = dw_{l^*}, T_2^{*,\prime} = dt_2,\cdots,T_{l^*-1}^{*,\prime} = dt_{l^*-1}\Big) = 0 \tag{B.56}$$

which implies

$$\mathbb{P}\Big(h(W_1^*,\cdots,W_j^*,T_2^*,\cdots,T_{l^*}^*) = s\Big) = 0$$

hence

$$\mu\Big(\partial E(\epsilon,A)\Big) = 0.$$

By Portmanteau theorem (see Theorem 2.1 of [1]) we have

$$\lim_{\eta\downarrow 0} \mathbb{Q}_{\eta,\delta}(E(\epsilon,A)) = \mu_{\delta,2\bar{t}}(E(\epsilon,A)).$$

Collecting the results we have and using (B.55), we can see that

$$\begin{aligned} \limsup_{\eta\downarrow 0} \frac{p(\epsilon,\delta,\eta;A)}{\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}\delta^\alpha} &= \limsup_{\eta\downarrow 0} \frac{(2\bar{t})^{l^*-1}\cdot p(\epsilon,\delta,\eta;A)}{\delta^{\alpha l^*}\cdot\Big(\mathbb{P}(T_1'\leq 2\bar{t}/\eta)\Big)^{l^*-1}} \cdot \Big(\frac{\delta^\alpha}{2\bar{t}}\cdot\frac{\mathbb{P}(T_1'\leq 2\bar{t}/\eta)}{H(1/\eta)/\eta}\Big)^{l^*-1} \\ &\leq \limsup_{\eta\downarrow 0} \frac{(2\bar{t})^{l^*-1}\cdot p(\epsilon,\delta,\eta;A)}{\delta^{\alpha l^*}\cdot\Big(\mathbb{P}(T_1'\leq 2\bar{t}/\eta)\Big)^{l^*-1}} \cdot \limsup_{\eta\downarrow 0}\Big(\frac{\delta^\alpha}{2\bar{t}}\cdot\frac{\mathbb{P}(T_1'\leq 2\bar{t}/\eta)}{H(1/\eta)/\eta}\Big)^{l^*-1} \\ &\leq \mu(E(\epsilon,A))\cdot\limsup_{\eta\downarrow 0}\Big(\frac{\delta^\alpha}{2\bar{t}}\cdot\frac{\mathbb{P}(T_1'\leq 2\bar{t}/\eta)}{H(1/\eta)/\eta}\Big)^{l^*-1}. \end{aligned}$$

Fix some $\kappa > 1$. From Lemma B.2 and the regularly varying nature of function H, we get

$$\limsup_{\eta\downarrow 0}\Big(\frac{\delta^\alpha}{2\bar{t}}\cdot\frac{\mathbb{P}(T_1'\le 2\bar{t}/\eta)}{H(1/\eta)/\eta}\Big)^{l^*-1}\le\kappa^{l^*-1}\limsup_{\eta\downarrow 0}\Big(\frac{\delta^\alpha}{2\bar{t}}\cdot\frac{2\bar{t}H(\delta/\eta)/\eta}{H(1/\eta)/\eta}\Big)^{l^*-1}=\kappa^{l^*-1}.$$

Due to the arbitrariness of $\kappa > 1$, we have established that

$$\limsup_{\eta\downarrow 0}\frac{p(\epsilon,\delta,\eta;A)}{\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}\delta^\alpha}\le\mu(E(\epsilon)).$$

The lower bound can be shown by an argument symmetric to the one for upper bound. □

The following result is an immediate corollary of Lemma B.14.

Corollary B.15. *For any $\epsilon\in(-\bar{\epsilon},\bar{\epsilon}),\delta\in(0,b\wedge\bar{\delta})$, we have*

$$\lim_{\eta\downarrow 0}\frac{p(\epsilon,\delta,\eta)}{\delta^\alpha\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}}=\mu\big(E(\epsilon)\big)$$

where μ is the measure defined in (B.52)*, and $p(\cdot,\cdot,\cdot)$ is the conditional probability defined in* (B.53)*.*

Define the following stopping times:

$$\sigma(\eta)=\min\{n\ge 0: X_n^\eta\notin\Omega\};\tag{B.57}$$
$$R(\epsilon,\delta,\eta)=\min\{n\ge T_1^\eta(\delta): X_n^\eta\in[-2\epsilon,2\epsilon]\}.\tag{B.58}$$

σ indicate the time that the iterates escape the current attraction field, while R denotes the time the SGD iterates return to a small neighborhood of the local minimum after first exit from this small neighborhood. In the next few results, we study the probability of several atypical scenarios when SGD iterates make attempts to escape Ω or return to local minimum after the attempt fails. First, we show that, when starting from the local minimum, it is very unlikely to escape with less than l^* big jumps.

Lemma B.16. *Given $\epsilon\in(0,\bar{\epsilon}),N>0$, the following claim holds for any sufficiently small $\delta>0$:*

$$\sup_{x\in[-2\epsilon,2\epsilon]}\mathbb{P}_x\Big(\sigma(\eta)<R(\epsilon,\eta),\ \sigma(\eta)<T_{l^*}^\eta(\delta)\Big)=o(\eta^N)$$

as $\eta\downarrow 0$.

Proof. Based on the given $\epsilon>0$, fix some $\widetilde{\epsilon}=\frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}$. Recall the definition of $\hat{t}(\epsilon)$ in (B.24).

First, using Lemma B.12, we know that for sufficiently small δ, we have

$$\sup_{x\in[-2\epsilon,2\epsilon]}\mathbb{P}\Big(A_1^\times(\epsilon,\delta,\eta)\Big)=o(\eta^N)\tag{B.59}$$

where

$$A_1^\times(\epsilon,\delta,\eta)=\Big\{\exists n<T_1^\eta(\delta)\ s.t.\ |X_n^\eta(x)|>3\epsilon\Big\}.$$

Define event

$$A_2^\times(\widetilde{\epsilon},\delta,\eta)\triangleq\Big\{\exists j=2,\cdots,l^*\ s.t.\max_{k=1,2,\cdots,T_j^\eta(\delta)-T_{j-1}^\eta(\delta)-1}\eta|Z_{T_{j-1}^\eta(\delta)+1}+\cdots+Z_{T_{j-1}^\eta(\delta)+j}|>\widetilde{\epsilon}\Big\}.$$

From Corollary B.13, we know that for sufficiently small $\delta > 0$,

$$\mathbb{P}\Big(A_2^{\times}(\widetilde{\epsilon}, \delta, \eta)\Big) = o(\eta^N). \tag{B.60}$$

From now on, we only consider such δ that (B.59)(B.60) hold.

On event $\Big(A_1^{\times} \cup A_2^{\times}\Big)^c \cap \{\sigma(\eta) < R(\epsilon, \eta)\} \cap \{\sigma(\eta) > T_1^{\eta}(\delta)\}$, we must have $\sigma(\eta) > T_1^{\eta}(\delta)$ and

$$T_2^{\eta}(\delta) \wedge \sigma(\eta) - T_1^{\eta}(\delta) < 2\hat{t}(\epsilon)/\eta.$$

Otherwise, due to Lemma B.4 and B.7, we know that at step $\widetilde{t} = T_1^{\eta}(\delta) + \lfloor \hat{t}(\epsilon)/\eta \rfloor$, we have

$$|X_{\widetilde{t}}^{\eta}| < 2\epsilon, \text{ and } |X_n^{\eta}| \leq \bar{\epsilon} \ \ \forall n \leq \widetilde{t}$$

for any sufficiently small η. By repeating this argument inductively, we obtain the following result: define

$$J = \min\{j = 1, 2, \cdots : \sigma(\eta) \in [T_j^{\eta}(\delta), T_{j+1}^{\eta}(\delta))\},$$

then on event $\Big(A_1^{\times} \cup A_2^{\times}\Big)^c \cap \{\sigma(\eta) < R(\epsilon, \eta),\ \sigma(\eta) < T_{l^*}^{\eta}(\delta)\}$, we must have

$$T_j^{\eta}(\delta) \wedge \sigma(\eta) - T_{j-1}^{\eta}(\delta) \wedge \sigma(\eta) < 2\hat{t}(\epsilon)/\eta \ \ \forall j = 2, 3, \cdots, J. \tag{B.61}$$

Furthermore, using this bound and Lemma B.7, we know that on event $\Big(A_1^{\times} \cup A_2^{\times}\Big)^c \cap \{\sigma(\eta) < R(\epsilon, \eta),\ \sigma(\eta) < T_{l^*}^{\eta}(\delta)\}$,

- $|X_{T_j^{\eta}(\delta)}^{\eta}| \leq |X_{T_{j-1}^{\eta}(\delta)}^{\eta}| + b + \epsilon + \bar{\epsilon}$ for all $j = 2, 3, J-1$,
- $|X_{\sigma(\eta)}^{\eta}| \leq |X_{T_{J-1}^{\eta}(\delta)}^{\eta}| + \epsilon + \bar{\epsilon}$

However, this implies

$$|X_{\sigma(\eta)}^{\eta}| \leq l^*(\bar{\epsilon} + \epsilon) + (l^* - 1)b < r$$

and contradicts the definition of $\sigma(\eta)$. In summary,

$$\sup_{x \in [-2\epsilon, 2\epsilon]} \mathbb{P}_x\Big(\sigma(\eta) < R(\epsilon, \eta),\ \sigma(\eta) < T_{l^*}^{\eta}(\delta)\Big) \leq \mathbb{P}\Big(A_1^{\times}(\epsilon, \delta, \eta) \cup A_2^{\times}(\widetilde{\epsilon}, \delta, \eta)\Big) = o(\eta^N).$$

□

The following two results follow immediately from the proof above, especially the inductive argument leading to bound (B.61), and we state them without repeating the deatils of the proof.

Corollary B.17. *Given $\epsilon \in (0, \bar{\epsilon}), N > 0$, the following claim holds for any sufficiently small $\delta > 0$:*

$$\sup_{x \in [-2\epsilon, 2\epsilon]} \mathbb{P}_x\Big(T_{l^*}^{\eta}(\delta) \leq \sigma(\eta) \wedge R(\epsilon, \eta),\ \textit{and } \exists j = 2, 3, \cdots, l^* \textit{ s.t. } T_j^{\eta}(\delta) - T_{j-1}^{\eta}(\delta) > 2\hat{t}(\epsilon)/\eta\Big) = o(\eta^N)$$

as $\eta \downarrow 0$.

Corollary B.18. *Given $\epsilon \in (0, \bar{\epsilon}), N > 0$, the following claim holds for any sufficiently small $\delta > 0$:*

$$\sup_{x \in [-2\epsilon, 2\epsilon]} \mathbb{P}_x\Big(R(\epsilon, \eta) < T_{l^*}^{\eta}(\delta) \wedge \sigma(\eta),\ R(\epsilon, \eta) - T_1^{\eta}(\delta) > 2l^*\hat{t}(\epsilon)/\eta\Big) = o(\eta^N)$$

as $\eta \downarrow 0$.

In the next result, we show that, if the inter-arrival time between some large jumps are too long, or some large jumps are still not *large enough*, then it is very unlikely that the SGD iterates could escape at the time of l^*-th large jump (or even get close enough to the boundary of the attraction field).

Lemma B.19. *Given $\epsilon \in (0, \bar{\epsilon})$ and any $N > 0$, the following holds for all $\delta > 0$ that are sufficiently small:*

$$\sup_{x \in [-2\epsilon, 2\epsilon]} \mathbb{P}_x(B_2^{\times}(\epsilon, \delta, \eta)) = o(\eta^N).$$

where

$$B_2^{\times}(\epsilon, \delta, \eta) = \{T_{l^*}^{\eta}(\delta) \leq \sigma(\eta) \wedge R(\epsilon, \eta)\} \cap \Big\{\exists j = 2, 3, \cdots, l^* \text{ s.t. } T_j^{\eta}(\delta) - T_{j-1}^{\eta}(\delta) > \bar{t}/\eta \\ \text{or } \exists j = 1, 2, \cdots, l^* \text{ s.t. } \eta |W_j^{\eta}(\delta)| \leq \bar{\delta}\Big\} \cap \{|X_{T_{l^*}^{\eta}}^{\eta}| \geq r - \bar{\epsilon}\}.$$

Proof. Let $A_1^{\times}, A_2^{\times}$ be the events defined in the proof of Lemma B.16. Based on the given $\epsilon > 0$, fix some $\widetilde{\epsilon} = \frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}$.

Let $J = \min\{j = 2, 3, \cdots : T_j^{\eta}(\delta) - T_{j-1}^{\eta}(\delta) > \bar{t}/\eta\}$. On event $\Big(A_1^{\times}(\epsilon, \delta, \eta) \cup A_2^{\times}(\widetilde{\epsilon}, \delta, \eta)\Big)^c \cap \{J \leq l^*\}$, from Lemma B.4 and B.7 and the definition of constant $\bar{t}$, we know that

- $|X_{T_j^{\eta}(\delta)}^{\eta}| \leq |X_{T_{j-1}^{\eta}(\delta)}^{\eta}| + b + \epsilon + \bar{\epsilon}$ for all $j = 2, 3, J-1$;
- $|X_{T_J^{\eta}(\delta)}^{\eta}| \leq 2\bar{\epsilon}$
- $|X_n^{\eta}| < s - \bar{\epsilon} \ \ \forall n \leq T_J^{\eta}(\delta)$

Now starting from step $T_J^{\eta}(\delta)$, by using Lemma B.4 and B.7 again one can see that

- $|X_{T_j^{\eta}(\delta)}^{\eta}| \leq |X_{T_{j-1}^{\eta}(\delta)}^{\eta}| + b + \epsilon + \bar{\epsilon}$ for all $j = J+1, \cdots, l^*$.

Combining these results, we have that $|X_{T_{l^*}^{\eta}}^{\eta}| < r - \bar{\epsilon}$ on event $\Big(A_1^{\times}(\epsilon, \delta, \eta) \cup A_2^{\times}(\widetilde{\epsilon}, \delta, \eta)\Big)^c \cap \{J \leq l^*\}$.

Next, define $J' = \min\{j = 1, 2, \cdots; \eta |W_j^{\eta}(\delta)| \leq \bar{\delta}\}$. Similarly, on event $\Big(A_1^{\times}(\epsilon, \delta, \eta) \cup A_2^{\times}(\widetilde{\epsilon}, \delta, \eta)\Big)^c \cap \{J > l^*\} \cap \{J' \leq l^*\}$, using Lemma B.4 and B.7 again one can see that

- $|X_{T_j^{\eta}(\delta)}^{\eta}| \leq |X_{T_{j-1}^{\eta}(\delta)}^{\eta}| + b + \epsilon + \bar{\epsilon}$ for all $j = 1, 2, \cdots, l^*, j \neq J'$;
- $|X_{T_J^{\eta}(\delta)}^{\eta}| \leq |X_{T_{J-1}^{\eta}(\delta)}^{\eta}| + \bar{\delta} + \epsilon + \bar{\epsilon}$ for all $j = 1, 2, \cdots, l^*, j \neq J'$.

Since $\bar{\delta} \in (0, \bar{\epsilon})$, we have $|X_{T_{l^*}^{\eta}}^{\eta}| < r - \bar{\epsilon}$ on this event.

In summary, the following bound

$$\sup_{x \in [-2\epsilon, 2\epsilon]} \mathbb{P}_x(B_2^{\times}) \leq \mathbb{P}(A_1^{\times}(\epsilon, \delta, \eta) \cup A_2^{\times}(\widetilde{\epsilon}, \delta, \eta)) = o(\eta^N)$$

holds for any δ that is sufficiently small, which is established in Lemma B.16. This conclude the proof. □

In the next lemma, we show that, starting from the local minimum, it is unlikely that the SGD iterates will be right at the boundary of the attraction field after l^* large jumps. Recall that there are $n_{\min}$ attraction fields on f, and excluding $s_0 = -\infty, s_{n_{\min}} = \infty$ the remaining points $s_1, \cdots, s_{n_{\min}-1}$ are the boundaries of the attraction fields.

Lemma B.20. *There exists a function $\Psi(\cdot):\mathbb{R}^+\mapsto\mathbb{R}^+$ satisfying $\lim_{\epsilon\downarrow 0}\Psi(\epsilon)=0$ such that the following claim folds. Given $\epsilon\in\big(0,\bar{\epsilon}/(3\bar{\rho}+3\widetilde{\rho}+9)\big)$, we have*

$$\limsup_{\eta\downarrow 0}\frac{\sup_{x\in[-2\epsilon,2\epsilon]}\mathbb{P}_x(B_3^{\times}(\epsilon,\delta,\eta))}{\Big(H(1/\eta)/\eta\Big)^{l^*-1}}\leq\delta^{\alpha}\Psi(\epsilon)$$

for all δ sufficiently small, where $\bar{\rho}$ and $\widetilde{\rho}$ are the constants defined in Corollary B.5 and B.8, and the event is defined as

$$B_3^{\times}(\epsilon,\delta,\eta)=\Big\{T_{l^*}^{\eta}(\delta)\leq\sigma(\eta)\wedge R(\epsilon,\eta)\Big\}\cap\Big\{\exists k\in[n_{min}-1]\text{ such that }X^{\eta}_{T^{\eta}_{l^*}(\delta)}\in[s_k-\epsilon,s_k+\epsilon]\Big\}.$$

Proof. Let $A_1^{\times},A_2^{\times}$ be the events defined in the proof of Lemma B.16. Based on the given $\epsilon>0$, fix some $\widetilde{\epsilon}=\frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}$. Fix some $N>\alpha l^*$.

Choose δ small enough so that claim in Lemma B.19 holds for the ϵ prescribed. Using the same arguments in Lemma B.19, we have the following inclusion of events:

$$\begin{aligned}&B_3^{\times}(\epsilon,\delta,\eta)\cap\Big(A_1^{\times}(\epsilon,\delta,\eta)\cup A_2^{\times}(\widetilde{\epsilon},\delta,\eta)\Big)^c\\ \subseteq&\Big\{\forall j=2,3,\cdots,l^*,\ T_j^{\eta}(\delta)-T_{j-1}^{\eta}(\delta)\leq\bar{t}/\eta\Big\}\cap\Big\{\forall j=1,2,3,\cdots,l^*,\ \eta|W_1^{\eta}(\delta)|>\bar{\delta}\Big\}.\end{aligned}$$

Therefore, on event $B_3^{\times}(\epsilon,\delta,\eta)\cap\Big(A_1^{\times}(\epsilon,\delta,\eta)\cup A_2^{\times}(\widetilde{\epsilon},\delta,\eta)\Big)^c$, we can apply Corollary B.5 and B.8 and conclude that $Z_{T_1^{\eta}(\delta)}$ has $(-\bar{\epsilon},\delta,\eta)-$overflow, and its $(-\bar{\epsilon},\delta,\eta)-$overflow endpoint lies

$$(s_k-3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,s_k+3(\bar{\rho}+\widetilde{\rho}+3)\epsilon)$$

for some $k\in[n_{\min}-1]$. Using Lemma B.14 and Corollary B.15, we have that (for any sufficiently small η)

$$\begin{aligned}&\mathbb{P}\Big(B_3^{\times}(\epsilon,\delta,\eta)\cap\Big(A_1^{\times}(\epsilon,\delta,\eta)\cup A_2^{\times}(\widetilde{\epsilon},\delta,\eta)\Big)^c\Big)\\ \leq&\delta^{\alpha}\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}\cdot\sum_{k=1}^{n_{\min}-1}\mu\big(E(-\bar{\epsilon})\cap h^{-1}\Big((s_k-\hat{\epsilon},s_k+\hat{\epsilon})\Big)\Big)\end{aligned}$$

where $\hat{\epsilon}=3(\bar{\rho}+\widetilde{\rho}+3)\epsilon$. Besides, as established in the proof of Lemma B.16, we have

$$\mathbb{P}\Big(A_1^{\times}(\epsilon,\delta,\eta)\cup A_2^{\times}(\widetilde{\epsilon},\delta,\eta)\Big)=o(\eta^N)$$

for all sufficiently small δ. In conclusion, we only need to choose

$$\Psi(\epsilon)=\sum_{k=1}^{n_{\min}-1}\mu\Big(E(-\bar{\epsilon})\cap h^{-1}\big((s_k-3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,s_k+3(\bar{\rho}+\widetilde{\rho}+3)\epsilon)\big)\Big).$$

To conclude the proof, just note that by combining the continuity of measure with the conditional probability argument leading to (B.56), we can show that $\lim_{\epsilon\downarrow 0}\Psi(\epsilon)=0$. □

Lastly, we establish the lower bound for the probability of the *most likely* way for SGD iterates to exit the current attraction field: making l^* large jumps in a relatively short period of time. Recall that $\bar{\epsilon}$ is the fixed constant in (B.13)-(B.16).

Lemma B.21. *Given $\epsilon \in (0, \bar{\epsilon}/3)$, it holds for any sufficiently small $\delta > 0$ such that*

$$\liminf_{\eta \downarrow 0} \frac{\inf_{|x| \le 2\epsilon} \mathbb{P}_x(A^\circ(\epsilon, \delta, \eta))}{\big(H(1/\eta)/\eta\big)^{l^*-1}} \ge c_* \delta^\alpha$$

where the event is defined as

$$A^\circ(\epsilon, \delta, \eta) \triangleq \Big\{\sigma(\eta) < R(\epsilon, \eta),\ \sigma(\eta) = T^\eta_{l^*}(\delta),\ X^\eta_{T^\eta_{l^*}} \notin [s_- - \epsilon, s_+ + \epsilon]\Big\}$$
$$\cap \Big\{T^\eta_j(\delta) - T^\eta_{j-1}(\delta) \le \frac{\bar{\epsilon}}{2M}\lceil 1/\eta \rceil\ \forall j = 2, 3, \cdots, l^*\Big\}$$

and the constant

$$c_* = \frac{1}{2}(\frac{1}{2b})^{l^*\alpha}(\frac{\bar{\epsilon}}{4M})^{l^*-1}$$

is strictly positive and does not vary with ϵ, δ.

Proof. Let $A^\times_1, A^\times_2$ be the events defined in the proof of Lemma B.16. Fix some N such that $N > \alpha l^*$. Based on the given $\epsilon > 0$, fix some $\widetilde{\epsilon} = \frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}$. We only consider $\delta < M$. Furthermore, choose δ small enough so that (B.59) and (B.60) hold for the chosen N and ϵ. Also, we only consider η small enough so that $\eta M < b \wedge \bar{\epsilon}$.

Due to (B.13)-(B.16), we can, without loss of generality, assume that $r = s_+$, and in this case we will have

$$l^* b - 100 l^* \bar{\epsilon} > s_+ + 100 l^* \bar{\epsilon}.$$

Under this assumption, we will now focus on providing a lower bound for the following event that describes the exit from the right side of Ω (in other words, by crossing s_+)

$$A^\circ_\rightarrow(\epsilon, \delta, \eta) \triangleq \Big\{\sigma(\eta) < R(\epsilon, \eta),\ \sigma(\eta) = T^\eta_{l^*}(\delta),\ X^\eta_{T^\eta_{l^*}} > s_+ + \epsilon\Big\}$$
$$\cap \Big\{T^\eta_j(\delta) - T^\eta_{j-1}(\delta) \le \frac{\bar{\epsilon}}{2M}\lceil 1/\eta \rceil\ \forall j = 2, 3, \cdots, l^*\Big\}.$$

First, define event

$$A^\circ_3(\delta, \eta) = \Big\{W^\eta_j(\delta) \ge 2b\ \forall j = 1, \cdots, l^*,\ T^\eta_j(\delta) - T^\eta_{j-1}(\delta) \le \frac{\bar{\epsilon}}{2M}\lceil 1/\eta \rceil\ \forall j = 2, \cdots, l^*\Big\},$$

and observe some facts on event $A^\circ_3(\delta, \eta) \cap \Big(A^\times_1(\epsilon, \delta, \eta) \cup A^\times_2(\widetilde{\epsilon}, \delta, \eta)\Big)^c$.

- $|X^\eta_k| \le 3\epsilon \forall n < T^\eta_1(\delta)$; (due to $A^\times_1$ not occurring)
- $X^\eta_{T^\eta_1(\delta)} \in [b - 3\epsilon, b + 3\epsilon]$; (due to $W^\eta_1 \ge 2b$ and the effect of gradient clipping at step T^η_1, as well as the fact that $X^\eta_{T^\eta_1 - 1} \in [-3\epsilon, 3\epsilon]$ from the previous bullet point)
- Due to $|f'(\cdot)| \le M$ and $\delta < M$, one can see that (for any $n \ge 1$)

  $$\sup_{x \in [-L, L]} |\eta f'(x)| + |\eta Z^{\le \delta, \eta}_n| \le 2\eta M;$$

  this provides an upper bound for the change in SGD iterates at each step, and gives us

  $$X^\eta_n \in [b - 3\epsilon - \bar{\epsilon}, b + 3\epsilon + \bar{\epsilon}]\ \ \forall T^\eta_1(\delta) \le n < T^\eta_2(\delta)$$

  where we also used $T^\eta_2(\delta) - T^\eta_1(\delta) \le \frac{\bar{\epsilon}}{2M}\lceil 1/\eta \rceil$

- Therefore, at the arrival time of the second large jump, we must have $X^\eta_{T^\eta_2(\delta)} \geq 2b - 3\epsilon - \bar{\epsilon}$;
- By repeating the argument above inductively, we can show that (for all $j = 1, 2, \cdots, l^*$)

$$X^\eta_n \in [(j-1)b - 3\epsilon - (j-1)\bar{\epsilon}, (j-1)b + 3\epsilon + (j-1)\bar{\epsilon}] \quad \forall T^\eta_{j-1} \leq n < T^\eta_j$$
$$X^\eta_{T^\eta_j} \in [jb - 3\epsilon - (j-1)\bar{\epsilon}, jb + 3\epsilon + (j-1)\bar{\epsilon}];$$

  In particular, we know that $X^\eta_n \in \Omega$ for any $n < T^\eta_{l^*}$ (so the exit does not occur before $T^\eta_{l^*}$), and at the arrival of the l^*-th large jump, we have (using $3\epsilon < \bar{\epsilon}$)

$$X^\eta_{T^\eta_{l^*}(\delta)} \geq l^* b - l^* \bar{\epsilon} > s_+ + \epsilon.$$

In summary, we have shown that

$$A^\circ_3(\delta, \eta) \cap \Big(A^\times_1(\epsilon, \delta, \eta) \cup A^\times_2(\widetilde{\epsilon}, \delta, \eta) \Big)^c \subseteq A^\circ_{\to}(\epsilon, \delta, \eta).$$

To conclude the proof, just notice that (for sufficiently small η)

$$\begin{aligned}
&\mathbb{P}\Big(A^\circ_3(\delta, \eta) \cap \Big(A^\times_1(\epsilon, \delta, \eta) \cup A^\times_2(\widetilde{\epsilon}, \delta, \eta) \Big)^c \Big) \\
&\geq \mathbb{P}(A^\circ_3(\delta, \eta)) - \mathbb{P}(A^\times_1(\epsilon, \delta, \eta)) - \mathbb{P}(A^\times_2(\widetilde{\epsilon}, \delta, \eta)) \\
&\geq \mathbb{P}(A^\circ_3(\delta, \eta)) - \eta^N \qquad \text{due to (B.59) and (B.60)} \\
&\geq \Big(\frac{H(2b/\eta)}{H(\delta/\eta)} \Big)^{l^*} \Big(\frac{\bar{\epsilon}}{4M} H(\delta/\eta)/\eta \Big)^{l^*-1} - \eta^N \qquad \text{due to Lemma B.2} \\
&\geq 2c_* \delta^\alpha (H(1/\eta)/\eta)^{l^*-1} - \eta^N \qquad \text{for all } \eta \text{ sufficiently small, due to } H \in \mathcal{RV}_{-\alpha} \\
&\geq c_* \delta^\alpha (H(1/\eta)/\eta)^{l^*-1}.
\end{aligned}$$

□

In order to present the main result of this section, we need to take into account the loss landscape outside of the current attraction field Ω. Recall that there are $n_{\min}$ attraction fields on f. For all the attraction fields different from Ω, we call them $(\widetilde{\Omega}_k)_{k=1}^{n_{\min}-1}$ where, for each $k \in [n_{\min} - 1]$, the attraction field $\widetilde{\Omega}_k = (s^-_k, s^+_k)$ with the corresponding local minimum located at $\widetilde{m}_k$. Also, recall that $\sigma(\eta)$ is the first time X^η_n exits from Ω. Building upon these concepts, we can define a stopping time

$$\tau(\eta, \epsilon) \triangleq \min\{n \geq 0 : \ X^\eta_n \in \bigcup_{k=1}^{n_{\min}-1} [\widetilde{m}_k - 2\epsilon, \widetilde{m}_k + 2\epsilon]\} \tag{B.62}$$

as the first time the SGD iterates visit a minimizer in an attraction field that is different from Ω. Besides, let index $J_\sigma(\eta)$ be such that

$$J_\sigma(\eta) = j \iff X^\eta_{\sigma(\eta)} \in \widetilde{\Omega}_j \quad \forall j \in [n_{\min} - 1]. \tag{B.63}$$

In other words, it is the label of the attraction field that X^η_n escapes to. Lastly, define

$$\lambda(\eta) \triangleq H(1/\eta) \Big(H(1/\eta)/\eta \Big)^{l^*-1}, \tag{B.64}$$
$$\nu^\Omega \triangleq \mu\big(E(0)\big), \tag{B.65}$$
$$\nu^\Omega_k \triangleq \mu\big(E(0) \cap h^{-1}(\widetilde{\Omega}_k)\big) \quad \forall k \in [n_{\min} - 1]. \tag{B.66}$$

For definitions of the measure μ, set E, and mapping h, see (B.51) and (B.52).

Now we are ready to state Proposition B.22, the most important technical tool in this section. In (B.67) and (B.68), we provide upper and lower bounds for the joint distribution of first exit time σ and the label J_σ indexing the attraction field we escape to; it is worth noticing that the claims hold uniformly for all $u > C$. In (B.69) and (B.70), we provide upper and lower bounds for the joint distribution of when we first visit a different local minimum (which is equal to τ) and which one we visit (indicated by X^η_τ). The similarity between (B.67) (B.68) and (B.69)(B.70) suggests a strong correlation between the behavior of the SGD iterates at time $\sigma(\eta)$ and that of time $\tau(\eta, \epsilon)$, and this is corroborated by (B.71): we show that it is almost always the case that τ is very close to σ, and on the short time interval $[\sigma(\eta), \tau(\eta, \epsilon)]$ the SGD iterates stay within the same attraction field.

Proposition B.22. *Given $C > 0$ and some $k' \in [n_{min} - 1]$, the following claims hold for all $\epsilon > 0$ that is sufficiently small:*

$$\limsup_{\eta\downarrow 0} \sup_{u\in(C,\infty)} \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\nu^\Omega\lambda(\eta)\sigma(\eta) > u,\ J_\sigma(\eta) = k'\Big) \le 2C + \exp(-(1-C)^3 u)\frac{\nu^\Omega_{k'} + C}{\nu^\Omega}, \tag{B.67}$$

$$\liminf_{\eta\downarrow 0} \inf_{u\in(C,\infty)} \inf_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\nu^\Omega\lambda(\eta)\sigma(\eta) > u,\ J_\sigma(\eta) = k'\Big) \ge -2C + \exp(-(1+C)^3 u)\frac{\nu^\Omega_{k'} - C}{\nu^\Omega}, \tag{B.68}$$

$$\limsup_{\eta\downarrow 0} \sup_{u\in(C,\infty)} \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\nu^\Omega\lambda(\eta)\tau(\eta,\epsilon) > u,\ X^\eta_{\tau(\eta,\epsilon)} \in B(\widetilde{m}_{k'}, 2\epsilon)\Big) \le 4C + \exp(-(1-C)^3 u)\frac{\nu^\Omega_{k'} + C}{\nu^\Omega}, \tag{B.69}$$

$$\liminf_{\eta\downarrow 0} \inf_{u\in(C,\infty)} \inf_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\nu^\Omega\lambda(\eta)\tau(\eta,\epsilon) > u,\ X^\eta_{\tau(\eta,\epsilon)} \in B(\widetilde{m}_{k'}, 2\epsilon)\Big) \ge -4C + \exp(-(1+C)^3 u)\frac{\nu^\Omega_{k'} - C}{\nu^\Omega}, \tag{B.70}$$

$$\liminf_{\eta\downarrow 0} \inf_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\lambda(\eta)\big(\tau(\eta,\epsilon) - \sigma(\eta)\big) < C,\ X^\eta_n \in \widetilde{\Omega}_{J_\sigma(\eta)}\ \ \forall n \in [\sigma(\eta), \tau(\eta,\epsilon)]\Big) \ge 1 - C. \tag{B.71}$$

Before presenting the proof to Proposition B.22, we make some preparations. First, we introduce stopping times (for all $k \ge 1$)

$$\tau_k(\epsilon,\delta,\eta) = \min\{n > \widetilde{\tau}_{k-1}(\epsilon,\delta,\eta) : \eta|Z_n| > \delta\}$$
$$\widetilde{\tau}_k(\epsilon,\delta,\eta) = \min\{n \ge \tau_k(\epsilon,\delta,\eta) : |X^\eta_n| \le 2\epsilon\}$$

with the convention that $\tau_0(\epsilon,\delta,\eta) = \widetilde{\tau}_0(\epsilon,\delta,\eta) = 0$. The intuitive interpretation is as follows. For the fixed ϵ we treat $[-2\epsilon, 2\epsilon]$ as a small neighborhood of the local minimum of the attraction field Ω. All the $\widetilde{\tau}_k$ partitioned the entire timeline into different *attempts* of escaping Ω. The interval $[\widetilde{\tau}_{k-1}, \widetilde{\tau}_k]$ can be viewed as the $k-$th *attempt*. If for $\sigma(\eta)$, the first exit time defined in (B.57), we have $\sigma(\eta) > \widetilde{\tau}_k$, then we consider the $k-$th attempt of escape as a *failure* because the SGD iterates returned to this small neighborhood of the local minimum again without exiting the attraction field. On the other hand, the stopping times τ_{k-1} indicate the arrival time of the first large jump during the $k-$th *attempt*. The proviso that $\widetilde{\tau}_k \ge \tau_{k-1}$ can be interpreted, intuitively, as that an *attempt* is considered failed only if, after some significant efforts to exit (for instance, a large jump) has been observed, the SGD iterates still returned to the small neighborhood $[-2\epsilon, 2\epsilon]$. Regarding the notations, we add a remark that when there is no ambiguity we will drop the dependency on ϵ, δ, η and simply write $\tau_k, \widetilde{\tau}_k$.

To facilitate the characterization of events during each *attempt*, we introduce the following definitions. First, for all $k \ge 1$, let

$$\mathbf{j}_k \triangleq \#\{n = \tau_{k-1}(\epsilon,\delta,\eta), \tau_{k-1}(\epsilon,\delta,\eta) + 1, \cdots, \widetilde{\tau}_k(\epsilon,\delta,\eta) \wedge \sigma(\eta) : \eta|Z_n| > \delta\}$$

be the number of large jumps during the $k-$th *attempt*. Two implications of this definition:

- First, for any k with $\sigma(\eta) < \widetilde{\tau}_k$, we have $\mathbf{j}_k = 0$. Note that this proposition concerns the dynamics of SGD up until $\sigma(\eta)$, the first time the SGD iterates escaped from Ω, so there is no need to consider an attempt that is after $\sigma(\eta)$, and we will not do so in the analysis below;
- Besides, the random variable $\mathbf{j}_k$ is measurable w.r.t. $\mathcal{F}_{\widetilde{\tau}_k \wedge \sigma(\eta)}$, the stopped $\sigma-$algebra generated by the stopping time $\widetilde{\tau}_k \wedge \sigma(\eta)$.

Furthermore, for each $k = 1, 2, \cdots$, let

$$\begin{aligned} T_{k,1}(\epsilon, \delta, \eta) &= \tau_{k-1}(\epsilon, \delta, \eta) \wedge \sigma(\eta), \\ T_{k,j}(\epsilon, \delta, \eta) &= \min\{n > T_{k,j-1}(\epsilon, \delta, \eta) : \eta |Z_n| > \delta\} \wedge \sigma(\eta) \wedge \widetilde{\tau}_k \quad \forall j \geq 2, \\ W_{k,j}(\epsilon, \delta, \eta) &= Z_{T_{k,j}(\epsilon,\delta,\eta)} \quad \forall j \geq 1 \end{aligned}$$

with the convention $T_{k,0}(\epsilon, \delta, \eta) = \widetilde{\tau}_{k-1}(\epsilon, \delta, \eta)$. Note that for any $k \geq 1, j \geq 1$, $T_{k,j}$ is a stopping time. Besides, from the definition of $\mathbf{j}_k$, one can see that

$$\widetilde{\tau}_{k-1} + 1 \leq T_{k,j} \leq \widetilde{\tau}_k \wedge \sigma(\eta) \quad \forall j \in [\mathbf{j}_k], \tag{B.72}$$

and the sequences $\big(T_{k,j}\big)_{j=1}^{\mathbf{j}_k}$ and $\big(W_{k,j}\big)_{j=1}^{\mathbf{j}_k}$ are the arrival times and sizes of *large* jumps during the $k-$th attempt, respectively. Again, when there is no ambiguity we will drop the dependency on ϵ, δ, η and simply write $T_{k,j}$ and $W_{k,j}$.

In order to prove Proposition B.22, we analyze the most likely scenario that the exit from Ω would happen. Specifically, we will introduce a series of events with superscript $\times$ or $\circ$, where $\times$ indicates that the event is *atypical* or unlikely to happen and $\circ$ means that it is a *typical* event and is likely to be observed before the first exit from the attraction field Ω. Besides, the subscript k indicates that the event in discussion concerns the dynamics of SGD during the $k-$th attempt. Our goal is to show that for some event $\mathbf{A}^{\times}(\epsilon, \delta, \eta)$ its probability becomes sufficiently small as learning rate η tends to 0, so the escape from Ω almost always occurs in the manner described by $(\mathbf{A}^{\times}(\epsilon, \delta, \eta))^c$. In particular, the definition of this *atypical* scenario $\mathbf{A}^{\times}$ involves the union of some *atypical* events $\mathbf{A}_k^{\times}, \mathbf{B}_k^{\times}$ that occur in the $k-$th attempt. In other words, the intuition of $\mathbf{A}^{\times}$ is that something *abnormal* happened during one of the attempts before the final exit.

Here is one more comment for the general naming convention of these events. Events with label $\mathbf{A}$ often describe the "efforts" made in an attempt to get out of Ω (such as large noises), while those with label $\mathbf{B}$ concern how the SGD iterates return to $[-2\epsilon, 2\epsilon]$ (and how this attempt fails). For instance, $\mathbf{A}_k^{\times}$ discusses the unlikely scenario before T_{k,l^*}, the arrival of the l^*-th large jump in this attempt, while $\mathbf{B}_k^{\times}$ in general discusses the abnormal cases after T_{k,l^*} and before the return to $[-2\epsilon, 2\epsilon]$. On the other hand, $\mathbf{A}_k^{\circ}$ describes a successful escape during $k-$th attempt, while $\mathbf{B}_k^{\circ}$ means that during this attempt the iterates return to without spending too much time.

Now we proceed and provide a formal definition and analysis of the aforementioned series of events. As building blocks, we inspect the process $(X_n^{\eta})_{n \geq 1}$ at a even finer granularity, and bound the probability of some events $(\mathbf{A}_{k,i}^{\times})_{i \geq 0}$, $(\mathbf{B}_{k,i}^{\times})_{i \geq 1}$ detailing several cases that are *unlikely* to occur during the escape from or return to local minimum in the $k-$th attempt. First, for each $k \geq 1$, define the event

$$\mathbf{A}_{k,0}^{\times}(\epsilon, \delta, \eta) \triangleq \Big\{ \exists i = 0, 1, \cdots, l^* \wedge \mathbf{j}_k \ s.t. \max_{j = T_{k,i}+1, \cdots, (T_{k,i+1}-1) \wedge \widetilde{\tau}_k \wedge \sigma(\eta)} \eta |Z_{T_{k,i}+1} + \cdots + Z_j| > \widetilde{\epsilon} \Big\}. \tag{B.73}$$

Intuitively speaking, the event characterizes the atypical scenario where, during the $k-$th attempt, there is some large fluctuations (compared to $\widetilde{\epsilon}$) between any of the first l^* large jumps (or the first $\mathbf{j}_k$ large jumps in case that $\mathbf{j}_k < l^*$). Similarly, consider event (for all $k \geq 1$)

$$\mathbf{A}_{k,1}^{\times}(\epsilon, \delta, \eta) \triangleq \Big\{ \sigma(\eta) < \widetilde{\tau}_k, \ \mathbf{j}_k < l^* \Big\} \tag{B.74}$$

that describes the atypical case where the exit occurs during the $k-$th attempt with less than l^* large jumps. Next, for all $k \geq 1$ we have another atypical event (note that from (B.72) we can see that, for any $j \geq 1$, $\mathbf{j}_k \geq j$ implies $T_{k,j} \leq \sigma(\eta) \wedge \widetilde{\tau}_k$)

$$\mathbf{A}^{\times}_{k,2} \triangleq \Big\{ \mathbf{j}_k \geq l^*,\ \exists j = 2, 3, \cdots, l^*\ s.t.\ T_{k,j} - T_{k,j-1} > 2\hat{t}(\epsilon)/\eta \Big\}. \tag{B.75}$$

representing the case where we have at least l^* large noises during the $k-$th attempt, but for some of the large noise (from the 2nd to the l^*-th), the inter-arrival time is unusually long. Moving on, we consider the following events (defined for all $k \geq 1$)

$$\mathbf{A}^{\times}_{k,3} \triangleq \Big\{ \mathbf{j}_k < l^*,\ \widetilde{\tau}_k < \sigma(\eta),\ \widetilde{\tau}_k - T_{k,1} > 2l^*\hat{t}(\epsilon)/\eta \Big\} \tag{B.76}$$

that describes the atypical case where the $k-$th attempt failed but the return to the small neighborhood $[-2\epsilon, 2\epsilon]$ took unusually long time.

The following event also concerns the scenario where there are at least l^* large noises during the $k-$th attempt:

$$\mathbf{A}^{\times}_{k,4} \triangleq \Big\{ \mathbf{j}_k \geq l^*,\ |X^{\eta}_{T_{k,l^*}}| \geq r - \bar{\epsilon},\ \exists j = 1, 2, \cdots, l^*\ s.t.\ \eta|W_{k,j}| \leq \bar{\delta} \Big\}; \tag{B.77}$$

specifically, it describes the atypical case where, during this attempt, right after the l^*- large noise the SGD iterate is far enough from the local minimum yet some of the large noises are not that *large*. Lastly, by defining events

$$\mathbf{A}^{\times}_{k,5} \triangleq \Big\{ \mathbf{j}_k \geq l^*,\ T_{k,l^*} \leq \sigma(\eta) \wedge \widetilde{\tau}_k,\ X^{\eta}_{T_{k,l^*}} \in \bigcup_{j \in [n_{\min}-1]} [s_j - \epsilon, s_j + \epsilon] \Big\}, \tag{B.78}$$

we analyze an atypical case where the SGD iterates arrive at somewhere too close to the boundaries of Ω at the arrival time of the l^* large noise during this attempt. As an amalgamation of these atypical scenarios, we let

$$\mathbf{A}^{\times}_{k}(\epsilon, \delta, \eta) \triangleq \bigcup_{i=0}^{5} \mathbf{A}^{\times}_{k,i}(\epsilon, \delta, \eta). \tag{B.79}$$

Also, we analyze the probability of some events $(\mathbf{B}^{\times}_k)_{k \geq 1}$ that concern the SGD dynamics after the l^*-th large noise during the $k-$th attempt. Let us define

$$\begin{aligned}
\mathbf{B}^{\times}_{k,1}(\epsilon, \delta, \eta) &\triangleq \Big\{ \mathbf{j}_k \geq l^*,\ X^{\eta}_{T_{k,l^*}} \in [s_- + \epsilon, s_+ - \epsilon],\ T_{k,j} - T_{k,j-1} \leq 2\frac{\hat{t}(\epsilon)}{\eta}\ \forall j = 2, 3, \cdots, l^* \Big\} \\
\mathbf{B}^{\times}_{k,2}(\epsilon, \delta, \eta) &\triangleq \{\widetilde{\tau}_k - T_{k,l^*} > \rho(\epsilon)/\eta\} \cup \{\sigma(\eta) < \widetilde{\tau}_k\} \\
\mathbf{B}^{\times}_{k}(\epsilon, \delta, \eta) &\triangleq \mathbf{B}^{\times}_{k,1} \cap \mathbf{B}^{\times}_{k,2}
\end{aligned} \tag{B.80}$$

where $\rho(\cdot)$ is the function in Lemma B.11. From the definition of $\mathbf{B}^{\times}_k$, in particular the inclusion of $\mathbf{B}^{\times}_{k,2}$, one can see that the intuitive interpretation of event $\mathbf{B}^{\times}_k$ is that the SGD iterates *did not return* to local minimum efficiently (or simply escaped from the attraction field) after the l^*-th large noise during the $k-$th attempt. In comparison, the following events will characterize what would typically happen during each attempt:

$$\mathbf{A}^{\circ}_{k}(\epsilon, \delta, \eta) \triangleq \{\mathbf{j}_k \geq l^*,\ \sigma(\eta) = T_{k,l^*},\ X^{\eta}_{T_{k,l^*}} \notin [s_- - \epsilon, s_+ + \epsilon],\ T_{k,j} - T_{k,j-1} \leq \frac{2\hat{t}(\epsilon)}{\eta}\ \forall j = 2, 3, \cdots, l^*\}, \tag{B.81}$$

$$\mathbf{B}^{\circ}_{k}(\epsilon, \delta, \eta) \triangleq \{\sigma(\eta) > \widetilde{\tau}_k,\ \widetilde{\tau}_k - T_{k,1} \leq \frac{2l^*\hat{t}(\epsilon) + \rho(\epsilon)}{\eta}\}. \tag{B.82}$$

Intuitively speaking, $\mathbf{A}_k^\circ$ tells us that the exit happened right at T_{k,l^*}, the arrival time of the l^*-th large noise during the $k-$th attempt, and $\mathbf{B}_k^\circ$ tells us that the first exit from Ω did not occur during the $k-$th attempt, and the SGD iterates returned to local minimum rather efficiently. All the preparations above allow use to define

$$\mathbf{A}^\times(\epsilon,\delta,\eta) \triangleq \bigcup_{k\geq 1}\left(\bigcap_{i=1}^{k-1}\left(\mathbf{A}_i^\times \cup \mathbf{B}_i^\times \cup \mathbf{A}_i^\circ\right)^c\right)\cap\left(\mathbf{A}_k^\times \cup \mathbf{B}_k^\times\right). \tag{B.83}$$

We need the next lemma in the proof of Proposition B.22. As mentioned earlier, the takeaway is that $\mathbf{A}^\times$ is indeed *atypical* in the sense that we will almost always observe $(\mathbf{A}^\times)^c$.

Lemma B.23. *Given any $C>0$, the following claim holds for all $\epsilon>0,\delta>0$ sufficiently small:*

$$\limsup_{\eta\downarrow 0}\sup_{|x|\leq 2\epsilon}\mathbb{P}_x(\boldsymbol{A}^\times(\epsilon,\delta,\eta)) < C.$$

Proof. We fix some parameters for the proof. First, with out loss of generality we only consider $C\in(0,1)$, and we fix some $N>\alpha l^*$. Next we discuss the valid range of ϵ for the claim to hold. We only consider $\epsilon>0$ such that

$$\epsilon < \frac{\bar{\epsilon}}{6(\bar{\rho}+\widetilde{\rho}+3)}\wedge\frac{\epsilon_0}{3}$$

where $\bar{\rho}$ and $\widetilde{\rho}$ are the constants in Corollary B.5 and Corollary B.8 respectively, and ϵ_0 is the constant in (B.1). Moreover, recall function Ψ in Lemma B.20 and the constant $c_*>0$ in Lemma B.21. Due to $\lim_{\epsilon\downarrow 0}\Psi(\epsilon)=0$, it holds for all ϵ small enough such that

$$\frac{3\Psi(\epsilon)}{c_*} < C \tag{B.84}$$

In our proof we only consider ϵ small enough so the inequality above holds, and the claim in Lemma B.11 holds. Now we specify the valid range of parameter δ that will be used below:

- For all sufficiently small $\delta>0$, the claim in Lemma B.12 will hold for the prescribed ϵ and with $N_0=N$;
- For all sufficiently small $\delta>0$, the claims in Lemma B.16, Corollary B.17, Corollary B.18 and Lemma B.19 will hold with the prescribed ϵ and N;
- For all sufficiently small $\delta>0$, the inequalities in Lemma B.20 and B.21 will hold for the ϵ we fixed at the beginning;
- Lastly, let

$$\widetilde{\epsilon} = \min\left\{\frac{\epsilon}{2\exp(2\bar{t}M)}, \frac{\epsilon}{4\exp(2M\hat{t}(\epsilon))}\right\}$$
$$\hat{\epsilon} = 3(\bar{\rho}+\widetilde{\rho}+3)\epsilon.$$

  where the function $\hat{t}(\cdot)$ is defined in (B.24); due to Corollary B.13 we know that for all sufficiently small $\delta>0$ the claim (B.50) holds with the chosen $\widetilde{\epsilon}$ and N.

We show that the claim holds for any ϵ,δ small enough to satisfy the conditions above.

First, recall that

$$\mathbf{A}_{k,0}^\times(\epsilon,\delta,\eta) \triangleq \left\{\exists i=0,1,\cdots,l^*\wedge\mathbf{j}_k \ s.t. \max_{j=T_{k,i}+1,\cdots,(T_{k,i+1}-1)\wedge\widetilde{\tau}_k\wedge\sigma(\eta)}\eta|Z_{T_{k,i}+1}+\cdots+Z_j|>\widetilde{\epsilon}\right\}.$$

Due to our choice of δ stated earlier and Corollary B.13, there exists some $\eta_0 > 0$ such that for all $\eta \in (0, \eta_0)$,

$$\mathbb{P}\big(\mathbf{A}^{\times}_{k,0}(\epsilon,\delta,\eta)\big) \le \eta^N \quad \forall k \ge 1. \tag{B.85}$$

Similarly, recall that $\mathbf{A}^{\times}_{k,1}(\epsilon,\delta,\eta) \triangleq \Big\{\sigma(\eta) < \widetilde{\tau}_k,\ \mathbf{j}_k < l^*\Big\}$. Let us temporarily focus on the first attempt (namely the case $k = 1$). From Lemma B.16 and our choice of ϵ and δ, we know the existence of some $\eta_1 > 0$ such that

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\big(\mathbf{A}^{\times}_{1,1}(\epsilon,\delta,\eta)\big) \le \eta^N \quad \forall \eta \in (0,\eta_1). \tag{B.86}$$

Next, for $\mathbf{A}^{\times}_{k,2} \triangleq \Big\{\mathbf{j}_k \ge l^*,\ \exists j = 2,3,\cdots,l^*\ s.t.\ T_{k,j} - T_{k,j-1} > 2\hat{t}(\epsilon)/\eta\Big\}$, from Corollary B.17 and our choice of δ at the beginning, we have the existence of some $\eta_2 > 0$ such that

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\big(\mathbf{A}^{\times}_{1,2}(\epsilon,\delta,\eta)\big) \le \eta^N \quad \forall \eta \in (0,\eta_2). \tag{B.87}$$

Moving on, for $\mathbf{A}^{\times}_{k,3} \triangleq \Big\{\mathbf{j}_k < l^*,\ \widetilde{\tau}_k < \sigma(\eta),\ \widetilde{\tau}_k - T_{k,1} > 2l^*\hat{t}(\epsilon)/\eta\Big\}$, due to Corollary B.18 and our choice of ϵ, δ, we have the existence of some $\eta_3 > 0$ such that

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\big(\mathbf{A}^{\times}_{1,3}(\epsilon,\delta,\eta)\big) \le \eta^N \quad \forall \eta \in (0,\eta_3). \tag{B.88}$$

As for $\mathbf{A}^{\times}_{k,4} \triangleq \Big\{\mathbf{j}_k \ge l^*,\ |X^{\eta}_{T_{k,l^*}}| \ge r - \bar{\epsilon},\ \exists j = 1,2,\cdots,l^*\ s.t.\ \eta|W_{k,j}| \le \bar{\delta}\Big\}$, from Lemma B.19, one can see the existence of $\eta_4 > 0$ such that

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\big(\mathbf{A}^{\times}_{1,4}(\epsilon,\delta,\eta)\big) \le \eta^N \quad \forall \eta \in (0,\eta_4). \tag{B.89}$$

Lastly, for $\mathbf{A}^{\times}_{k,5} \triangleq \Big\{\mathbf{j}_k \ge l^*,\ T_{k,l^*} \le \sigma(\eta) \wedge \widetilde{\tau}_k,\ X^{\eta}_{T_{k,l^*}} \in \bigcup_{j\in[n_{\min}-1]}[s_j - \epsilon, s_j + \epsilon]\Big\}$, from Lemma B.20 we see the existence of $\eta_5 > 0$ such that

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\big(\mathbf{A}^{\times}_{1,5}(\epsilon,\delta,\eta)\big) \le 2\delta^{\alpha}\Psi(\epsilon)\Big(H(1/\eta)/\eta\Big)^{l^*-1} \quad \forall \eta \in (0,\eta_5). \tag{B.90}$$

Recall that $\mathbf{A}^{\times}_k(\epsilon,\delta,\eta) = \bigcup_{i=0}^{5}\mathbf{A}^{\times}_{k,i}(\epsilon,\delta,\eta)$. Also, for definitions of $\mathbf{B}^{\times}_k, \mathbf{A}^{\circ}_k, \mathbf{B}^{\circ}_k$, see (B.80),(B.81),(B.82) respectively. Our next goal is to establish bounds regarding the probabilities of these events. First, if we consider the event $\bigcap_{j=1}^{k}(\mathbf{A}^{\times}_j \cup \mathbf{B}^{\times}_j)^c \cap \mathbf{B}^{\circ}_j$, then the inclusion of the $(\mathbf{B}^{\circ}_j)_{j=1}^{k}$ implies that during the first k attempts the SGD iterates have never left the attraction field, so

$$\bigcap_{j=1}^{k}(\mathbf{A}^{\times}_j \cup \mathbf{B}^{\times}_j)^c \cap \mathbf{B}^{\circ}_j = \big(\bigcap_{j=1}^{k}(\mathbf{A}^{\times}_j \cup \mathbf{B}^{\times}_j)^c \cap \mathbf{B}^{\circ}_j\big) \cap \{\sigma(\eta) > \widetilde{\tau}_k\}.$$

Next, note that

$$\begin{aligned}
&\mathbb{P}_x\big(\mathbf{B}^{\times}_k \mid \bigcap_{j=1}^{k-1}(\mathbf{A}^{\times}_j \cup \mathbf{B}^{\times}_j)^c \cap \mathbf{B}^{\circ}_j\big)\\
=&\mathbb{P}_x\big(\mathbf{B}^{\times}_{k,1} \mid \bigcap_{j=1}^{k-1}(\mathbf{A}^{\times}_j \cup \mathbf{B}^{\times}_j)^c \cap \mathbf{B}^{\circ}_j\big)\mathbb{P}_x\Big(\mathbf{B}^{\times}_{k,2} \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}^{\times}_j \cup \mathbf{B}^{\times}_j)^c \cap \mathbf{B}^{\circ}_j\big) \cap \mathbf{B}^{\times}_{k,1}\Big)
\end{aligned}$$

$$\leq \mathbb{P}_x\big(\mathbf{j}_k \geq l^*,\ T_{k,j} - T_{k,j-1} \leq \frac{2\hat{t}(\epsilon)}{\eta}\ \forall j = 2,3,\cdots,l^* \mid \bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big)$$
$$\cdot\, \mathbb{P}_x\Big(\mathbf{B}_{k,2}^\times \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times\Big).$$

From the definition of the events $\mathbf{A}_j^\times, \mathbf{B}_j^\times, \mathbf{B}_j^\circ$, one can see that $\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ \in \mathcal{F}_{\widetilde{\tau}_{k-1}\wedge\sigma(\eta)}$, and on this event $\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ$ we have $\sigma(\eta) > \widetilde{\tau}_{k-1}$. So by applying strong Markov property at stopping time $\widetilde{\tau}_{k-1} \wedge \sigma(\eta)$, we have

$$\mathbb{P}_x\big(\mathbf{B}_k^\times \mid \bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big)$$
$$\leq \mathbb{P}\Big(T_j^\eta(\delta) - T_{j-1}^\eta(\delta) \leq 2\hat{t}(\epsilon)/\eta\ \forall j \in [l^*-1]\Big) \cdot \mathbb{P}_x\Big(\mathbf{B}_{k,2}^\times \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times\Big)$$
$$\leq 2\Big(H(\delta/\eta)\hat{t}(\epsilon)/\eta\Big)^{l^*-1} \cdot \mathbb{P}_x\Big(\mathbf{B}_{k,2}^\times \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times\Big) \quad \text{for all } \eta \text{ sufficiently small due to Lemma B.2,}$$
$$\leq 4\Big(\frac{\hat{t}(\epsilon)}{\delta^\alpha}\Big)^{l^*-1}\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1} \cdot \mathbb{P}_x\Big(\mathbf{B}_{k,2}^\times \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times\Big)$$

for all sufficiently small η, due to $H \in \mathcal{RV}_{-\alpha}(\eta)$. Meanwhile, note that

- $\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times \in \mathcal{F}_{T_{k,l^*}}$;
- on this event $\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times$ we have $\sigma(\eta) \wedge \widetilde{\tau}_k > T_{k,l^*}$ and $X_{T_{k,l^*}}^\eta \in [s_- + \epsilon, s_+ - \epsilon]$.

Therefore, using Lemma B.11 and strong Markov property again (at stopping time T_{k,l^*}), we know the following inequality holds for all η sufficiently small:

$$\sup_{k\geq 1}\sup_{|x|\leq 2\epsilon} \mathbb{P}_x\Big(\mathbf{B}_{k,2}^\times \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times\Big)$$
$$= \sup_{k\geq 1}\sup_{|x|\leq 2\epsilon} \mathbb{P}_x\Big(\big\{\sigma(\eta) > \widetilde{\tau}_k,\ \widetilde{\tau}_k - T_{k,l^*} \leq \rho(\epsilon)/\eta\big\}^c \,\Big|\, \big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap \mathbf{B}_{k,1}^\times\Big)$$
$$\leq \Psi(\epsilon)\frac{\delta^\alpha}{4\big(\hat{t}(\epsilon)/\delta^\alpha\big)^{l^*-1}}.$$

Therefore, we know the existence of some $\eta_6 > 0$ such that

$$\sup_{|x|\leq 2\epsilon} \mathbb{P}_x\big(\mathbf{B}_k^\times \mid \bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \leq \Psi(\epsilon)\delta^\alpha\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1} \quad \forall \eta \in (0,\eta_6),\ \forall k \geq 1. \tag{B.91}$$

Similarly, we can bound conditional probabilities of the form $\mathbb{P}_x\big(\mathbf{A}_k^\times \mid \bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big)$. To be specific, recall that $\mathbf{A}_k^\times = \cup_{i=0}^5 \mathbf{A}_{k,i}^\times$. By combining (B.85)-(B.90) with Markov property, we know the existence of some $\eta_7 > 0$ such that

$$\sup_{|x|\leq 2\epsilon} \mathbb{P}_x\big(\mathbf{A}_k^\times \mid \bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \leq 5\eta^N + 2\Psi(\epsilon)\delta^\alpha\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1} \quad \forall \eta \in (0,\eta_7),\ \forall k \geq 1. \tag{B.92}$$

On the other hand, a lower bound can be established for conditional probability involving $\mathbf{A}_k^\circ$, the event defined in (B.81) describing the exit from Ω during an attempt with exactly l^* large noises. Using Lemma B.21 and Markov property of $(X_n^\eta)_{n\geq 1}$, one can see the existence of some $\eta_8 > 0$ such that

$$\inf_{|x|\leq 2\epsilon} \mathbb{P}_x\big(\mathbf{A}_k^\circ \mid \bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \geq c_* \delta^\alpha \Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1} \quad \forall \eta \in (0, \eta_8). \tag{B.93}$$

In order to apply the bounds (B.91)-(B.93), we make use of the following inclusion relationship:

$$\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \subseteq \mathbf{A}_k^\circ \cup \mathbf{B}_k^\circ. \tag{B.94}$$

To see why this is true, let us consider a decomposition of the event on the L.H.S. of (B.94). As mentioned above, on event $\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ$ we know that $\sigma(\eta) > \widetilde{\tau}_{k-1}$, so the $k-$th attempt occurred and there are only three possibilities on this event:

- $\mathbf{j}_k < l^*$;
- $\mathbf{j}_k \geq l^*,\ X^\eta_{T_{k,l^*}} \notin \Omega$;
- $\mathbf{j}_k \geq l^*,\ X^\eta_{T_{k,l^*}} \in \Omega$.

Let us partition the said event accordingly and analyze them one by one.

- On $\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \cap \{\mathbf{j}_k < l^*\}$, due to the exclusion of $\mathbf{A}_k^\times$ (especially $\mathbf{A}_{k,1}^\times$ and $\mathbf{A}_{k,3}^\times$), we can see that if $\mathbf{j}_k < l^*$, then we must have $\sigma(\eta) > \widetilde{\tau}_k$ and $\widetilde{\tau}_k - T_{k,1} \leq 2l^*\hat{t}(\epsilon)/\eta$. Therefore,

$$\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \cap \{\mathbf{j}_k < l^*\} \subseteq \mathbf{B}_k^\circ.$$

- On $\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \cap \{\mathbf{j}_k \geq l^*,\ X^\eta_{T_{k,l^*}} \notin \Omega\}$, then the exclusion of $\mathbf{A}_{k,2}^\times$ implies that $T_{k,j} - T_{k,j-1} \leq 2\hat{t}(\epsilon)/\eta$ for all $j = 2, \cdots, l^*$, and the exclusion of $\mathbf{A}_{k,5}^\times$ tells us that if $X^\eta_{T_{k,l^*}} \notin \Omega$, then we have $X^\eta_{T_{k,l^*}} \notin [s_- - \epsilon, s_+ + \epsilon]$. In summary,

$$\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \cap \{\mathbf{j}_k \geq l^*,\ X^\eta_{T_{k,l^*}} \notin \Omega\} \subseteq \mathbf{A}_k^\circ.$$

- On $\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \cap \{\mathbf{j}_k \geq l^*,\ X^\eta_{T_{k,l^*}} \in \Omega\}$, the exclusion of $\mathbf{A}_{k,2}^\times$ again implies that $T_{k,j} - T_{k,j-1} \leq 2\hat{t}(\epsilon)/\eta$ for all $j = 2, \cdots, l^*$, hence $T_{k,l^*} - T_{k,1} \leq 2l^*\hat{t}(\epsilon)/\eta$. Similarly, the exclusion of $\mathbf{A}_{k,5}^\times$ tells us that if $X^\eta_{T_{k,l^*}} \in \Omega$, then we have $X^\eta_{T_{k,l^*}} \in [s_- + \epsilon, s_+ - \epsilon]$. Now since $\mathbf{B}_k^\times$ did not occur (see the definition in (B.80)), we must have $\sigma(\eta) > \widetilde{\tau}_k$ and $\widetilde{\tau}_k - T_{k,l^*} \leq \rho(\epsilon)/\eta$, hence $\widetilde{\tau}_k - T_{k,1} \leq \frac{2l^*\hat{t}(\epsilon)+\rho(\epsilon)}{\eta}$. Therefore,

$$\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^\times \cup \mathbf{B}_j^\times)^c \cap \mathbf{B}_j^\circ\big) \cap (\mathbf{A}_k^\times \cup \mathbf{B}_k^\times)^c \cap \{\mathbf{j}_k \geq l^*,\ X^\eta_{T_{k,l^*}} \in \Omega\} \subseteq \mathbf{B}_k^\circ.$$

Collecting results above, we have (B.94). Now we discuss some of its implications. First, from (B.94) we can immediately get that

$$\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^{\times}\cup\mathbf{B}_j^{\times})^c\cap\mathbf{B}_j^{\circ}\big)\cap(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times})^c=\big(\bigcap_{j=1}^{k-1}(\mathbf{A}_j^{\times}\cup\mathbf{B}_j^{\times})^c\cap\mathbf{B}_j^{\circ}\big)\cap(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times})^c\cap(\mathbf{A}_k^{\circ}\cup\mathbf{B}_k^{\circ}). \tag{B.95}$$

Next, recall the definitions of $\mathbf{A}_k^{\circ}$ in (B.81) and $\mathbf{B}_k^{\circ}$ in (B.82), and one can see that $\mathbf{A}_k^{\circ}$ and $\mathbf{B}_k^{\circ}$ are mutually exclusive, since the former implies that the first exit occurs during the $k-$th attempt while the latter implies that this attempt fails. This fact and (B.95) allow us to conclude that

$$\bigcap_{i=1}^{k}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\cup\mathbf{A}_i^{\circ}\big)^c=\bigcap_{i=1}^{k}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}=\Big(\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\cap(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\cup\mathbf{A}_k^{\circ})^c. \tag{B.96}$$

Now we use the results obtained so far to bound the probability of

$$\mathbf{A}^{\times}(\epsilon,\delta,\eta)\triangleq\bigcup_{k\geq1}\Big(\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\cup\mathbf{A}_i^{\circ}\big)^c\Big)\cap\big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\big).$$

Using (B.96), we can see that (for any $x\in[-2\epsilon,2\epsilon]$)

$$\begin{aligned}
&\mathbb{P}_x(\mathbf{A}^{\times}(\epsilon,\delta,\eta))\\
=&\sum_{k\geq1}\mathbb{P}_x\bigg(\Big(\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\cup\mathbf{A}_i^{\circ}\big)^c\Big)\cap\big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\big)\bigg)\\
=&\sum_{k\geq1}\mathbb{P}_x\bigg(\Big(\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\cap\big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\big)\bigg)\\
=&\sum_{k\geq1}\mathbb{P}_x\Big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\mid\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\cdot\prod_{j=1}^{k-1}\mathbb{P}_x\Big(\bigcap_{i=1}^{j}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\;\Big|\;\bigcap_{i=1}^{j-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\\
=&\sum_{k\geq1}\mathbb{P}_x\Big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\mid\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\\
&\qquad\cdot\prod_{j=1}^{k-1}\mathbb{P}_x\Big(\big(\bigcap_{i=1}^{j-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\big)\cap(\mathbf{A}_j^{\times}\cup\mathbf{B}_j^{\times}\cup\mathbf{A}_j^{\circ})^c\;\Big|\;\bigcap_{i=1}^{j-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\\
=&\sum_{k\geq1}\mathbb{P}_x\Big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\mid\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\cdot\prod_{j=1}^{k-1}\mathbb{P}_x\Big(\big(\mathbf{A}_j^{\times}\cup\mathbf{B}_j^{\times}\cup\mathbf{A}_j^{\circ}\big)^c\;\Big|\;\bigcap_{i=1}^{j-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\\
\leq&\sum_{k\geq1}\mathbb{P}_x\Big(\mathbf{A}_k^{\times}\cup\mathbf{B}_k^{\times}\mid\bigcap_{i=1}^{k-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\cdot\prod_{j=1}^{k-1}\bigg(1-\mathbb{P}_x\Big(\mathbf{A}_j^{\circ}\;\Big|\;\bigcap_{i=1}^{j-1}\big(\mathbf{A}_i^{\times}\cup\mathbf{B}_i^{\times}\big)^c\cap\mathbf{B}_i^{\circ}\Big)\bigg).
\end{aligned}$$

This allows us to apply (B.91)-(B.93) and conclude that (here we only consider $\eta<\min\{\eta_i:i\in[8]\}$),

$$\begin{aligned}
&\sup_{|x|\leq2\epsilon}\mathbb{P}_x(\mathbf{A}^{\times}(\epsilon,\delta,\eta))\\
\leq&\sum_{k\geq1}\Big(5\eta^N+2\Psi(\epsilon)\delta^{\alpha}\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}\Big)\cdot\Big(1-c_*\delta^{\alpha}\Big(\frac{H(1/\eta)}{\eta}\Big)^{l^*-1}\Big)^{k-1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{5\eta^N + 2\Psi(\epsilon)\delta^\alpha \left(\frac{H(1/\eta)}{\eta}\right)^{l^*-1}}{c_*\delta^\alpha \left(\frac{H(1/\eta)}{\eta}\right)^{l^*-1}} \\
&\le \frac{2\Psi(\epsilon) + 5\eta^\alpha}{c_*} \quad \text{for sufficiently small } \eta, \text{ due to } H \in \mathcal{RV}_{-\alpha}(\eta) \text{ and our choice of } N > \alpha l^* \\
&\le \frac{3\Psi(\epsilon)}{c_*} < C \quad \text{for all } \eta \text{ small enough such that } 5\eta^\alpha < \Psi(\epsilon).
\end{aligned}
$$

The last inequality follows from our choice of ϵ in (B.84). This concludes the proof. □

Having established Lemma B.23, we return to Proposition B.22 and give a proof. Recall that, aside from the attraction field $\Omega = (s_-, s_+)$, there are $n_{\min} - 1$ other attraction fields $\widetilde{\Omega}_k = (s_k^-, s_k^+)$ (for each $k \in [n_{\min} - 1]$). Besides, the function $\lambda(\cdot)$ and constants ν^Ω, ν_k^Ω are defined in (B.64)-(B.66).

Proof of Proposition B.22. We fix some parameters for the proof. First, with out loss of generality we only need to consider $C \in (0, 1)$. Next we discuss the valid range of ϵ for the claim to hold. We only consider $\epsilon > 0$ such that

$$
\epsilon < \frac{\bar{\epsilon}}{6(\bar{\rho} + \widetilde{\rho} + 3)} \wedge \frac{\epsilon_0}{3}
$$

where $\bar{\rho}$ and $\widetilde{\rho}$ are the constants in Corollary B.5 and Corollary B.8 respectively, and ϵ_0 is the constant in (B.1). Due to continuity of measure μ, it holds for all ϵ small enough such that (let $\hat{\epsilon} = 3(\bar{\rho}+\widetilde{\rho}+3)\epsilon$)

$$
\frac{\mu(E(0))}{\mu\big(E\big(\hat{\epsilon}\big)\big)} < 1/(1 - C), \tag{B.97}
$$

$$
\frac{\mu(E(0))}{\mu\big(E\big(-\hat{\epsilon}\big)\big)} > 1/(1 + C), \tag{B.98}
$$

$$
\frac{\mu\Big(h^{-1}\big((s_- - 2\hat{\epsilon}, s_- + 2\hat{\epsilon}) \cup (s_+ - 2\hat{\epsilon}, s_+ + 2\hat{\epsilon})\big)\Big)}{\mu(E(-\hat{\epsilon}))} \le C \tag{B.99}
$$

$$
\frac{\mu\Big(E(\hat{\epsilon}) \cap \big(s_{k'}^- - \hat{\epsilon}, s_{k'}^+ + \hat{\epsilon}\big)\Big)}{\mu(E(\hat{\epsilon}))} \le \frac{\nu_{k'}^\Omega + C}{\nu^\Omega} \tag{B.100}
$$

$$
\frac{\mu\Big(E(-\hat{\epsilon}) \cap \big(s_{k'}^- + 2\hat{\epsilon}, s_{k'}^+ - 2\hat{\epsilon}\big)\Big)}{\mu(E(-\hat{\epsilon}))} \ge \frac{\nu_{k'}^\Omega - C}{\nu^\Omega} \tag{B.101}
$$

In our proof we only consider ϵ small enough so the inequality above holds, and the claims in Lemma B.11 hold. Moreover, we only consider ϵ and δ small enough so that Lemma B.23 hold and we have

$$
\lim_{\eta \downarrow 0} \sup_{|x| \le 2\epsilon} \mathbb{P}_x(\mathbf{A}^\times(\epsilon, \delta, \eta)) < C. \tag{B.102}
$$

We show that the desired claims hold for all ϵ, δ sufficiently small that satisfy conditions above.

First, in order to show (B.71), we define event

$$
\begin{aligned}
&\widetilde{\mathbf{A}}^\times(\epsilon, \delta, \eta) \\
&\triangleq \big(\mathbf{A}^\times(\epsilon, \delta, \eta)\big)^c \cap \Big\{\lambda(\eta)\big(\tau(\eta, \epsilon) - \sigma(\eta)\big) \ge C \text{ or } \exists n = \sigma(\eta) + 1, \cdots, \tau(\eta, \epsilon) \text{ such that } X_n^\eta \notin \widetilde{\Omega}_{J_\sigma(\eta)}\Big\}
\end{aligned}
$$

Since $\lambda \in \mathcal{RV}_{-1-l^*(\alpha-1)}$ and $\alpha > 1$, for the ϵ we fixed at the beginning of this proof, $\rho(\epsilon)$ is a fixed constant as well (the function ρ is defined in Lemma B.11) and we have $\lim_{\eta\downarrow 0} \lambda(\eta)\rho(\epsilon)/\eta = 0$. Next, the occurrence of $\big(\mathbf{A}^\times(\epsilon, \delta, \eta)\big)^c$ (in particular, the exclusion of all the $\mathbf{A}_{k,5}^\times$ defined in (B.78)),

we know that $X^{\eta}_{\sigma(\eta)} \notin [s^-_{J_\sigma} - \epsilon, s^-_{J_\sigma} + \epsilon] \cup [s^+_{J_\sigma} - \epsilon, s^+_{J_\sigma} + \epsilon]$ (recall that for any $k \in [n_{\min} - 1]$, we have $\widetilde{\Omega}_j = (s^-_j, s^+_j)$; for definition of J_σ see (B.63)). Meanwhile, for all η sufficiently small, we have $\epsilon/\lambda(\eta) > \rho(\epsilon)/\eta$. Therefore, using Lemma B.11 we can see that (for all η sufficiently small)

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\Big(\widetilde{\mathbf{A}}^{\times}(\epsilon,\delta,\eta) \mid \big(\mathbf{A}^{\times}(\epsilon,\delta,\eta)\big)^c\Big) \le C. \tag{B.103}$$

Lastly, observe that

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\lambda(\eta)\big(\tau(\eta,\epsilon)-\sigma(\eta)\big) \ge C \text{ or } \exists n = \sigma(\eta)+1,\cdots,\tau(\eta,\epsilon) \text{ such that } X^{\eta}_n \notin \widetilde{\Omega}_{J_\sigma(\eta)}\Big\}\Big)\\
\le&\mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \Big\{\lambda(\eta)\big(\tau(\eta,\epsilon)-\sigma(\eta)\big) \ge C \text{ or } \exists n = \sigma(\eta)+1,\cdots,\tau(\eta,\epsilon) \text{ such that } X^{\eta}_n \notin \widetilde{\Omega}_{J_\sigma(\eta)}\Big\}\Big)\\
&+\mathbb{P}_x(\mathbf{A}^{\times})
\end{aligned}$$

so by combining (B.102) with (B.103), we can obtain (B.71).

Moving on, we discuss the upper bounds (B.67) and (B.69). Recall that the fixed constant $k' \in [n_{\min} - 1]$ is prescribed in the description of this proposition. Let us observe some facts on event $(\mathbf{A}^{\times}(\epsilon,\eta,\delta))^c \cap \{J_\sigma(\eta) = k'\}$: If we let $J(\epsilon,\delta,\eta) \triangleq \sup\{k \ge 0 : \widetilde{\tau}_k < \sigma(\eta)\}$ be the number of attempts it took to escape, and

$$J^{\uparrow}(\epsilon,\delta,\eta) \triangleq \min\{k \ge 1 : T_{k,1} \text{ has } \Big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\Big)\text{−overflow}\},$$

then for all η sufficiently small, we must have $J \le J^{\uparrow}$ on event $(\mathbf{A}^{\times}(\epsilon,\eta,\delta))^c \cap \{J_\sigma(\eta) = k'\}$. To see this via a proof of contradiction, let us assume that, for some arbitrary positive integer j, there exists some sample path on $(\mathbf{A}^{\times})^c \cap \{J_\sigma(\eta) = k'\}$ such that $J^{\uparrow} = j < J$. Then from the definition of $(\mathbf{A}^{\times})^c$, in particular the exclusion of event $\mathbf{A}^{\times}_{j,0}$ (see the definition in (B.73)), for all sufficiently small η, we are able to apply Corollary B.8 and B.5 and conclude that $X^{\eta}_{T_{j,l^*}} \notin \Omega$: indeed, using Corollary B.8 and B.5 we can show that the distance between $X^{\eta}_{T_{j,l^*}}$ and the perturbed ODE

$$\widetilde{\mathbf{x}}^{\eta}\Big(T_{j,l^*} - T_{j,1}, 0; \big(0, T_{j,2} - T_{j,1}, \cdots, T_{j,l^*} - T_{j,1}\big), \big(\eta W_{j,1}, \cdots, \eta W_{j,l^*}\big)\Big)$$

is strictly less than $3(\bar{\rho}+\widetilde{\rho}+3)\epsilon$; on the other hand, the definition of $\Big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\Big)$−overflow implies that

$$\begin{aligned}
&\widetilde{\mathbf{x}}^{\eta}\Big(T_{j,l^*} - T_{j,1}, 0; \big(0, T_{j,2} - T_{j,1}, \cdots, T_{j,l^*} - T_{j,1}\big), \big(\eta W_{j,1}, \cdots, \eta W_{j,l^*}\big)\Big)\\
&\quad\notin [s_- - 3(\bar{\rho}+\widetilde{\rho}+3)\epsilon, s_+ + 3(\bar{\rho}+\widetilde{\rho}+3)\epsilon].
\end{aligned}$$

Therefore, we must have $X^{\eta}_{T_{j,l^*}} \notin \Omega$, which contradicts our assumption $j = J^{\uparrow} < J$. In summary, we have shown that, on $(\mathbf{A}^{\times})^c \cap \{J_\sigma = k'\}$, we have $J^{\uparrow}(\epsilon,\delta,\eta) \ge J(\epsilon,\delta,\eta)$. Similarly, if we consider

$$J^{\downarrow}(\epsilon,\delta,\eta) \triangleq \min\{k \ge 1 : T_{k,1} \text{ has } \big(-3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\big)\text{−overflow}\},$$

then by the same argument above we can show that $J^{\downarrow}(\epsilon,\delta,\eta) \le J(\epsilon,\delta,\eta)$. Now consider the following decomposition of events.

- On $\{J^{\downarrow} < J^{\uparrow}\}$, we know that for the first k such that $T_{k,1}$ has $\big(-3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\big)$−overflow, it does not have $\big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\big)$−overflow. Now we analyze the probability that Z_0 does not have $(\hat{\epsilon},\delta,\eta)$−overflow conditioning on that it does have $(-\hat{\epsilon},\delta,\eta)$−overflow (recall that we let $\hat{\epsilon} = 3(\bar{\rho}+\widetilde{\rho}+3)$). Using Lemma B.14 and the bound (B.99), we know that for all η sufficiently small,

$$\sup_{|x|\le 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J^{\downarrow} < J^{\uparrow}\}\Big) \le \frac{\mu\Big(h^{-1}\big((s_- - 2\hat{\epsilon}, s_- + 2\hat{\epsilon}) \cup (s_+ - 2\hat{\epsilon}, s_+ + 2\hat{\epsilon})\big)\Big)}{\mu(E(-\hat{\epsilon}))} \le C. \tag{B.104}$$

- On $(\mathbf{A}^{\times})^c \cap \{J_\sigma = k'\} \cap \{J^\uparrow = J^\downarrow\}$, due to $J^\uparrow = J^\downarrow = J$ we know that $T_{J(\epsilon,\delta,\eta),1}$ is the first among all $T_{k,1}$ to have $(\hat{\epsilon},\delta,\eta)-$overflow. Moreover, due to $\{J_\sigma = k'\}$ and using Corollary B.8 and B.5 again as we did above, we know that the overflow endpoint of $T_{J(\epsilon,\delta,\eta),1}$ is in $(s^-_{k'} - \hat{\epsilon}, s^+_{k'} + \hat{\epsilon})$ (recall that $\widetilde{\Omega}_{k'} = (s^-_{k'}, s^+_{k'})$). In summary, for any $n \geq 0$

$$\begin{aligned}
&(\mathbf{A}^{\times})^c \cap \{J_\sigma = k'\} \cap \{J^\uparrow = J^\downarrow > n\} \\
&\subseteq (\mathbf{A}^{\times})^c \cap \{J^\uparrow > n\} \cap \Big\{T_{J^\uparrow,1} \text{ has overflow endpoint in } (s^-_{k'} - \hat{\epsilon}, s^+_{k'} + \hat{\epsilon})\Big\}
\end{aligned}$$

so using Lemma B.14, we obtain that (for all η sufficiently small)

$$\begin{aligned}
&\sup_{|x|\leq 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J_\sigma = k'\} \cap \{J^\uparrow = J^\downarrow > n\}\Big) \\
&\leq \sup_{|x|\leq 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J^\uparrow > n\}\Big) \cdot \frac{p\big(\hat{\epsilon},\delta,\eta;(s^-_{k'} - \hat{\epsilon}, s^+_{k'} + \hat{\epsilon})\big)}{p(\hat{\epsilon},\delta,\eta)} \\
&\leq \sup_{|x|\leq 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J^\uparrow > n\}\Big) \cdot \frac{\nu^\Omega_{k'} + C}{\nu^\Omega}.
\end{aligned} \tag{B.105}$$

uniformly for any $n = 0, 1, 2, \cdots$ due to (B.100).

- On the other hand, on $(\mathbf{A}^{\times})^c$, if $T_{J^\downarrow,1}$ has overflow endpoint in $(s^-_{k'} + 2\hat{\epsilon}, s^+_{k'} - 2\hat{\epsilon})$, then from Definition B.1 we know that $T_{J^\downarrow,1}$ also has $(\hat{\epsilon},\delta,\eta)-$overflow, hence $J^\downarrow = J^\uparrow = J$. Moreover, using Corollary B.8 and B.5 again, we know that $X^\eta_{T_{J^\downarrow,l^*}} \in (s^-_{k'}, s^+_{k'})$ so $J_\sigma = k'$. In summary, for any $n \geq 0$,

$$\begin{aligned}
&(\mathbf{A}^{\times})^c \cap \{J_\sigma = k'\} \cap \{J^\uparrow = J^\downarrow > n\} \\
&\supseteq (\mathbf{A}^{\times})^c \cap \{J^\downarrow > n\} \cap \Big\{T_{J^\downarrow,1} \text{ has overflow endpoint in } (s^-_{k'} + 2\hat{\epsilon}, s^+_{k'} - 2\hat{\epsilon})\Big\}
\end{aligned}$$

so using Lemma B.14, we obtain that (for all η sufficiently small)

$$\begin{aligned}
&\inf_{|x|\leq 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J_\sigma = k'\} \cap \{J^\uparrow = J^\downarrow > n\}\Big) \\
&\geq \inf_{|x|\leq 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J^\downarrow > n\}\Big) \cdot \frac{p\big(-\hat{\epsilon},\delta,\eta;(s^-_{k'} + 2\hat{\epsilon}, s^+_{k'} - 2\hat{\epsilon})\big)}{p(-\hat{\epsilon},\delta,\eta)} \\
&\geq \inf_{|x|\leq 2\epsilon} \mathbb{P}_x\Big((\mathbf{A}^{\times})^c \cap \{J^\downarrow > n\}\Big) \cdot \frac{\nu^\Omega_{k'} - C}{\nu^\Omega}.
\end{aligned} \tag{B.106}$$

uniformly for any $n = 0, 1, 2, \cdots$ due to (B.101).

Besides, the following claim holds on event $(\mathbf{A}^{\times})^c$.

- From (B.96), the definition of $\mathbf{B}^\circ_k$ as well as the definition of event $\mathbf{A}^\circ_k$ (see (B.81)), one can see that for any $j = 1, 2, \cdots, J$, we have

$$\widetilde{\tau}_j \wedge \sigma(\eta) - T_{j,1} \leq \frac{2l^*\hat{t}(\epsilon) + \rho(\epsilon)}{\eta}.$$

- Now if we turn to the interval $(\widetilde{\tau}_{j-1}, T_{j,1}]$ (the time between the start of the $j-$th attempt and the arrival of the first large noise during this attempt) for each $j = 1, 2, \cdots, J$, and the following sequence constructed by concatenating these intervals

$$\mathbf{S}(\epsilon,\delta,\eta)$$

$\triangleq \big(1, 2, \cdots, T_{1,1}, \widetilde{\tau}_1 + 1, \widetilde{\tau}_1 + 2, \cdots, T_{2,1}, \cdots, \widetilde{\tau}_k + 1, \widetilde{\tau}_k + 2, \cdots, T_{k+1,1}, \widetilde{\tau}_{k+1} + 1, \widetilde{\tau}_{k+2} + 1, \cdots\big)$,

then the discussion above have shown that, for

$$\min\{n \in \mathbf{S}(\epsilon, \delta, \eta) : \ Z_n \ \text{ has } \Big(3(\bar{\rho} + \widetilde{\rho} + 3)\epsilon, \delta, \eta\Big)\text{−overflow}\} \geq T_{J,1}.$$

Meanwhile, from the definition of overflow we know that the probability that Z_1 has $\Big(3(\bar{\rho} + \widetilde{\rho} + 3)\epsilon, \delta, \eta\Big)$−overflow is equal to

$$H(\delta/\eta)p\Big(3(\bar{\rho} + \widetilde{\rho} + 3)\epsilon, \delta, \eta\Big).$$

- Therefore, if, within the duration of each attempt, we split the attempt into two parts at the arrival time of the first large jump $(T_{k,1})_{k\geq 1}$ at each attempt, and define (here the subscript *before* or *after* indicates that we are counting the steps before or after the first large jump in an attempt)

$$\begin{aligned}
\mathbf{S}_{\text{before}}(\epsilon, \delta, \eta) &\triangleq \{n \in \mathbf{S}(\epsilon, \delta, \eta) : \ n \leq \sigma(\eta)\}, \ I_{\text{before}}(\epsilon, \delta, \eta) \triangleq \#\mathbf{S}_{\text{before}}(\epsilon, \delta, \eta), \\
\mathbf{S}_{\text{after}}(\epsilon, \delta, \eta) &\triangleq \{n \notin \mathbf{S}(\epsilon, \delta, \eta) : \ n \leq \sigma(\eta)\}, \ I_{\text{after}}(\epsilon, \delta, \eta) \triangleq \#\mathbf{S}_{\text{after}}(\epsilon, \delta, \eta),
\end{aligned}$$

then we have $\sigma(\eta) = I_{\text{before}} + I_{\text{after}}$. Moreover, the discussion above implies that

$$\begin{aligned}
I_{\text{after}} &\leq J\big(2l^*\hat{t}(\epsilon) + \rho(\epsilon)\big)/\eta \\
I_{\text{before}} &\leq \min\{n \in \mathbf{S}(\epsilon, \delta, \eta) : \ Z_n \ \text{ has } \Big(3(\bar{\rho} + \widetilde{\rho} + 3)\epsilon, \delta, \eta\Big)\text{−overflow}\}
\end{aligned}$$

and on event $(\mathbf{A}^{\times})^c$.

Define geometric random variables with the following success rates

$$\begin{aligned}
U_1(\epsilon, \delta, \eta) &\sim \text{Geom}\Big(p\big(3(\bar{\rho} + \widetilde{\rho} + 3)\epsilon, \delta, \eta\big)\Big) \\
U_2(\epsilon, \delta, \eta) &\sim \text{Geom}\Big(H(\delta/\eta)p\big(3(\bar{\rho} + \widetilde{\rho} + 3)\epsilon, \delta, \eta\big)\Big).
\end{aligned}$$

Using results above to bound I_{before} and I_{after} separately on event $(\mathbf{A}^{\times})^c$, we can show that (for all η sufficiently small and any $u > 0$)

$$\begin{aligned}
&\sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(v^{\Omega}\lambda(\eta)\sigma(\eta) > u, J_\sigma(\eta) = k'\Big) \\
\leq &\sup_{|x|\leq 2\epsilon} \mathbb{P}_x(\mathbf{A}^{\times}(\epsilon, \delta, \eta)) + \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)\sigma(\eta) > u, J_\sigma(\eta) = k'\} \cap \big(\mathbf{A}^{\times}(\epsilon, \delta, \eta)\big)^c\Big) \\
\leq &C + \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)\sigma(\eta) > u, J_\sigma(\eta) = k'\} \cap \big(\mathbf{A}^{\times}(\epsilon, \delta, \eta)\big)^c\Big) \quad \text{due to (B.102)} \\
\leq &C + \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)I_{\text{before}}(\epsilon, \delta, \eta) > (1-C)u, J_\sigma(\eta) = k'\} \cap \big(\mathbf{A}^{\times}(\epsilon, \delta, \eta)\big)^c\Big) \\
&\quad + \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)I_{\text{after}}(\epsilon, \delta, \eta) > Cu\} \cap \big(\mathbf{A}^{\times}(\epsilon, \delta, \eta)\big)^c\Big) \\
\leq &C + \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)I_{\text{before}}(\epsilon, \delta, \eta) > (1-C)u, J_\sigma(\eta) = k'\} \cap \big(\mathbf{A}^{\times}(\epsilon, \delta, \eta)\big)^c\Big) \\
&+ \mathbb{P}\Big(v^{\Omega}\lambda(\eta)\frac{\rho(\epsilon) + 2l^*\hat{t}(\epsilon)}{\eta} \cdot U_1(\epsilon, \delta, \eta) > Cu\Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq C + \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\Big(\{v^\Omega\lambda(\eta)I_{\text{before}}(\epsilon,\delta,\eta) > (1-C)u, J_\sigma(\eta)=k'\}\cap\big(\mathbf{A}^\times(\epsilon,\delta,\eta)\big)^c\cap\{J^\downarrow = J^\uparrow\}\Big)\\
&+ \sup_{x\in[-2\epsilon,2\epsilon]} \mathbb{P}_x\big((\mathbf{A}^\times)^c\cap\{J^\downarrow < J^\uparrow\}\big) + \mathbb{P}\Big(v^\Omega\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}\cdot U_1(\epsilon,\delta,\eta) > Cu\Big)\\
&\leq 2C + \mathbb{P}\Big(v^\Omega\lambda(\eta)U_2(\epsilon,\delta,\eta) > (1-C)u\Big)\frac{\nu^\Omega_{k'}+C}{\nu^\Omega} + \mathbb{P}\Big(v^\Omega\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}\cdot U_1(\epsilon,\delta,\eta) > Cu\Big)
\end{aligned}
\tag{B.107}
$$

where the last inequality follows from (B.104) and (B.105). Now let us analyze the probability terms on the last row of the display above. For the first term, let $a(\eta) = H(\delta/\eta)p\Big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\Big)$. Due to Lemma B.14, we have (recall that $\nu^\Omega = \mu(E(0))$)

$$
\lim_{\eta\downarrow 0}\frac{a(\eta)}{\lambda(\eta)\mu\big(E\big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon\big)\big)} = 1.
$$

Combining this with (B.97), one can see that for all η sufficiently small,

$$
\mathbb{P}\Big(v^\Omega\lambda(\eta)U_2(\epsilon,\delta,\eta) > (1-C)u\Big) \leq \mathbb{P}\Big(a(\eta)\text{Geom}\big(a(\eta)\big) > (1-C)^2u\Big)\ \forall u > 0.
$$

Next, let $b(\eta,u) = \mathbb{P}\Big(a(\eta)\text{Geom}\big(a(\eta)\big) > (1-C)^2u\Big) = \mathbb{P}\Big(\text{Geom}\big(a(\eta)\big) > \frac{(1-C)^2u}{a(\eta)}\Big)$. For $g(y) = \log(1-y)$, we know the existence of some $y_0 > 0$ such that for all $y\in(0,y_0)$, we have $\log(1-y)\leq -(1-C)y$. So one can see that for all η sufficiently small,

$$
\begin{aligned}
&\log b(u,\eta) \leq \frac{(1-C)^2u}{a(\eta)}\log(1-a(\eta)) \leq -(1-C)^3u\\
&\Rightarrow b(u,\eta) \leq \exp\big(-(1-C)^3u\big)
\end{aligned}
\tag{B.108}
$$

uniformly for all $u > 0$.

For the second probability term, if we only consider $u\geq C$, then

$$
\mathbb{P}\Big(v^\Omega\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}\cdot U_1(\epsilon,\delta,\eta) > Cu\Big) \leq \mathbb{P}\Big(v^\Omega\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}\cdot U_1(\epsilon,\delta,\eta) > C^2\Big).
$$

Using $H\in\mathcal{RV}_{-\alpha}(\eta)$ with $\alpha > 1$, we get

$$
p\Big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\Big)U_1(\epsilon,\delta,\eta) \xrightarrow{d} \text{Exp}(1) \quad \text{as } \eta\downarrow 0
$$

due to the nature of the Geometric random variable U_1. Besides, due to $H\in\mathcal{RV}_{-\alpha}(\eta)$ with $\alpha > 1$ and Lemma B.14, it is easy to show that

$$
\lim_{\eta\downarrow 0}\frac{\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}}{p\Big(3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\Big)} = 0.
$$

Combining these results with Slutsky's theorem, we now obtain

$$
\mu(E(0))\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}\cdot U_1(\epsilon,\delta,\eta) \xrightarrow{d} 0 \quad \text{as } \eta\downarrow 0.
$$

Therefore,

$$
\limsup_{\eta\downarrow 0}\sup_{u\geq C}\mathbb{P}\Big(\mu(E(0))\lambda(\eta)\frac{\rho(\epsilon)+2l^*\hat{t}(\epsilon)}{\eta}\cdot U_1(\delta,\eta) > Cu\Big) = 0. \tag{B.109}
$$

Plugging (B.108) and (B.109) back into (B.107), we can establish the upper bound in (B.67). To show (B.69), note that for event

$$E(\epsilon,\eta)=\{\nu^{\Omega}\lambda(\eta)\tau(\eta,\epsilon)>u,\ X^{\eta}_{\tau(\eta,\epsilon)}\in B(\widetilde{m}_{k'},2\epsilon)\},$$

we have (for definitions of τ, see (B.62))

$$\begin{aligned}E(\epsilon,\eta)\supseteq&\{v^{\Omega}\lambda(\eta)\sigma(\eta)>u,J_{\sigma}(\eta)=k'\}\cap\{X^{\eta}_{n}\in\widetilde{\Omega}_{J_{\sigma}(\eta)}\ \ \forall n\in[\sigma(\eta),\tau(\eta,\epsilon)]\},\\ E(\epsilon,\eta)\cap&\{v^{\Omega}\lambda(\eta)\sigma(\eta)>u,J_{\sigma}(\eta)=j\}\cap\{X^{\eta}_{n}\in\widetilde{\Omega}_{J_{\sigma}(\eta)}\ \ \forall n\in[\sigma(\eta),\tau(\eta,\epsilon)]\}=\emptyset\ \forall j\neq k'.\end{aligned}$$

Therefore, for all η sufficiently small,

$$\begin{aligned}&\sup_{|x|\leq 2\epsilon}\mathbb{P}_x(E(\epsilon,\eta))\\ \leq&\sup_{|x|\leq 2\epsilon}\mathbb{P}_x(\mathbf{A}^{\times})+\sup_{|x|\leq 2\epsilon}\mathbb{P}_x\big((\mathbf{A}^{\times})^c\cap\{X^{\eta}_{n}\notin\widetilde{\Omega}_{J_{\sigma}(\eta)}\text{ for some }n\in[\sigma(\eta),\tau(\eta,\epsilon)]\}\big)\\ +&\sup_{|x|\leq 2\epsilon}\mathbb{P}_x\Big((\mathbf{A}^{\times})^c\cap\{v^{\Omega}\lambda(\eta)\sigma(\eta)>u,J_{\sigma}(\eta)=k'\}\cap\{X^{\eta}_{n}\in\widetilde{\Omega}_{J_{\sigma}(\eta)}\ \ \forall n\in[\sigma(\eta),\tau(\eta,\epsilon)]\}\Big)\\ \leq&4C+\exp\big(-(1-C)^3u\big)\frac{\nu^{\Omega}_{k'}+C}{\nu^{\Omega}}\end{aligned}$$

uniformly for all $u\geq C$, due to (B.102), (B.71) and (B.107).

The lower bound can be shown by an almost identical approach. In particular, analogous to (B.107), we can show that (for any $u>0$)

$$\begin{aligned}&\inf_{x\in[-2\epsilon,2\epsilon]}\mathbb{P}_x\Big(v^{\Omega}\lambda(\eta)\sigma(\eta)>u,J_{\sigma}(\eta)=k'\Big)\\ \geq&\inf_{x\in[-2\epsilon,2\epsilon]}\mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)\sigma(\eta)>u,J_{\sigma}(\eta)=k'\}\cap\big(\mathbf{A}^{\times}(\epsilon,\delta,\eta)\big)^c\Big)\\ \geq&\inf_{x\in[-2\epsilon,2\epsilon]}\mathbb{P}_x\Big(\{v^{\Omega}\lambda(\eta)I_{\text{before}}(\epsilon,\delta,\eta)>u,J_{\sigma}(\eta)=k'\}\cap\big(\mathbf{A}^{\times}(\epsilon,\delta,\eta)\big)^c\Big)\\ \geq&\mathbb{P}\Big(v^{\Omega}\lambda(\eta)U_2'(\epsilon,\delta,\eta)>(1-C)u\Big)\frac{\nu^{\Omega}_{k'}+C}{\nu^{\Omega}}-2C\text{ due to }\mathbb{P}(E\backslash F)\geq\mathbb{P}(E)-\mathbb{P}(F)\text{ and (B.102)(B.104)(B.106)}\end{aligned}$$

where

$$U_2'(\epsilon,\delta,\eta)\sim\text{Geom}\Big(H(\delta/\eta)p\big(-3(\bar{\rho}+\widetilde{\rho}+3)\epsilon,\delta,\eta\big)\Big).$$

Using the similar argument leading to (B.108), we are able to show (B.68), (B.70) and conclude the proof. □

Recall that $\sigma_i(\eta)=\min\{n\geq 0:X_n\notin\Omega_i\}$ and that value of constants $q_i,q_{i,j}$ are specified via (5)-(10). Define

$$\tau_i^{\min}(\eta,\epsilon)\triangleq\min\{n\geq\sigma_i(\eta):X^{\eta}_{n}\in\bigcup_{j}[m_j-2\epsilon,m_j+2\epsilon]\},\tag{B.110}$$

$$J_i(\eta)=j\iff X^{\eta}_{\sigma_i(\eta)}\in\Omega_j\ \forall j\in[n_{\min}].\tag{B.111}$$

The following result is simply a restatement of Proposition B.22 under the new system of notations. Despite the reiteration, we still state it here because this is the version that will be used to prove Lemma 8, which is the key tool for establishing Theorem 1, as well as many other results in Appendix C.

Proposition B.24. *Given $C > 0$ and $i, j \in [n_{min}]$ such that $i \neq j$, the following claims hold for all $\epsilon > 0$ that are sufficiently small:*

$$\limsup_{\eta\downarrow 0} \sup_{u\in(C,\infty)} \sup_{x\in(m_i-2\epsilon,m_i+2\epsilon)} \mathbb{P}_x\Big(q_i\lambda_i(\eta)\sigma_i(\eta) > u,\ X^\eta_{\sigma_i(\eta)} \in \Omega_j\Big) \le C + \exp\big(-(1-C)u\big)\frac{q_{i,j}+C}{q_i},$$

$$\liminf_{\eta\downarrow 0} \inf_{u\in(C,\infty)} \inf_{x\in(m_i-2\epsilon,m_i+2\epsilon)} \mathbb{P}_x\Big(q_i\lambda_i(\eta)\sigma_i(\eta) > u,\ X^\eta_{\sigma_i(\eta)} \in \Omega_j\Big) \ge -C + \exp\big(-(1+C)u\big)\frac{q_{i,j}-C}{q_i},$$

$$\limsup_{\eta\downarrow 0} \sup_{u\in(C,\infty)} \sup_{x\in(m_i-2\epsilon,m_i+2\epsilon)} \mathbb{P}_x\Big(q_i\lambda_i(\eta)\tau_i^{min}(\eta,\epsilon) > u,\ X^\eta_{\tau_i^{min}(\eta,\epsilon)} \in \Omega_j\Big) \le C + \exp\big(-(1-C)u\big)\frac{q_{i,j}+C}{q_i},$$

$$\liminf_{\eta\downarrow 0} \inf_{u\in(C,\infty)} \inf_{x\in(m_i-2\epsilon,m_i+2\epsilon)} \mathbb{P}_x\Big(q_i\lambda_i(\eta)\tau_i^{min}(\eta,\epsilon) > u,\ X^\eta_{\tau_i^{min}(\eta,\epsilon)} \in \Omega_j\Big) \ge -C + \exp\big(-(1+C)u\big)\frac{q_{i,j}-C}{q_i},$$

$$\liminf_{\eta\downarrow 0} \inf_{x\in(m_i-2\epsilon,m_i+2\epsilon)} \mathbb{P}_x\Big(q_i\lambda_i(\eta)\big(\tau_i^{min}(\eta,\epsilon)-\sigma_i(\eta)\big) < C,\ X^\eta_n \in \Omega_{J_i(\eta)}\ \forall n \in [\sigma_i(\eta), \tau_i^{min}(\eta,\epsilon)]\Big) \ge 1 - C.$$

Concluding this section, we apply Proposition B.24 and prove Lemma 8.

Proof of Lemma 8. Fix some $C \in (0,1)$, $u > 0$, and some $k, l \in [n_{\min}]$ with $k \neq l$. Let $q_i, q_{i,j}$ be the constants defined in (10).

Fix some $C_0 \in \big(0, \frac{C}{n_{\min}} \wedge \frac{q_k}{n_{\min}} C\big)$. Using Proposition B.24, we know that for all ϵ sufficiently small, we have

$$\limsup_{\eta\downarrow 0} \sup_{x\in(m_k-2\epsilon,m_k+2\epsilon)} \mathbb{P}_x\Big(q_k\lambda_k(\eta)\sigma_k(\eta) > u,\ X^\eta_{\sigma_k(\eta)} \in \Omega_j\Big) \le C_0 + \exp\big(-(1-C)u\big)\frac{q_{k,j}+C_0}{q_k}\ \forall j \in [n_{\min}].$$

Summing up the inequality above over all $j \in [n_{\min}]$, we can obtain (17). The lower bound (18) can be established using an identical approach.

In order to show (20), note that we can find $C_1 \in (0, u)$ sufficiently small so that

$$-C_1 + \exp\big(-(1+C_1)\cdot 2C_1\big)\frac{q_{k,l}-C_1}{q_k} \ge \frac{q_{k,l}-C}{q_k}.$$

Fix such C_1. From Proposition B.24, we also know that for all ϵ small enough, we have

$$\liminf_{\eta\downarrow 0} \inf_{x\in(m_k-2\epsilon,m_k+2\epsilon)} \mathbb{P}_x\Big(q_k\lambda_k(\eta)\sigma_k(\eta) > u, X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big) \ge -C_1 + \exp\big(-(1+C_1)\cdot 2C_1\big)\frac{q_{k,l}-C_1}{q_k}.$$

Then using $\mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big) \ge \mathbb{P}\Big(q_k\lambda_k(\eta)\sigma_k(\eta) > u, X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big)$ we conclude the proof for (20).

Moving on, we show (19) in the following way. Note that we can find $C_2 \in (0, u)$ small enough so that

$$2C_2 + \frac{q_{k,l}+C_2}{q_k} < \frac{q_{k,l}+C}{q_k}. \tag{B.112}$$

Fix such C_2. Since (18) has been established already, we can find some $u_2 > 0$ such that for all ϵ small enough,

$$\limsup_{\eta\downarrow 0} \sup_{x\in(m_k-2\epsilon,m_k+2\epsilon)} \mathbb{P}_x\Big(q_k\lambda_k(\eta)\sigma_k(\eta) \le u_2\Big) < C_2 \tag{B.113}$$

Fix such u_2. Meanwhile, fix some $C_3 \in (0, C_2 \wedge u_2)$. From Proposition B.24 we know that for all ϵ sufficiently small,

$$\limsup_{\eta\downarrow 0} \sup_{x\in(m_k-2\epsilon,m_k+2\epsilon)} \mathbb{P}_x\Big(q_k\lambda_k(\eta)\sigma_k(\eta) > u_2,\ X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big) \le C_3 + \exp\big(-(1-C_3)u_2\big)\frac{q_{k,l}+C_3}{q_k}$$

$$\leq C_2 + \frac{q_{k,l} + C_2}{q_k}. \tag{B.114}$$

Lastly, observe the following decomposition of events (for any $x \in \Omega_k$)

$$\mathbb{P}_x\Big(X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big) \leq \mathbb{P}_x\Big(q_k\lambda_k(\eta)\sigma_k(\eta) \leq u_2\Big) + \mathbb{P}_x\Big(q_k\lambda_k(\eta)\sigma_k(\eta) > u_2,\ X^\eta_{\sigma_k(\eta)} \in \Omega_l\Big).$$

Combining this bound with (B.112)-(B.114), we complete the proof. □

C Proofs for Technical Lemmas in Section 2.6

C.1 Proof of Lemma 9

First, we introduce another dichotomy for *small* and *large* noises. For any $\widetilde{\gamma} > 0$ and any learning rate $\eta > 0$, we say that a noise Z_n is small if

$$\eta|Z_n| > \eta^{\widetilde{\gamma}}$$

and we say Z_n is large otherwise. For this new classification of small and large noises, we introduce the following notations and definitions:

$$Z_n^{\leq,\widetilde{\gamma},\eta} = Z_n \mathbb{1}\{\eta|Z_n| \leq \eta^{\widetilde{\gamma}}\}, \tag{C.1}$$

$$Z_n^{>,\widetilde{\gamma},\eta} = Z_n \mathbb{1}\{\eta|Z_n| > \eta^{\widetilde{\gamma}}\}, \tag{C.2}$$

$$\widetilde{T}_1^\eta(\widetilde{\gamma}) \triangleq \min\{n \geq 1 :\ \eta|Z_n| > \eta^{\widetilde{\gamma}}\}. \tag{C.3}$$

Similar to Lemma B.10, the following result is a direct application of Lemma B.9, and shows that it is rather unlikely to observe large perturbation that are caused only by *small* noises. Specifically, since $\alpha > 1$ we can always find

$$\widetilde{\gamma} \in (0, (1 - \frac{1}{\alpha} \wedge \frac{1}{2})$$
$$\beta \in \big(1, (2 - 2\widetilde{\gamma}) \wedge \alpha(1 - \widetilde{\gamma})\big).$$

Now in Lemma B.9, if we let $\Delta = \widetilde{\gamma}, \widetilde{\Delta} = \Delta/2$ and $\epsilon = \delta = 1$ (in other words, $u(\eta) = 1/\eta^{1-\widetilde{\gamma}},\ v(\eta) = \eta^{\widetilde{\gamma}/2}$), then for any positive integer j the condition (B.38) is satisfied, allowing us to draw the following conclusion immediately as a corollary from Lemma B.9.

Lemma C.1. *Given $N > 0$ and*

$$\widetilde{\gamma} \in (0, (1 - \frac{1}{\alpha}) \wedge \frac{1}{2}),\ \beta \in \big(1, (2 - 2\widetilde{\gamma}) \wedge (\alpha - \alpha\widetilde{\gamma})\big),$$

we have (as $\eta \downarrow 0$)

$$\mathbb{P}\Big(\max_{j=1,2,\cdots,\lceil(1/\eta)^\beta\rceil} \eta|Z_1^{\leq,\widetilde{\gamma},\eta} + \cdots + Z_j^{\leq,\widetilde{\gamma},\eta}| > \eta^{\widetilde{\gamma}/2}\Big) = o(\eta^N).$$

The flavor of the next lemma is similar to that of Lemma B.11. Specifically, we show that, with high probability, the SGD iterates would quickly return to the local minimum as long as they start from somewhere that are not too close the boundary of an attraction field (namely, the points $s_1, s_2, \cdots, s_{n_{\min}}$). To this end, we consider a refinement of function $\hat{t}(\cdot)$ defined in (B.24). For any $i = 1, 2, \cdots, n_{\min}$, any $x \in \Omega_i$ and any $\eta > 0, \gamma \in (0, 1)$, we can define the return time to $\eta^\gamma-$neighborhood for the ODE $\mathbf{x}^\eta$ as

$$\hat{t}_\gamma^{(i)}(x, \eta) \triangleq \min\{t \geq 0 : |\mathbf{x}^\eta(t, x) - m_i| \leq \eta^\gamma\}.$$

Given the bound in (B.23) (which is stated for a specific attraction field) and the fact that there only exists finitely many attraction fields, we know the existence of some $c_2 < \infty$ such that for any $i = 1, 2, \cdots, n_{\min}$, any $\eta > 0$, any $\gamma \in (0, 1)$ and any $x \in \Omega_i$ such that $|x - s_i| \vee |x - s_{i-1}| > \eta^\gamma$, we have

$$\hat{t}_\gamma^{(i)}(x, \eta) \leq c_2 \gamma \log(1/\eta)/\eta$$

and define function $t^\uparrow$ as

$$t^\uparrow(\eta, \gamma) \triangleq c_2 \gamma \log(1/\eta).$$

Lastly, define the following stopping time for any $i = 1, 2, \cdots, n_{\min}$, any $x \in \Omega_i$ and any $\Delta > 0$

$$T_{\text{return}}^{(i)}(\eta, \Delta) \triangleq \min\{n \geq 0 : X_n^\eta(x) \in B(m_i, 2\Delta)\}$$

where we adopt the notation $B(u, v) \triangleq [u - v, u + v]$ for the $v-$neighborhood around point u.

Lemma C.2. *Given*

$$\widetilde{\gamma} \in (0, (1 - \frac{1}{\alpha}) \wedge (\frac{1}{2})), \ \gamma \in (0, \frac{\widetilde{\gamma}}{16 M c_2} \wedge \frac{\widetilde{\gamma}}{4}),$$

and any $i = 1, 2, \cdots, n_{min}$, any $\Delta > 0$, we have

$$\liminf_{\eta \downarrow 0} \inf_{x \in \Omega_i : |x - s_{i-1}| \vee |x - s_i| \geq 2\eta^\gamma} \mathbb{P}_x \Big(T_{return}^{(i)}(\eta, \Delta) \leq 2c_2 \gamma \log(1/\eta)/\eta, \ X_n^\eta \in \Omega_i \ \forall n \leq T_{return}^{(i)}(\eta, \Delta) \Big) = 1.$$

Proof. Throughout this proof, we only consider η small enough such that

$$2c_2 \gamma \log(1/\eta)/\eta < \lceil (1/\eta)^\beta \rceil, \quad \eta M \leq \eta^{2\widetilde{\gamma}}, \quad 2\eta^\gamma < \Delta/2, \quad 2\eta^{\widetilde{\gamma}/4} < \eta^\gamma. \tag{C.4}$$

The condition above holds for all $\eta > 0$ sufficiently small because $\beta > 1$, $2\widetilde{\gamma} < 1$, and $\gamma < \widetilde{\gamma}/4$. Also, fix some $\beta \in \big(1, (2 - 2\widetilde{\gamma}) \wedge (\alpha - \alpha\widetilde{\gamma})\big)$

Define the following events

$$A_1^\times(\eta) \triangleq \Big\{ \max_{j = 1, 2, \cdots, \lceil (1/\eta)^\beta \rceil} \eta | Z_1^{\leq, \widetilde{\gamma}, \eta} + \cdots + Z_j^{\leq, \widetilde{\gamma}, \eta} | > \eta^{\widetilde{\gamma}/2} \Big\}$$

$$A_2^\times(\eta) \triangleq \{ \widetilde{T}_1^\eta(\widetilde{\gamma}) \leq \lceil (1/\eta)^\beta \rceil \} \quad \text{(see (C.3) for definition of the stopping time involved)}$$

and fix some $N > 0$. From Lemma C.1, we see that (for all sufficiently small η)

$$\mathbb{P}(A_1^\times(\eta)) \leq \eta^N.$$

Besides, using Lemma B.2 together with the fact that $\beta < \alpha(1 - \widetilde{\gamma})$, we know the existence of some constant $\theta > 0$ such that

$$\mathbb{P}(A_2^\times(\eta)) \leq \eta^\theta$$

for all sufficiently small η.

Now we focus on the behavior of the SGD iterates on event $\Big(A_1^\times(\eta) \cap A_2^\times(\eta) \Big)^c$. Let us arbitrarily choose some $x \in \Omega_i$ such that $|x - s_i| \vee |x - s_{i-1}| > 2\eta^\gamma$. First, from Lemma B.4 and (C.4), we know that

$$|\mathbf{x}_t^\eta(x) - \mathbf{y}_{\lfloor t \rfloor}^\eta(x)| \leq 2\eta M \exp\big(2Mc_2 \gamma \log(1/\eta)\big) \leq 2\eta^{2\widetilde{\gamma} - 2Mc_2\gamma} \leq 2\eta^{\widetilde{\gamma}} \leq \eta^\gamma \quad \forall t \leq 2c_2 \gamma \log(1/\eta)/\eta \tag{C.5}$$

Next, from the definition of the function $t^{\uparrow}(\cdot)$ and (C.5), we know that for

$$T_{\mathrm{GD,return}}(x;\eta,\Delta) \triangleq \min\{n \geq 0 : \mathbf{y}_n^\eta(x) \in B(m_i, \frac{\Delta}{2} + \eta^\gamma)\}, \tag{C.6}$$

we have

$$T_{\mathrm{GD,return}}(x;\eta,\Delta) \leq 2c_2\gamma\log(1/\eta)/\eta \tag{C.7}$$

$$\mathbf{y}_n^\eta > s_{i-1} + \eta^\gamma,\ \mathbf{y}_n^\eta < s_i - \eta^\gamma \quad \forall n \leq 2c_2\gamma\log(1/\eta)/\eta. \tag{C.8}$$

Furthermore, on event $\Big(A_1^\times(\eta) \cap A_2^\times(\eta)\Big)^c$, due to Lemma B.3 and (C.4), we have that

$$|X_n^\eta(x) - \mathbf{y}_n^\eta(x)| \leq \eta^{\widetilde{\gamma}/2}\exp\big(2Mc_2\log(1/\eta)\big) = \eta^{\frac{\widetilde{\gamma}}{2}-2Mc_2\gamma} \leq \eta^{\widetilde{\gamma}/4} < \eta^\gamma \quad \forall n \leq 2c_2\gamma\log(1/\eta)/\eta.$$

Combining this with (C.7), (C.8), we can conclude that (recall that due to (C.4) we have $2\eta^\gamma < \Delta/2$)

$$\begin{aligned} T_{\mathrm{return}}^{(i)}(\eta,\Delta) &\leq 2c_2\gamma\log(1/\eta)/\eta \\ X_n^\eta &\in \Omega_i \quad \forall n \leq 2c_2\gamma\log(1/\eta)/\eta \end{aligned}$$

on event $\Big(A_1^\times(\eta) \cap A_2^\times(\eta)\Big)^c$. Therefore,

$$\begin{aligned} &\liminf_{\eta\downarrow 0} \inf_{x\in\Omega_i : |x-s_{i-1}|\vee|x-s_i|\geq 2\eta^\gamma} \mathbb{P}_x\Big(T_{\mathrm{return}}^{(i)}(\eta,\Delta) \leq 2c_2\gamma\log(1/\eta)/\eta,\ X_n^\eta \in \Omega_i\ \forall n \leq T_{\mathrm{return}}^{(i)}(\eta,\Delta)\Big) \\ &\geq \liminf_{\eta\downarrow 0} \mathbb{P}\Big(\Big(A_1^\times(\eta) \cap A_2^\times(\eta)\Big)^c\Big) \geq \liminf_{\eta\downarrow 0} 1 - \eta^N - \eta^\theta = 1. \end{aligned}$$

This concludes the proof. □

The takeaway of the next lemma is that, almost always, the SGD iterates will quickly escape from the neighborhood of any s_i, the boundaries of each attraction fields.

Lemma C.3. *Given any $\gamma \in (0,1), t > 0$, we have*

$$\liminf_{\eta\downarrow 0} \inf_{x\in[-L,L]} \mathbb{P}_x\Big(\min\big\{n \geq 0 : X_n^\eta \notin \bigcup_i B(s_i, 2\eta^\gamma)\big\} \leq \frac{t}{H(1/\eta)}\Big) = 1.$$

Proof. We only consider η small enough so that

$$\begin{aligned} \min_{i=2,3,\cdots,n_{\min}-1} |s_i - s_{i-1}| &> 3\eta^{\frac{1+\gamma}{2}}, \\ \eta M &< \eta^\gamma. \end{aligned}$$

Also, the claim is trivial if $x \notin \cup_j B(s_j, 2\eta^\gamma)$, so without loss of generality we only consider the case where there is some $j \in [n_{\min}]$ and $x \in [-L, L], x \in B(s_j, 2\eta^\gamma)$. Let us define stopping times

$$T^\gamma \triangleq \min\{n \geq 1 : \eta|Z_n| > 5\eta^\gamma\}; \tag{C.9}$$

$$T_{\mathrm{escape}}^\gamma \triangleq \min\{n \geq 0 : X_n^\eta \notin \cup_j B(s_j, 2\eta^\gamma)\}, \tag{C.10}$$

and the following two events

$$\begin{aligned} A_1^\times(\eta) &\triangleq \{T^\gamma > \frac{t}{H(1/\eta)}\}, \\ A_2^\times(\eta) &\triangleq \{\eta|Z_{T^\gamma}| > \eta^{\frac{1+\gamma}{2}}\}. \end{aligned}$$

First, using Lemma B.1 and the regularly varying nature of $H(\cdot) = \mathbb{P}(|Z_1| > \cdot)$, we know the existence of some $\theta > 0$ such that

$$\mathbb{P}(A_1^{\times}(\eta)) \leq \exp(-1/\eta^{\theta})$$

for all $\eta > 0$ sufficiently small. Next, by definition of T^{γ}, one can see that (for any $\eta \in (0,1)$)

$$\mathbb{P}(A_2^{\times}(\eta)) = \frac{H(1/\eta^{\frac{1-\gamma}{2}})}{H(5/\eta^{1-\gamma})}.$$

Again, due to $H \in \mathcal{RV}_{-\alpha}$ and $1-\gamma > 0$, we know the existence of some $\theta_1 > 0$ such that

$$\mathbb{P}(A_2^{\times}(\eta)) < \eta^{\theta_1}$$

for all $\eta > 0$ sufficiently small. To conclude the proof, we only need to note the following fact on event $\big(A_1^{\times}(\eta) \cup A_2^{\times}(\eta)\big)^c$. There are only two possibilities on this event: $T^{\gamma}_{\text{escape}} \leq T^{\gamma} - 1$, or $T^{\gamma}_{\text{escape}} \geq T^{\gamma}$. Now we analyze the two cases respectively.

- On $\big(A_1^{\times}(\eta) \cup A_2^{\times}(\eta)\big)^c \cap \{T^{\gamma}_{\text{escape}} \leq T^{\gamma} - 1\}$, we must have $T^{\gamma}_{\text{escape}} < T^{\gamma} \leq t/H(1/\eta)$.
- On $\big(A_1^{\times}(\eta) \cup A_2^{\times}(\eta)\big)^c \cap \{T^{\gamma}_{\text{escape}} \geq T^{\gamma}\}$, we know that at $n = T^{\gamma} - 1$, there exists an integer $j \in \{1, 2, \cdots, n_{\min} - 1\}$ such that $X_n^{\eta} \in B(s_j, 2\eta^{\gamma})$. Now since $\eta M < \eta^{\gamma}$ and $\eta |Z_{T^{\gamma}}| > 5\eta^{\gamma}$, we must have

  $$|X^{\eta}_{T^{\gamma}} - X^{\eta}_{T^{\gamma}-1}| > 4\eta^{\gamma} \Rightarrow X^{\eta}_{T^{\gamma}} \notin B(s_j, 2\eta^{\gamma}).$$

  On the other hand, the exclusion of event $A_2^{\times}(\eta)$ tells us that $|X^{\eta}_{T^{\gamma}} - X^{\eta}_{T^{\gamma}-1}| < 2\eta^{\frac{1+\gamma}{2}}$. Due to (C.9), we then have $X_n^{\eta} \notin \cup_i B(s_i, 2\eta^{\gamma})$.

In summary, $\big(A_1^{\times}(\eta) \cup A_2^{\times}(\eta)\big)^c \subseteq \{T^{\gamma}_{\text{return}} \leq t/H(1/\eta)\}$ and this conclude the proof. □

In the next lemma, we analyze the number of transitions needed to visit a certain local minimizer in the loss landscape. In general, we focus on a communication class G and, for now, assume it is absorbing. Next, we introduce the following concepts to record the transitions between different local minimum. To be specific, for any $\eta > 0$ and any $\Delta > 0$ small enough so that $B(m_j, \Delta) \cap \Omega_j^c = \emptyset$ for all j, define

$$T_0(\eta, \Delta) = \min\{n \geq 0 : \ X_n^{\eta} \in \cup_j B(m_j, 2\Delta)\}; \tag{C.11}$$

$$I_0(\eta, \Delta) = j \text{ iff } X^{\eta}_{T_0(\eta,\Delta)} \in B(m_j, 2\Delta); \tag{C.12}$$

$$T_k(\eta, \Delta) = \min\{n > T_{k-1}(\eta, \Delta) : \ X_n^{\eta} \in \cup_{j \neq I_{k-1}(\eta,\Delta)} B(m_j, 2\Delta)\} \ \ \forall k \geq 1 \tag{C.13}$$

$$I_k(\eta, \Delta) = j \text{ iff } X^{\eta}_{T_k(\eta,\Delta)} \in B(m_j, 2\Delta) \ \ \forall k \geq 1. \tag{C.14}$$

As mentioned earlier, the next goal is to analyze the transitions between attraction fields it takes to visit m_j when starting from m_i when $m_i, m_j \in G$. Define

$$K_i(\eta, \Delta) \triangleq \min\{k \geq 0 : \ I_k(\eta, \Delta) = i\}.$$

Lemma C.4. *Assume that G is an absorbing communication class on the graph $\mathcal{G}$. Then there exists some constant $p > 0$ such that for any i with $m_i \in G$, any $\epsilon > 0$, and any $\Delta > 0$,*

$$\sup_{j:\ m_j \in G;\ x \in B(m_j, 2\Delta)} \mathbb{P}_x(K_i(\eta, \Delta) > u \cdot n_{min}) \leq \mathbb{P}(Geom(p) \geq u) + \epsilon \ \ \forall u = 1, 2, \cdots,$$

$$\sup_{j:\ m_j \in G;\ x \in B(m_j, 2\Delta)} \mathbb{P}_x\Big(\exists k \in [K_i(\eta, \Delta)] \ s.t. \ m_{I_k(\eta,\Delta)} \notin G\Big) \leq \epsilon$$

hold for all $\eta > 0$ sufficiently small.

Proof. The claim is trivial if, for the initial condition, we have $x \in B(m_i, 2\Delta)$. Next, let us observe the following facts.

- Define (recall the definitions of measure μ_i and sets $E_i, E_{i,j}$ in (4)(8)(9))

$$J(j) \triangleq \arg \min_{\widetilde{j}:\mu_i(E_{j,\widetilde{j}})>0} |i - \widetilde{j}| \quad \forall j \neq i$$

$$p^* \triangleq \min_{j:\ j\neq i,\ m_j \in G} \frac{\mu_j(E_{j,J(j)})}{\mu_j(E_j)}.$$

- From the definition of $J(j)$ and the fact that there are only finitely many attraction fields we can see that $p^* > 0$. Moreover, G being a communication class implies that

$$|J(j) - i| < |j - i| \quad \forall j \neq i,\ m_j \in G.$$

  Indeed, if $i < j$, then since G is a communication class and there are some $m_i \in G$ with $i < j$, we will at least have $\mu_j(E_{j,j-1}) > 0$, so $|J(j) - i| \leq |i - j| - 1$; the case that $i > j$ can be approached analogously.

- Now from the definition of $J(j)$ and Proposition B.24, together with the previous bullet point, we know that for all η sufficiently small,

$$\inf_{x\in[-L,L]} \mathbb{P}_x(|I_{k+1} - i| \leq |I_k - i| - 1,\ m_{I_{k+1}} \in G \mid K_i(\eta, \Delta) > k,\ m_{I_k} \in G) \geq p^*/2$$

  uniformly for all $k \geq 0$.

- Meanwhile, since $p^* > 0$, we are able to fix some $\delta > 0$ small enough such that

$$\frac{n_{\min}\delta}{(p^*/2)^{n_{\min}}} < \epsilon.$$

- On the other hand, for any $\widetilde{j}$ with $m_{\widetilde{j}} \notin G$, by definition of the typical transition graph we must have $\mu_j(E_{j,\widetilde{j}}) = 0$ for any j with $m_j \in G$. Then due to Proposition B.24 again, one can see that for all $\eta > 0$ that is sufficiently small,

$$\sup_{x\in[-L,L]} \mathbb{P}_x(m_{I_{k+1}} \notin G \mid K_i(\eta, \Delta) > k,\ m_{I_k} \in G) < \delta$$

  uniformly for all $k \geq 0$.

- Repeat this argument for $n_{\min}$ times, and we can see that for all η sufficiently small

$$\inf_{x\in[-L,L]} \mathbb{P}_x(K_i(\eta, \Delta) \leq k + n_{\min} \mid K_i(\eta, \Delta) > k,\ m_{I_k} \in G) \geq \left(\frac{p^*}{2}\right)^{n_{\min}}$$

$$\sup_{x\in[-L,L]} \mathbb{P}_x(\exists l \in [n_{\min}] \text{ s.t. } m_{I_{k+l}} \notin G \mid K_i(\eta, \Delta) > k,\ m_{I_k} \in G) \leq n_{\min}\delta$$

  uniformly for all $k \geq 1$.

- Lastly, to apply the bounds established above, we will make use of the following expression of several probabilities. For any j with $m_j \in G$ and $x \in B(m_j, 2\Delta)$ and any $u = 1, 2, \cdots$,

$$\begin{aligned}
&\mathbb{P}_x(\exists l \in [un_{\min}] \text{ s.t. } m_{I_l} \notin G,\ K_i(\eta, \Delta) > u \cdot n_{\min}) \\
&= \sum_{v=0}^{u-1} \mathbb{P}_x\Big(\exists k \in [n_{\min}] \text{ s.t. } m_{I_{k+vn_{\min}}} \notin G \ \Big|\ K_i(\eta, \Delta) > v \cdot n_{\min},\ m_{I_l} \in G\ \forall l \leq vn_{\min}\Big)
\end{aligned}$$

$$\cdot \prod_{w=0}^{v-1} \mathbb{P}_x\Big(K_i(\eta,\Delta) > (w+1)n_{\min},\ m_{I_{k+wn_{\min}}} \in G\ \forall k \in [n_{\min}] \ \Big|\ K_i(\eta,\Delta) > w\cdot n_{\min},\ m_{I_l} \in G\ \forall l \le wn_{\min}\Big)$$

$$\begin{aligned}
&\mathbb{P}_x(m_{I_l} \in G\ \forall l \in [un_{\min}],\ K_i(\eta,\Delta) > u\cdot n_{\min})\\
=&\prod_{v=0}^{u-1} \mathbb{P}\Big(K_i(\eta,\Delta) > (v+1)n_{\min},\ m_{I_{k+vn_{\min}}} \in G\ \forall k \in [n_{\min}] \ \Big|\ K_i(\eta,\Delta) > vn_{\min},\ m_{I_k} \in G\ \forall k \in [vn_{\min}]\Big)
\end{aligned}$$

In summary, now we can see that (for sufficiently small η)

$$\begin{aligned}
&\sup_{j:\ m_j\in G;\ x\in B(m_j,2\Delta)} \mathbb{P}_x\Big(\exists k \in [K_i(\eta,\Delta)]\ s.t.\ m_{I_k(\eta,\Delta)} \notin G\Big)\\
\le&\sum_{u=0}^{\infty} \sup_{j:\ m_j\in G;\ x\in B(m_j,2\Delta)} \mathbb{P}_x\Big(\exists v \in [n_{\min}] \text{ such that } m_{I_{v+un_{\min}}} \notin G,\ K_i(\eta,\Delta) > un_{\min}, m_{I_k} \in G\ \forall k \in [un_{\min}]\Big)\\
\le&\sum_{u=0}^{\infty} \sup_{j:\ m_j\in G;\ x\in B(m_j,2\Delta)} \mathbb{P}_x\Big(\exists v \in [n_{\min}] \text{ such that } m_{I_{v+un_{\min}}} \notin G \ \Big|\ K_i(\eta,\Delta) > un_{\min}, m_{I_k} \in G\ \forall k \in [un_{\min}]\Big)\\
\cdot&\prod_{v=0}^{u-1} \sup_{j:\ m_j\in G;\ x\in B(m_j,2\Delta)} \mathbb{P}_x\Big(K_i(\eta,\Delta) > (v+1)n_{\min}, m_{I_{k+vn_{\min}}} \in G\ \forall k \in [n_{\min}]\Big|K_i(\eta,\Delta) > vn_{\min}, m_{I_k} \in G\ \forall k \in [vn_{\min}]\Big)\\
\le&\sum_{u\ge 0} n_{\min}\delta\Big(1-(\frac{p^*}{2})^{n_{\min}}\Big)^{u-1} = \frac{n_{\min}\delta}{(p^*/2)^{n_{\min}}} \le \epsilon
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{j:m_j\in G,\ x\in B(m_j,2\Delta)} \mathbb{P}_x(K_i(\eta,\Delta) > u\cdot n_{\min})\\
\le&\sup_{j:m_j\in G,\ x\in B(m_j,2\Delta)} \mathbb{P}_x(\exists l \in [un_{\min}] \text{ s.t. } m_{I_l} \notin G,\ K_i(\eta,\Delta) > u\cdot n_{\min})\\
+&\sup_{j:m_j\in G,\ x\in B(m_j,2\Delta)} \mathbb{P}_x(m_{I_l} \in G\ \forall l \in [un_{\min}],\ K_i(\eta,\Delta) > u\cdot n_{\min})\\
\le&\sum_{v=1}^{u} n_{\min}\delta\Big(1-(\frac{p^*}{2})^{n_{\min}}\Big)^{v-1} + \Big(1-(\frac{p^*}{2})^{n_{\min}}\Big)^{u}\\
\le&\frac{n_{\min}\delta}{(p^*/2)^{n_{\min}}} + \Big(1-(\frac{p^*}{2})^{n_{\min}}\Big)^{u}\\
\le&\epsilon + \Big(1-(\frac{p^*}{2})^{n_{\min}}\Big)^{u}
\end{aligned}$$

uniformly for all $u = 1, 2, \cdots$. To conclude the proof, it suffices to set $p = (\frac{p^*}{2})^{n_{\min}}$. □

The proof above can be easily adapted to the case when the communication class G is transient. Define

$$K_i^G(\eta,\Delta) \triangleq \min\{k \ge 0 :\ I_k(\eta,\Delta) = i \text{ or } m_{I_k(\eta,\Delta)} \notin G\}.$$

Lemma C.5. *Assume that G is a transient communication class on the graph $\mathcal{G}$. Then there exists some constant $p > 0$ such that for any i with $m_i \in G$ and any $\Delta \in (0, \bar{\epsilon}/3)$,*

$$\sup_{j:\ m_j\in G;\ x\in B(m_j,2\Delta)} \mathbb{P}_x(K_i^G(\eta,\Delta) > u\cdot n_{min}) \le \mathbb{P}(Geom(p) \ge u)\ \ \forall u = 1, 2, \cdots \tag{C.15}$$

hold for all $\eta > 0$ sufficiently small.

Proof. The structure of this proof is analogous to that of Lemma C.4. Again, the claim is trivial if, for the initial condition, we have $x \in B(m_i, 2\Delta)$. Next, let us observe the following facts.

- Define (recall the definitions of measure μ_i and sets $E_i, E_{i,j}$ in (4)(8)(9))

$$J(j) \triangleq \arg \min_{\widetilde{j}: \mu_i(E_{j,\widetilde{j}})>0} |i - \widetilde{j}| \quad \forall j \neq i$$

$$p^* \triangleq \min_{j:\ j\neq i,\ m_j \in G} \frac{\mu_j(E_{j,J(j)})}{\mu_j(E_j)}.$$

- From the definition of $J(j)$ and the fact that there are only finitely many attraction fields we can see that $p^* > 0$. Moreover, G being a communication class implies that

$$|J(j) - i| < |j - i| \quad \forall j \neq i,\ m_j \in G.$$

  Indeed, if $i < j$, then since G is a communication class and there are some $m_i \in G$ with $i < j$, we will at least have $\mu_j(E_{j,j-1}) > 0$, so $|J(j) - i| \leq |i - j| - 1$; the case that $i > j$ can be approached analogously.

- Now from the definition of $J(j)$ and Proposition B.24, together with the previous bullet point, we know that for all η sufficiently small,

$$\inf_{x\in[-L,L]} \mathbb{P}_x(|I_{k+1} - i| \leq |I_k - i| - 1,\ m_{I_{k+1}} \in G \mid K_i^G(\eta, \Delta) > k \geq p^*/2$$

  uniformly for all $k \geq 0$.

- Repeat this argument for $n_{\min}$ times, and we can see that for all η sufficiently small

$$\inf_{x\in[-L,L]} \mathbb{P}_x(K_i^G(\eta, \Delta) \leq k + n_{\min} \mid K_i^G(\eta, \Delta) > k) \geq \Big(\frac{p^*}{2}\Big)^{n_{\min}}$$

  uniformly for all $k \geq 1$.

- Lastly, for any $j \neq i$ with $m_j \in G$ and $x \in B(m_j, 2\Delta)$ and any $u = 1, 2, \cdots$,

$$\begin{aligned}
&\mathbb{P}_x(K_i^G(\eta, \Delta) > u \cdot n_{\min}) \\
&= \prod_{v=0}^{u-1} \mathbb{P}_x\Big(K_i^G(\eta, \Delta) > (v+1)n_{\min} \ \Big|\ K_i^G(\eta, \Delta) > v \cdot n_{\min}\Big) \\
&= \prod_{v=0}^{u-1} \left(1 - \mathbb{P}_x\Big(K_i^G(\eta, \Delta) \leq (v+1)n_{\min} \ \Big|\ K_i^G(\eta, \Delta) > v \cdot n_{\min}\Big)\right)
\end{aligned}$$

In summary, now we can see that (for suffciently small η)

$$\sup_{j:m_j\in G,\ x\in B(m_j,2\Delta)} \mathbb{P}_x(K_i^G(\eta, \Delta) \geq u \cdot n_{\min}) \leq \Big(1 - (\frac{p^*}{2})^{n_{\min}}\Big)^u$$

uniformly for all $u = 1, 2, \cdots$. To conclude the proof, it suffices to set $p = (\frac{p^*}{2})^{n_{\min}}$. □

We are now ready to prove Lemma 9, which, as demonstrated earlier, is the key tool in proof of Theorem 2.

Proof of Lemma 9. The claim is trivial if $l^{\text{large}} = 1$, so we focus on the case where $l^{\text{large}} \geq 2$. Fix some

$$\widetilde{\gamma} \in (0, (1 - \frac{1}{\alpha}) \wedge (\frac{1}{2})),\ \beta \in \big(1, (2 - 2\widetilde{\gamma}) \wedge (\alpha - \alpha\widetilde{\gamma})\big),\ \gamma \in (0, \frac{\widetilde{\gamma}}{16Mc_2} \wedge \frac{\widetilde{\gamma}}{4}).$$

Let $q^* = \max_j \mu_j(E_j(0))$. We show that for any $t \in (0, \frac{\delta}{4q^*})$ the claim is true.

Now we only consider $\Delta \in (0, \bar{\epsilon}/3)$ and η small enough so that $\eta M \leq \eta^\gamma$ and $\eta^\gamma < \Delta$. Consider the following stopping times

$$\begin{aligned} T^\gamma_{\text{escape}} &\triangleq \min\{n \geq 0 : X^\eta_n \notin \cup_j B(s_j, 2\eta^\gamma)\}; \\ T^\gamma_{\text{return}} &\triangleq \min\{n \geq 0 : X^\eta_n \in \cup_j B(m_j, 2\eta^\gamma)\}. \end{aligned}$$

First, from Lemma C.3, we know that

$$\sup_{x \in [-L,L]} \mathbb{P}_x(T^\gamma_{\text{escape}} > 1/H(1/\eta)) < \delta/2$$

for all η sufficiently small. Besides, by combining Lemma C.2 with Markov property (applied at T^γ_{escape}), we have

$$\sup_{x \in [-L,L]} \mathbb{P}_x\Big(T^\gamma_{\text{return}} - T^\gamma_{\text{escape}} > 2c_2\gamma \log(1/\eta)/\eta \ \Big|\ T^\gamma_{\text{escape}} \leq \frac{1}{H(1/\eta)}\Big) < \delta/2$$

for all η sufficiently small. Therefore, for all η sufficiently small,

$$\sup_{x \in [-L,L]} \mathbb{P}\Big(T^\gamma_{\text{return}} > \frac{1}{H(1/\eta)} + 2c_2\gamma \frac{\log(1/\eta)}{\eta}\Big) < \delta. \tag{C.16}$$

Let J be the unique index such that $X^\eta_{T^\gamma_{\text{return}}} \in \Omega_J$. Our next goal is to show that, almost always, the SGD iterates will visit the local minimum at some *large* attraction fields. Therefore, without loss of generality, we can assume that $m_J \notin M^{\text{large}}$, and define

$$T^\gamma_{\text{large}} \triangleq \min\{n \geq T^\gamma_{\text{return}} :\ X^\eta_n \in \bigcup_{i: m_i \in M^{\text{large}}} B(m_i, 2\Delta)\}$$

and introduce the following definitions:

$$\begin{aligned} &\tau_0 \triangleq T^\gamma_{\text{return}},\ J_0 \triangleq J \\ &\tau_k \triangleq \min\{n > \tau_{k-1} :\ X^\eta_n \in \bigcup_{j \neq J_{k-1}} B(m_j, 2\Delta)\} \\ &J_k = j \Leftrightarrow X^\eta_{\tau_k} \in \Omega_j \quad \forall k \geq 1 \\ &K \triangleq \min\{k \geq 0 :\ m_{J_k} \in M^{\text{large}}\}. \end{aligned}$$

In other words, the sequence of stopping times $(\tau_k)_{k \geq 1}$ is the time that, starting from T^γ_{return}, the SGD iterates visited a local minimum that is different from the one visited at τ_{k-1}, and $(J_k)_{k \geq 0}$ records the label of the visited local minima. The random variable K is the number of transitions required to visit a minimizer in a *large* attraction field. From Lemma C.4, we know the existence of some $p^* > 0$ such that (for all η sufficiently small)

$$\sup_{x \in [-L,L]} \mathbb{P}_x(K \geq u \cdot n_{\min}) \leq \mathbb{P}\Big(\text{Geom}(p^*) \geq u\Big) + \frac{\delta}{2} \quad \forall u = 1, 2, 3, \cdots.$$

where Geom(a) is a Geometric random variable with success rate $a \in (0,1)$. Therefore, one can find integer $N(\delta)$ such that (for all sufficiently small η)

$$\sup_{x\in[-L,L]} \mathbb{P}(K \geq N(\delta)) \leq \delta. \tag{C.17}$$

Next, given results in Proposition B.24 and the fact that there are only finitely many attraction fields, one can find a real number $u(\delta)$ such that (for all sufficiently small η)

$$\sup_{x\in[-L,L]} \mathbb{P}_x(\tau_k - \tau_{k-1} \leq \frac{u(\delta)}{\lambda_{J_{k-1}}(\eta)}) \leq \delta/N(\delta) \tag{C.18}$$

uniformly for all $k = 1, 2, \cdots, N(\delta)$. From (C.16), (C.17),(C.18), we now have

$$\sup_{x\in[-L,L]} \mathbb{P}_x\Big(X_n^\eta \notin \bigcup_{j:m_j\in M^{\text{large}}} B(m_j, 2\Delta)\ \forall n \leq N(\delta)u(\delta)\frac{H(1/\eta)/\eta}{\lambda^{\text{large}}(\eta)} + \frac{1}{H(1/\eta)} + 2c_2\gamma\frac{\log(1/\eta)}{\eta}\Big) \leq 3\delta \tag{C.19}$$

for any sufficiently small η. To conclude the proof we just observe the following facts. First, due to $H \in \mathcal{RV}_{-\alpha}$ and $l^{\text{large}} \geq 2$, we have

$$\lim_{\eta\downarrow 0} H(1/\eta)/\eta = 0, \ \ \lim_{\eta\downarrow 0} \frac{\lambda^{\text{large}}(\eta)}{H(1/\eta)} = 0, \ \ \lim_{\eta\downarrow 0} \frac{\log(1/\eta)}{\eta}\lambda^{\text{large}}(\eta) = 0.$$

Therefore, for sufficiently small η, we will have (note that ϵ, δ are fixed constants in this proof, so $N(\delta), u(\delta)$ are also fixed)

$$\frac{N(\delta)u(\delta)\frac{H(1/\eta)/\eta}{\lambda^{\text{large}}(\eta)} + \frac{1}{H(1/\eta)} + 2c_2\gamma\frac{\log(1/\eta)}{\eta}}{\lfloor t/\lambda^{\text{large}}(\eta)\rfloor} \leq \epsilon. \tag{C.20}$$

Second, recall that we fixed some $t \in (0, \frac{\delta}{4q^*})$ where $q^* = \max_j \mu_j(E_j)$. Also, choose some $C > 0$ small enough so that

$$C < \delta/2, \ 2(1+C)^2 < 4.$$

From Proposition B.22 and the fact that there are only finitely many attraction fields, there exists some $\bar{\eta}_0 > 0$ such that for any $\eta \in (0, \bar{\eta}_0)$ and any $\Delta > 0$ sufficiently small,

$$\begin{aligned}
&\sup_{i:m_i\in M^{\text{large}}} \sup_{x\in[m_i-2\Delta, m_i+2\Delta]} \mathbb{P}_x\Big(\sigma_i(\eta) \leq \frac{t}{\lambda^{\text{large}}(\eta)}\Big) \\
\leq &\sup_{i:m_i\in M^{\text{large}}} \sup_{x\in[m_i-2\Delta, m_i+2\Delta]} \mathbb{P}_x\Big(\mu_i(E_i)\lambda^{\text{large}}(\eta)\sigma_i(\eta) \leq q^*t\Big) \\
\leq &C + 2(1+C)^2q^*t \leq 2\delta.
\end{aligned}$$

Combine this bound with Markov property (applied at τ_K), and we obtain that

$$\sup_{x\in[-L,L]} \mathbb{P}_x\Big(\exists n \in \big[\lfloor t/\lambda^{\text{large}}(\eta)\rfloor\big] \ s.t.\ X^\eta_{n+\tau_K} \notin \bigcup_{i:m_i\in M^{\text{large}}} \Omega_i\Big) \leq 2\delta$$

for all η sufficiently small. Together with (C.19)(C.20), we have shown that

$$\sup_{x\in[-L,L]} \mathbb{P}_x\Big(V^{\text{small}}(\eta,\epsilon,t) > \epsilon\Big) \leq 5\delta$$

holds for all η sufficiently small. □

C.2 Proof of Lemma 10, 11

We shall return to the discussion about the dynamics of SGD iterates on a communication class G. Recall that

$$G^{\text{large}} = \{m_1^{\text{large}}, \cdots, m_{i_G}^{\text{large}}\},\ G^{\text{small}} = \{m_1^{\text{small}}, \cdots, m_{i'_G}^{\text{small}}\}.$$

If X_n^η is initialized at some sharp minimum on G, then we are interested in the behavior of X_n^η at the first visit to some large attraction fields on G. Define

$$T_G(\eta, \Delta) \triangleq \min\{n \geq 0 : X_n^\eta \in \bigcup_{i:\ m_i \in G^{\text{large}}} B(m_i, 2\Delta) \text{ or } X_n^\eta \notin \cup_{i:\ m_i \in G} \Omega_i\}. \tag{C.21}$$

Not only is this definition of T_G analogous to the one for T_G^{DTMC} in (34), but, as illustrated in the next lemma, T_G also behaves similarly as T_G on a communication class G in the following sense: the probabilities $p_{i,j}$ defined in (35) govern the dynamics regarding which large attraction field on G is the first one to be visited. Besides, T_G is usually rather small, meaning that the SGD iterates would efficiently arrive at a large attraction field on G or simply escape from G.

Lemma C.6. *Given any $\theta \in (0, (\alpha-1)/2)$, $\epsilon \in (0,1)$, $i, j \in [n_{min}]$ such that $m_i \in G^{small}$, $m_j \notin G^{large}$, the following claims hold for all $\Delta > 0$ that is sufficiently small:*

$$\limsup_{\eta \downarrow 0} \sup_{x \in B(m_i, 2\Delta)} \mathbb{P}_x\Big(T_G(\eta,\Delta) \leq \frac{\eta^\theta}{\lambda_G(\eta)},\ X_{T_G}^\eta \in B(m_j, 2\Delta)\Big) \leq p_{i,j} + 5\epsilon,$$

$$\liminf_{\eta \downarrow 0} \inf_{x \in B(m_i, 2\Delta)} \mathbb{P}_x\Big(T_G(\eta,\Delta) \leq \frac{\eta^\theta}{\lambda_G(\eta)},\ X_{T_G}^\eta \in B(m_j, 2\Delta)\Big) \geq p_{i,j} - 5\epsilon,$$

$$\limsup_{\eta \downarrow 0} \sup_{x \in B(m_i, 2\Delta)} \mathbb{P}_x\Big(T_G(\eta,\Delta) > \frac{\eta^\theta}{\lambda_G(\eta)}\Big) \leq 2\epsilon.$$

Proof. For $G^{\text{small}} \neq \emptyset$ to hold (and the discussion to be meaningful), we must have $l_G^* \geq 2$. Throughout the proof, we assume this is the case. Besides, we require that $\Delta \in (0, \bar{\epsilon}/3)$ so we have

$$B(m_i, 3\Delta) \cap \Omega_i^c = \emptyset\ \forall i \in [n_{\min}]$$

and the 3Δ-neighborhood of each local minimum will not intersect with each other. In this proof we will only consider Δ in this range.

From Lemma C.4 (if G is absorbing) or Lemma C.5 (if G is transient), we know the existence of some integer $N(\epsilon)$ such that for (see the definition of I_k in (C.11)-(C.14))

$$N_G(\eta, \Delta) \triangleq \min\{k \geq 0 :\ m_{I_k(\eta,\Delta)} \in G^{\text{large}} \text{ or } m_{I_k(\eta,\Delta)} \notin G\},$$

we have

$$\sup_{x \in B(m_i, 2\Delta)} \mathbb{P}_x\Big(N_G(\eta,\Delta) > N(\epsilon)\Big) < \epsilon$$

for all η sufficiently small. Fix such $N(\epsilon)$. Next, from Proposition B.24, we can find $u(\epsilon) \in (0, \infty)$ and $\bar{\Delta} \in (0, \bar{\epsilon}/3)$ such that for all $\Delta \in (0, \bar{\Delta})$, we have

$$\sup_{x \in B(m_i, 2\Delta)} \mathbb{P}_x\Big(T_k(\eta,\Delta) - T_{k-1}(\eta,\Delta) > u(\epsilon)/\Lambda\big(I_{k-1}(\eta,\Delta), \eta\big)\Big) \leq \epsilon/N(\epsilon)\ \ \forall k \in [N(\epsilon)]$$

for all η sufficiently small. Fix such $u(\epsilon)$ and $\bar{\Delta}$. Now note that on the event

$$A \triangleq \Big\{N_G \leq N(\epsilon)\Big\} \cap \Big\{T_k(\eta,\Delta) - T_{k-1}(\eta,\Delta) \leq u(\epsilon)/\Lambda\big(I_{k-1}(\eta,\Delta), \eta\big)\ \forall k \in [N(\epsilon)]\Big\},$$

due to the choice of $\theta \in (0,(\alpha-1)/2)$ and $H \in \mathcal{RV}_{-\alpha}$, we have (when $\eta \in (0,1)$)

$$T_k(\eta,\Delta) - T_{k-1}(\eta,\Delta) \leq \frac{\eta^{2\theta}}{\lambda_G(\eta)} \ \forall k < N_G(\eta,\Delta)$$
$$\Rightarrow T_G(\eta,\Delta) = T_{N_G(\eta,\Delta)}(\eta,\Delta) \leq N(\epsilon)u(\epsilon)\frac{\eta^{2\theta}}{\lambda_G(\eta)}.$$

For any η sufficiently small, we will have $N(\epsilon)u(\epsilon)\frac{\eta^{2\theta}}{\lambda_G(\eta)} < \frac{\eta^{\theta}}{\lambda_G(\eta)}$. In summary, we have established that for all $\Delta \in (0,\bar{\Delta})$,

$$\limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \mathbb{P}_x\Big(T_G > \frac{\eta^{\theta}}{\lambda_G(\eta)}\Big) \leq \limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \mathbb{P}_x(A^c) < 2\epsilon. \tag{C.22}$$

Next, let

$$\mathbf{S}(\epsilon) \triangleq \Big\{\big(m'_1,\cdots,m'_{N(\epsilon)}\big) \in \{m_1,\cdots,m_{n_{\min}}\}^{N(\epsilon)} : \ \exists k \in [N(\epsilon)] \text{ s.t. } m'_k = m_j\Big\}.$$

We can see that $\mathbf{S}(\epsilon)$ contains all the possible transition path for Y^{DTMC} where the state m_j is visited within the first $N(\epsilon)$ steps. Obviously, $|\mathbf{S}(\epsilon)| < \infty$. Let $\epsilon_1 = \epsilon/|\mathbf{S}(\epsilon)|$. If we are able to show the existence of some $\bar{\Delta}_1 > 0$ such that for all $\Delta \in (0,\bar{\Delta}_1)$, the following claim holds for any $(m'_k)_{k=1}^{N(\epsilon)} \in \mathbf{S}(\epsilon)$:

$$\limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \Big|\mathbb{P}_x\Big(m_{I_k} = m'_k \ \forall k \in [N(\epsilon)]\Big) - \mathbb{P}\Big(Y_k^{DTMC}(m_i) = m'_k \ \forall k \in [N(\epsilon)]\Big)\Big| < \epsilon_1, \tag{C.23}$$

then we must have (for all $\Delta \in (0,\bar{\Delta}\wedge\bar{\Delta}_1)$)

$$\begin{aligned}
&\limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \Big|\mathbb{P}_x\Big(X^{\eta}_{T_G} \in B(m_j,2\Delta)\Big) - p_{i,j}\Big| \\
=&\limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \Big|\mathbb{P}_x\Big(X^{\eta}_{T_G} \in B(m_j,2\Delta),\ T_G \leq N(\epsilon)\Big) + \mathbb{P}_x\Big(X^{\eta}_{T_G} \in B(m_j,2\Delta),\ T_G > N(\epsilon)\Big) \\
&\qquad - \mathbb{P}\Big(Y^{DTMC}_{T_G^{DTMC}}(m_i) = m_j,\ T_G^{DTMC} \leq N(\epsilon)\Big) - \mathbb{P}\Big(Y^{DTMC}_{T_G^{DTMC}}(m_i) = m_j,\ T_G^{DTMC} > N(\epsilon)\Big)\Big| \\
\leq&\limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \Big|\mathbb{P}_x\Big(X^{\eta}_{T_G} \in B(m_j,2\Delta),\ T_G \leq N(\epsilon)\Big) - \mathbb{P}\Big(Y^{DTMC}_{T_G^{DTMC}}(m_i) = m_j,\ T_G^{DTMC} \leq N(\epsilon)\Big)\Big| \\
&\quad + \limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \mathbb{P}_x(T_G > N(\epsilon)) + \mathbb{P}(T_G^{DTMC}(m_i) > N(\epsilon)) \\
\leq&|\mathbf{S}(\epsilon)|\epsilon_1 + \limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \mathbb{P}_x(T_G > N(\epsilon)) + \mathbb{P}(T_G^{DTMC}(m_i) > N(\epsilon)) \\
\leq&3\epsilon.
\end{aligned}$$

To show that (C.23) is true, we fix some $(m'_k)_{k=1}^{N(\epsilon)} \in \mathbf{S}(\epsilon)$ and let $(\mathbf{k}'(k))_{k=1}^{N(\epsilon)}$ be the sequence with $m_{\mathbf{k}'(k)} = m'_k$ for each $k \in [N(\epsilon)]$. From the definition of Y^{DTMC} we have (let $\mathbf{k}'(0) = i$)

$$\mathbb{P}\Big(Y_k^{DTMC}(m_i) = m'_k \ \forall k \in [N(\epsilon)]\Big) = \prod_{k=0}^{N(\epsilon)-1} \frac{\mu_{\mathbf{k}'(k)}(E_{\mathbf{k}'(k),\mathbf{k}'(k+1)})}{\mu_{\mathbf{k}'(k)}(E_{\mathbf{k}'(k)})}.$$

On the other hand, using Proposition B.24, we know that for any arbitrarily chosen $\epsilon' \in (0,1)$, we have

$$\limsup_{\eta\downarrow 0} \sup_{x\in B(m_i,2\Delta)} \mathbb{P}_x\Big(m_{I_k(\eta,\Delta)} = m'_k \ \forall k \in [N(\epsilon)]\Big) \leq \prod_{k=0}^{N(\epsilon)-1} \frac{\mu_{\mathbf{k}'(k)}(E_{\mathbf{k}'(k),\mathbf{k}'(k+1)})}{\mu_{\mathbf{k}'(k)}(E_{\mathbf{k}'(k)})} \cdot (1+\epsilon'),$$

$$\liminf_{\eta\downarrow 0}\inf_{x\in B(m_i,2\Delta)}\mathbb{P}_x\Big(m_{I_k(\eta,\Delta)}=m_k'\ \forall k\in[N(\epsilon)]\Big)\geq\prod_{k=0}^{N(\epsilon)-1}\frac{\mu_{\mathbf{k}'(k)}(E_{\mathbf{k}'(k),\mathbf{k}'(k+1)})}{\mu_{\mathbf{k}'(k)}(E_{\mathbf{k}'(k)})}\cdot(1-\epsilon'),$$

for all $\Delta>0$ sufficiently small. The arbitrariness of $\epsilon'>0$, together with $|\mathbf{S}(\epsilon)|<\infty$, allows us to see the existence of some $\bar{\Delta}_1>0$ such that with $\Delta\in(0,\bar{\Delta}_1)$, (C.23) holds for any $(m_k')_{k=1}^{N(\epsilon)}\in\mathbf{S}(\epsilon)$. To conclude the proof, observe that

$$\begin{aligned}&\limsup_{\eta\downarrow 0}\sup_{x\in B(m_i,2\Delta)}\Big|\mathbb{P}_x\Big(X^\eta_{T_G}\in B(m_j,2\Delta)\Big)-\mathbb{P}_x\Big(T_G(\eta,\Delta)\leq\frac{\eta^\theta}{\lambda_G(\eta)},\ X^\eta_{T_G}\in B(m_j,2\Delta)\Big)\Big|\\ \leq&\limsup_{\eta\downarrow 0}\sup_{x\in B(m_i,2\Delta)}\mathbb{P}_x\Big(T_G>\frac{\eta^\theta}{\lambda_G(\eta)}\Big)<\epsilon\end{aligned}$$

due to (C.22). □

Recall that continuous-time process $X^{*,\eta}$ is the *scaled* version of X^η defined in (31), and the mapping $\mathbf{T}^*(n,\eta)\triangleq n\lambda_G(\eta)$ returns the timestamp t for $X_t^{*,\eta}$ corresponding to the unscaled step n on the time horizon of X_n^η. As an *inverse* mapping of $\mathbf{T}^*$, we define the mapping $\mathbf{N}^*(t,\eta)=\lfloor t/\lambda_G(\eta)\rfloor$ that maps the scaled timestamp t back to the step number n for the unscaled process X^η.

In the next lemma, we show that, provided that $X^{*,\eta}$ stays on a communication class G before some time t, the scaled process $X_t^{*,\eta}$ is almost always in the largest attraction fields of a communication class G.

Lemma C.7. *Let G be a communication class on the graph $\mathcal{G}$. Given any $\epsilon_1>0$, $t>0$ and any $x\in\Omega_i$ with $m_i\in G$, the following claim holds for all $\Delta>0$ small enough:*

$$\limsup_{\eta\downarrow 0}\mathbb{P}_x\Big(\Big\{X_t^{*,\eta}\notin\bigcup_{j:m_j\in G^{large}}B(m_j,3\Delta)\Big\}\cap\Big\{X_s^{*,\eta}\in\bigcup_{k:\ m_k\in G}\Omega_k\ \forall s\in[0,t]\Big\}\Big)\leq 2\epsilon_1.$$

Proof. Let $\Delta\in(0,\bar{\epsilon}/3)$ for the constant $\bar{\epsilon}$ in (B.15)(B.16), so we are certain that each $B(m_i,2\Delta)$ lies entirely in Ω_i and would not intersect with each other since

$$B(m_i,3\Delta)\cap\Omega_i^c=\emptyset\quad\forall i\in[n_{\min}].$$

Besides, with $\epsilon=\Delta/3$, we know the existence of some $\delta>0$ such that claims in Lemma B.12 would hold of the chosen ϵ,δ. Fix such δ for the entirety of this proof. Lastly, fix some

$$\widetilde{\gamma}\in(0,(1-\frac{1}{\alpha})\wedge(\frac{1}{2})),\ \beta\in\big(1,(2-2\widetilde{\gamma})\wedge(\alpha-\alpha\widetilde{\gamma})\big),\ \gamma\in(0,\frac{\widetilde{\gamma}}{16Mc_2}\wedge\frac{\widetilde{\gamma}}{4}).$$

The blueprint of this proof is as follows. We will define a sequence of stopping times $(N_j)_{j=1}^6$ such that the corresponding scaled timestamps $\mathbf{T}_j^*=\mathbf{T}^*(N_j,\eta)$ gradually approach t. By analyzing the behavior of $X^{*,\eta}$ on a time interval $[t-\Delta_t,t]$ that is very close to t (in particular, on the aforementioned stopping times $\mathbf{T}_j^*$), we are able to establish the properties of a series of events $A_1\supseteq A_2\supseteq A_3$. Moreover, we will show that $A_3\subseteq\{X_t^{*,\eta,\Delta}\in\bigcup_{i:\ m_i\in G^{\text{large}}}B(m_i,3\Delta)\}$, so the properties about events A,A_2,A_3 can be used to bound the probability of the target event.

Arbitrarily choose some $\Delta_t\in(0,t)$. To proceed, let $N_0\triangleq\mathbf{N}^*(t-\Delta_t,\eta)$ be the stopping time corresponding to timestamp $t-\Delta_t$ for the scaled process. Using Lemma C.3, we know that for stopping time $N_1\triangleq\min\{n\geq N_0:\ X_n^\eta\notin\cup_j B(s_j,2\eta^\gamma)\}$, we have

$$\liminf_{\eta\downarrow 0}\inf_{x\in[-L,L]}\mathbb{P}_x(N_1-N_0<\frac{\Delta_t/4}{H(1/\eta)})=1.$$

Next, let $N_2 \triangleq \min\{n \geq N_1 : \ X_n^\eta \in \cup_j B(m_j, 2\Delta)\}$. From Lemma C.2 and $H \in \mathcal{RV}_{-\alpha}$ (so that $\log(1/\eta)/\eta = o(H(1/\eta))$, we have

$$\liminf_{\eta \downarrow 0} \inf_{x \in [-L,L]} \mathbb{P}_x(N_2 - N_1 < \frac{\Delta_t/4}{H(1/\eta)}) = 1.$$

Collecting results above, we have

$$\liminf_{\eta \downarrow 0} \inf_{x \in [-L,L]} \mathbb{P}_x(N_2 - N_0 < \frac{\Delta_t/2}{H(1/\eta)}) = 1. \tag{C.24}$$

Now note the following fact on the event $\{N_2 - N_0 < \frac{\Delta_t/2}{H(1/\eta)}\}$. The definition of the mapping $\mathbf{T}^*$ implies that, for any pair of positive integers $n_1 \leq n_2$, we have $\mathbf{T}^*(n_2, \eta) - \mathbf{T}^*(n_1, \eta) = (n_2 - n_1)\lambda_G(\eta) \leq (n_2 - n_1) \cdot H(1/\eta)$. Therefore, on $\{N_2 - N_0 \leq \frac{\Delta_t/2}{H(1/\eta)}\}$ we have

$$\mathbf{T}^*(N_2, \eta) - \mathbf{T}^*(N_0, \eta) < \Delta_t/2 \Rightarrow \mathbf{T}^*(N_2, \eta) < t - \frac{\Delta_t}{2}.$$

Besides, let $\mathbf{T}_2^* = \mathbf{T}^*(N_2, \eta)$. Now we can see that for event

$$A_0 \triangleq \Big\{X_s^{*,\eta} \in \bigcup_{k:\ m_k \in G} \Omega_k \ \forall s \in [0,t]\Big\} \cap \Big\{N_2 - N_0 < \frac{\Delta_t/2}{H(1/\eta)}\Big\},$$

we have

$$A_0 \subseteq A_1 \triangleq \Big\{\mathbf{T}_2^* < t - \frac{\Delta_t}{2}\Big\} \cap \Big\{X_s^{*,\eta} \in \bigcup_{k:\ m_k \in G} \Omega_k \ \forall s \in [0, \mathbf{T}_2^*]\Big\}.$$

Meanwhile, from (C.24) we obtain that

$$\limsup_{\eta \downarrow 0} \sup_{x \in [-L,L]} \mathbb{P}_x\Big(A_1^c \cap \{X_s^{*,\eta} \in \bigcup_{k:\ m_k \in G} \Omega_k \ \forall s \in [0,t]\}\Big) = 0. \tag{C.25}$$

Moving on, we consider the following stopping times

$$N_3 \triangleq \min\{n \geq N_2 : \ X_n^\eta \in \bigcup_{j:\ m_j \in G^{\text{large}}} B(m_j, 2\Delta) \text{ or } X_n^\eta \notin \bigcup_{j:\ m_j \in G} \Omega_j\},$$
$$\mathbf{T}_3^* \triangleq \mathbf{T}^*(N_3, \eta).$$

Using Lemma C.6, we have

$$\limsup_{\eta \downarrow 0} \mathbb{P}_x\Big(N_3 - N_2 > \frac{\Delta_t/4}{\lambda_G(\eta)} \ \Big| \ A_1\Big) \leq \epsilon_1. \tag{C.26}$$

Meanwhile, on event $A_1 \cap \Big\{N_3 - N_2 \leq \frac{\Delta_t/4}{\lambda_G(\eta)}\Big\} \cap \Big\{X_s^{*,\eta} \in \bigcup_{k:\ m_k \in G} \Omega_k \ \forall s \in [0,t]\Big\}$, we have $\mathbf{T}_3^* - \mathbf{T}_2^* \leq \Delta_t/4$, hence $\mathbf{T}_3^* \in [t - \Delta_t, t - \Delta_t/4]$. In summary,

$$\begin{aligned} &A_1 \cap \Big\{N_3 - N_2 \leq \frac{\Delta_t/4}{\lambda_G(\eta)}\Big\} \cap \Big\{X_s^{*,\eta} \in \bigcup_{k:\ m_k \in G} \Omega_k \ \forall s \in [0,t]\Big\} \\ &\subseteq \Big\{\mathbf{T}_3^* \in [t - \Delta_t, t - \Delta_t/4]\Big\} \cap \Big\{X_s^{*,\eta} \in \bigcup_{k:\ m_k \in G} \Omega_k \ \forall s \in [0, \mathbf{T}_3^*]\Big\} \end{aligned}$$

Moreover, on event $\Big\{X_s^{*,\eta} \in \bigcup_{k:\ m_k\in G}\Omega_k\ \forall s\in[0,t]\Big\}$, if we let J_3 be label of the local minimum visited at $\mathbf{T}_3^*$ such that $J_3=j \iff X_{T_3^*}^{*,\eta}\in B(m_j,2\Delta)$, then we must have $m_{J_3}\in G^{\text{large}}$. Meanwhile, consider the following stopping times

$$\mathbf{T}^\sigma \triangleq \min\{s>\mathbf{T}_3^*:\ X_s^{*,\eta}\notin\Omega_{J_3}\}.$$

From Proposition B.24, we know that

$$\begin{aligned}&\limsup_{\eta\downarrow 0}\mathbb{P}_x\Big(\mathbf{T}^\sigma-\mathbf{T}_3^*\le\Delta_t\ \Big|\ \Big\{\mathbf{T}_3^*\in[t-\Delta_t,t-\Delta_t/4]\Big\}\cap\Big\{X_s^{*,\eta}\in\bigcup_{k:\ m_k\in G}\Omega_k\ \forall s\in[0,\mathbf{T}_3^*]\Big\}\Big)\\ \le&\epsilon_1+1-\exp\big(-(1+\epsilon_1)q^*\Delta_t\big)\end{aligned}\tag{C.27}$$

where $q^*=\max_j\mu_j(E_j)$. Now we define the event

$$A_2\triangleq A_1\cap\Big\{N_3-N_2\le\frac{\Delta_t/4}{\lambda_G(\eta)}\Big\}\cap\Big\{\mathbf{T}^\sigma-\mathbf{T}_3^*>\Delta_t\Big\}\cap\Big\{X_s^{*,\eta}\in\bigcup_{k:\ m_k\in G}\Omega_k\ \forall s\in[0,\mathbf{T}_3^*]\Big\}.$$

Using (C.25)-(C.27), we get

$$\limsup_{\eta\downarrow0}\sup_{x\in[-L,L]}\mathbb{P}_x\Big(A_2^c\cap\{X_s^{*,\eta}\in\bigcup_{k:\ m_k\in G}\Omega_k\ \forall s\in[0,t]\}\Big)\le 2\epsilon_1+1-\exp\big(-(1+\epsilon_1)q^*\Delta_t\big).\tag{C.28}$$

Furthermore, on event A_2, due to $\mathbf{T}_3^*\in[t-\Delta_t,t-\Delta_t/4]$ as established above, we must have

$$X_s^{*,\eta}\in\Omega_{J_3}\ \forall s\in[\mathbf{T}_3^*,t].$$

Now let us focus on a timestamp $\mathbf{T}_4^*=t-\frac{\Delta_t/8}{H(1/\eta)}\lambda_G(\eta)$ and $N_4=\mathbf{N}^*(\mathbf{T}_4^*,\eta)$. Obviously, $\mathbf{T}_4^*>\mathbf{T}_3^*$ on event A_2. Next, define

$$\begin{aligned}N_5&\triangleq\min\{n\ge N_4:\ X_n^\eta\in\bigcup_j B(m_j,2\Delta)\}\\ \mathbf{T}_5^*&\triangleq\mathbf{T}^*(N_5,\eta).\end{aligned}$$

Using Lemma C.2 and C.3 again as we did above when obtaining (C.24), we can show that

$$\limsup_{\eta\downarrow0}\mathbb{P}_x\Big(N_5-N_4>\frac{\Delta_t/16}{H(1/\eta)}\Big)=0.\tag{C.29}$$

On the other hand, on event $A_2\cap\Big\{N_5-N_4\le\frac{\Delta_t/16}{H(1/\eta)}\Big\}$ we must have

- $\mathbf{T}_5^*-\mathbf{T}_4^*\le\frac{\Delta_t/16}{H(1/\eta)}\lambda_G(\eta)$, so $\mathbf{T}_5^*\in[t-\frac{\Delta_t/8}{H(1/\eta)}\lambda_G(\eta),t-\frac{\Delta_t/16}{H(1/\eta)}\lambda_G(\eta)]$;
- $X_{\mathbf{T}_5^*}^{*,\eta}\in\Omega_{J_3}$, due to $\mathbf{T}^\sigma-\mathbf{T}_3^*>\Delta_t$.

This implies that for event

$$\widetilde{A}\triangleq\Big\{\mathbf{T}_5^*\in[t-\frac{\Delta_t/8}{H(1/\eta)}\lambda_G(\eta),t-\frac{\Delta_t/16}{H(1/\eta)}\lambda_G(\eta)],\ X_{\mathbf{T}_5^*}^{*,\eta}\in\Omega_{J_3}\Big\}\cap\{X_s^{*,\eta}\in\bigcup_{k:\ m_k\in G}\Omega_j\ \forall s\in[0,\mathbf{T}_5^*]\},$$

we have $A_2\cap\Big\{N_5-N_4\le\frac{\Delta_t/16}{H(1/\eta)}\Big\}\subseteq\widetilde{A}$. Lastly, observe that

- From Lemma B.2, we know that for $N_6(\delta)\triangleq\min\{n>N_5:\eta|Z_n|>\delta\}$ we have

$$\limsup_{\eta\downarrow0}\mathbb{P}\big(N_6(\delta)-N_5\le\Delta_t/H(1/\eta)\big)\le\Delta_t/\delta^\alpha;$$

- As stated at the beginning of the proof, our choice of δ allows us to apply Lemma B.12 and show that

$$\limsup_{\eta\downarrow 0}\sup_{x\in[-L,L]}\mathbb{P}_x\big(\exists n=N_5,\cdots,N_6-1 \text{ s.t. } X_n^\eta\notin B(m_{J_3},3\Delta)\mid \widetilde{A}\big)=0;$$

- Combining the two bullet points above, we get

$$\limsup_{\eta\downarrow 0}\mathbb{P}_x\Big(\exists s\in[\mathbf{T}_5^*,t] \text{ such that } X_s^{*,\eta}\notin B(m_{J_3},3\Delta)\ \Big|\ \widetilde{A}\Big)\le \Delta_t/\delta^\alpha. \tag{C.30}$$

  On the other hand,

$$\widetilde{A}\cap\{X_s^{*,\eta}\in B(m_{J_3},3\Delta)\ \forall s\in[\mathbf{T}_5^*,t]\}\subseteq\{X_t^{*,\eta}\in\bigcup_{k:\ m_k\in G^{\text{large}}}B(m_k,3\Delta)\}.$$

In summary, for event

$$A_3\triangleq A_2\cap\Big\{N_5-N_4\le\frac{\Delta_t/16}{H(1/\eta)}\Big\}\cap\Big\{X_s^{*,\eta}\in B(m_{J_3},3\Delta)\ \forall s\in[\mathbf{T}_5^*,t]\Big\},$$

we have $A_3\subseteq\{X_t^{*,\eta}\in\bigcup_{k:\ m_k\in G^{\text{large}}}B(m_k,3\Delta)\}$. Besides, due to (C.28)(C.29)(C.30), we get

$$\begin{aligned}&\limsup_{\eta\downarrow 0}\sup_{x\in[-L,L]}\mathbb{P}_x\Big(A_3^c\cap\{X_s^{*,\eta}\in\bigcup_{k:\ m_k\in G}\Omega_k\ \forall s\in[0,t]\}\Big)\\ \le&2\epsilon_1+1-\exp\big(-(1+\epsilon_1)q^*\Delta_t\big)+\frac{\Delta_t}{\delta^\alpha}.\end{aligned}$$

Remember that δ,ϵ_1,q^* are fixed constants while Δ_t can be made arbitrarily small, so by driving Δ_t to 0 we can conclude the proof. □

Recall the definition of jump processes in Definition 2. Central to the proof of Lemma 10, the next result provides a set of sufficient conditions for the convergence of a sequence of such jump processes in the sense of finite dimensional distributions.

Lemma C.8. *For a sequence of processes $(Y^n)_{n\ge1}$ that, for each $n\ge1$, Y^n is a $\Big((U_j^n)_{j\ge0},(V_j^n)_{j\ge0}\Big)$ jump process, and a $\Big((U_j)_{j\ge0},(V_j)_{j\ge0}\Big)$ jump process Y, if*

- *$U_0\equiv0$;*
- *$(U_0^n,V_0^n,U_1^n,V_1^n,U_2^n,V_2^n,\cdots)$ converges in distribution to $(0,V_0,U_1,V_1,U_2,V_2,\cdots)$ as $n\to\infty$;*
- *For any $x>0$ and any $n\ge1$,*

$$\mathbb{P}(U_1+\cdots+U_n=x)=0;$$

- *For any $x>0$,*

$$\lim_{n\to\infty}\mathbb{P}(U_1+U_2+\cdots U_n>x)=1,$$

then the finite dimensional distribution of Y^n converges to that of Y in the following sense: for any $k\in\mathbb{N}$ and any $0<t_1<t_2<\cdots<t_k<\infty$, the random element $(Y_{t_1}^n,\cdots,Y_{t_k}^n)$ converges in distribution to $(Y_{t_1},\cdots,Y_{t_k})$ as $n\to\infty$.

Proof. Fix some $k \in \mathbb{N}$ and $0 < t_1 < t_2 < \cdots < t_k < \infty$. For notational simplicity, let $t = t_k$. Let $(\mathbb{D}, \mathbf{d})$ be the metric space where $\mathbb{D} = \mathbb{D}_{[0,t]}$, the space of all càdlàg functions in $\mathbb{R}$ on the time interval $[0, t]$, and $\mathbf{d}$ is the Skorokhod metric defined as

$$\mathbf{d}(\zeta_1, \zeta_2) \triangleq \inf_{\lambda \in \Lambda} \|\zeta_1 - \zeta_2 \circ \lambda\| \vee \|\lambda - I\|$$

where Λ is the set of all nondecreasing homeomorphism from $[0, t]$ onto itself, and $I(s) = s$ is the identity mapping. Also, we arbitrarily choose some $\epsilon \in (0, 1)$ and some open set $A \subseteq \mathbb{R}^k$.

From the assumption, we can find integer $J(\epsilon)$ such that $\mathbb{P}\Big(\sum_{j=1}^{J(\epsilon)} U_j \leq t\Big) < \epsilon$, as well as an integer $N(\epsilon)$ such that, for all $n \geq N(\epsilon)$, we have $\mathbb{P}\Big(\sum_{j=1}^{J(\epsilon)} U_j^n \leq t\Big) + \mathbb{P}\Big(U_0^n \geq t_1\Big) < \epsilon$. We fix such $J(\epsilon), N(\epsilon)$ (we may abuse the notation slightly and simply write J, N when there is no ambiguity).

Using Skorokhod's representation theorem, we can construct a probability space $(\mathbf{\Omega}, \mathcal{F}, \mathbb{Q})$ that supports random variables $(\widetilde{U}_0^n, \widetilde{V}_0^n, \cdots, \widetilde{U}_J^n, \widetilde{V}_J^n)_{n \geq 1}$ and $(\widetilde{U}_0, \widetilde{V}_0, \cdots, \widetilde{U}_J, \widetilde{V}_J)$ and satisfies the following conditions:

- $\mathcal{L}(U_0^n, V_0^n, \cdots, U_J^n, V_J^n) = \mathcal{L}(\widetilde{U}_0^n, \widetilde{V}_0^n, \cdots, \widetilde{U}_J^n, \widetilde{V}_J^n)$ for all $n \geq 1$;
- $\mathcal{L}(U_0, V_0, \cdots, U_J, V_J) = \mathcal{L}(\widetilde{U}_0, \widetilde{V}_0, \cdots, \widetilde{U}_J, \widetilde{V}_J)$;
- $U_j^n \xrightarrow{a.s.} U_j$ and $V_j^n \xrightarrow{a.s.} V_j$ as $n \to \infty$ for all $j \in [J]$.

Therefore, on $(\mathbf{\Omega}, \mathcal{F}, \mathbb{Q})$ we can define the following random elements (taking values in the space of càdlàg functions):

$$Y_s^{n,\downarrow J} = \begin{cases} \widetilde{V}_0^n & \text{if } s < \widetilde{U}_0^n \\ \sum_{j=0}^{J} \widetilde{V}_j^n \mathbb{1}_{[\widetilde{U}_0^n + \widetilde{U}_1^n + \cdots + \widetilde{U}_j^n,\ \widetilde{U}_0^n + \widetilde{U}_1^n + \cdots + \widetilde{U}_{j+1}^n)}(s) & \text{otherwise} \end{cases},$$

$$Y_s^{\downarrow J} = \sum_{j=0}^{J} \widetilde{V}_j \mathbb{1}_{[\widetilde{U}_1 + \cdots + \widetilde{U}_j,\ \widetilde{U}_1 + \cdots + \widetilde{U}_{j+1})}(s) \quad \forall s \geq 0.$$

Note that (1) for the first jump time of $Y^{\downarrow J}$ we have $\widetilde{U}_0 \equiv 0$, hence $Y_0^{\downarrow J} = \widetilde{V}_0$; (2) when defining $Y^{n,\downarrow J}$ we set its value on $[0, \widetilde{U}_0^n)$ to be $\widetilde{V}_0^n$ instead of 0.

Since $U_j^n \xrightarrow{a.s.} U_j$ and $V_j^n \xrightarrow{a.s.} V_j$ as $n \to \infty$ for all $j \in [J]$, we must have

$$\lim_n \mathbf{d}(Y_s^{n,\downarrow J}, Y_s^{\downarrow J}) = 0$$

almost surely, which further implies that $Y_s^{n,\downarrow J} \Rightarrow Y_s^{\downarrow J}$ as $n \to \infty$ on $(\mathbb{D}, \mathbf{d})$. Now from our assumption that, for the jump times $U_1 + \cdots + U_j$, we have $\mathbb{P}(U_1 + \cdots + U_j = x) = 0\ \forall x > 0, j \geq 1$, as well as (13.3) in [1], we then obtain

$$(Y_{t_1}^{n,\downarrow J}, \cdots, Y_{t_k}^{n,\downarrow J}) \Rightarrow (Y_{t_1}^{\downarrow J}, \cdots, Y_{t_k}^{\downarrow J}) \tag{C.31}$$

as $n \to \infty$. Recall that A is the open set we arbitrarily chose at the beginning of the proof, and $\epsilon > 0$ is also chosen arbitrarily. Now we observe the following facts.

- Using (C.31), we can see that

$$\liminf_n \mathbb{Q}\big((Y_{t_1}^{n,\downarrow J}, \cdots, Y_{t_k}^{n,\downarrow J}) \in A\big) \geq \mathbb{Q}\big(Y_{t_1}^{\downarrow J}, \cdots, Y_{t_k}^{\downarrow J}) \in A\big).$$

- The choice of $N(\epsilon)$ and $J(\epsilon)$ above implies that

$$\big|\mathbb{Q}\big((Y_{t_1}^{\downarrow J(\epsilon)}, \cdots, Y_{t_k}^{\downarrow J(\epsilon)}) \in A\big) - \mathbb{P}\big((Y_{t_1}, \cdots, Y_{t_k}) \in A\big)\big| \leq \mathbb{P}\Big(\sum_{j=1}^{J(\epsilon)} U_j \leq t\Big) < \epsilon$$

$$\big|\mathbb{Q}\big((Y^{n,\downarrow J(\epsilon)}_{t_1},\cdots,Y^{n,\downarrow J(\epsilon)}_{t_k})\in A\big)-\mathbb{P}\big((Y^n_{t_1},\cdots,Y^n_{t_k})\in A\big)\big|\le \mathbb{P}\Big(\sum_{j=1}^{J(\epsilon)}U^n_j\le t\Big)+\mathbb{P}\Big(U^n_0\ge t_1\Big)<\epsilon\ \ \forall n\ge N(\epsilon).$$

Collecting the two results above, we have established that

$$\liminf_n \mathbb{P}\big((Y^n_{t_1},\cdots,Y^n_{t_k})\in A\big)\ge \mathbb{P}\big((Y_{t_1},\cdots,Y_{t_k})\in A\big)-2\epsilon.$$

From Portmanteau theorem, together with arbitrariness of $\epsilon>0$ and open set A, we can now conclude that $(Y^n_{t_1},\cdots,Y^n_{t_k})$ converges in distribution to $(Y_{t_1},\cdots,Y_{t_k})$. □

The following lemma concerns the scaled version of the marker process $\hat{X}^{*,\eta,\Delta}$ defined in (31)-(32). Obviously, it is a jump process that complies with Definition 2. When there is no ambiguity about the sequences $(\eta_n)_{n\ge1}$ and $(\Delta_n)_{n\ge1}$, let $\hat{X}^{(n)}_t\triangleq\hat{X}^{*,\eta_n,\Delta_n}_t$. From (24)-(29) and (33), we know that for any $n\ge1$, $\hat{X}^{(n)}$ is a $\Big(\big(\tau^*_k(\eta_n,\Delta_n)-\tau^*_{k-1}(\eta_n,\Delta_n)\big)_{k\ge0},\ \big(m_{I_k(\eta_n,\Delta_n)}\big)_{k\ge0}\Big)$-jump process (with the convention that $\tau^*_{-1}=0$). Also, for clarity of the exposition, we let (for all $n\ge1,k\ge0$)

$$\begin{aligned}\widetilde{S}^{(n)}_k&=\sigma^*_k(\eta_n,\Delta_n)-\tau^*_{k-1}(\eta_n,\Delta_n),\\ S^{(n)}_k&=\tau^*_k(\eta_n,\Delta_n)-\tau^*_{k-1}(\eta_n,\Delta_n),\\ \widetilde{W}^{(n)}_k&=m_{\widetilde{I}^G_k(\eta_n,\Delta_n)},\\ W^{(n)}_k&=m_{I^G_k(\eta_n,\Delta_n)}.\end{aligned}$$

Lastly, remember that Y is the continuous-time Markov chain defined in (36)-(39) and $\pi_G(\cdot)$ is the random mapping defined in (40) that is used to initialize Y. Besides, Y is a $\big((S_k)_{k\ge0},(W_k)_{k\ge0}\big)$ jump process under Definition 2, with $S_0=0$ and $W_0=\pi_G(m_i)$ (here $x\in\Omega_i$ and $X^\eta_0=x$, so i is the index of the attraction field where the SGD iterate is initialized). The following result states that, given a sequence of learning rates $(\eta_n)_{n\ge1}$ that tend to 0, we are able to find a sequence of $(\Delta_n)_{n\ge1}$ to parametrize $\hat{X}^{(n)}=\hat{X}^{*,\eta_n,\Delta_n},X^{(n)}=X^{*,\eta_n,\Delta_n}$ so that they have several useful properties, one of which is that the jump times and locations of $\hat{X}^{(n)}$ converges in distribuiton to those of $Y(\pi_G(m_i))$.

Lemma C.9. *Assume the communication class G is absorbing. Given any $m_i\in G$, $x\in\Omega_i$, finitely many real numbers $(t_l)_{l=1}^{k'}$ such that $0<t_1<t_2<\cdots<t_{k'}$, and a sequence of strictly positive real numbers $(\eta_n)_{n\ge1}$ with $\lim_{n\to0}\eta_n=0$, there exists a sequence of strictly positive real numbers $(\Delta_n)_{n\ge1}$ with $\lim_n\Delta_n=0$ such that*

- *Under $\mathbb{P}_x$ (so $X^\eta_0=x$), as n tends to ∞,*

$$(S^{(n)}_0,W^{(n)}_0,S^{(n)}_1,W^{(n)}_1,S^{(n)}_2,W^{(n)}_2,\cdots)\Rightarrow(S_0,W_0,S_1,W_1,S_2,W_2,\cdots)\tag{C.32}$$

- *(Recall the definition of T_k,I_k in (C.11)-(C.14)) Given any $\epsilon>0$, the following claim holds for all n sufficiently large:*

$$\sup_{k\ge0}\mathbb{P}_x\Big(\exists j\in[T_k(\eta_n,\Delta_n),T_k(\eta_n,\Delta_n)]\ s.t.\ X^\eta_j\notin\bigcup_{l:\ m_l\in G}\Omega_l\ \mid\ m_{I_k(\eta_n,\Delta_n)}\in G\Big)<\epsilon;\tag{C.33}$$

- *Given any $\epsilon>0$, the following claim holds for all n sufficiently large,*

$$\sup_{k\ge0}\mathbb{P}_k\Big(m_{I_k(\eta_n,\Delta_n)+v}\notin G^{large}\ \forall v\in[un_{min}]\ \Big|\ m_{I_k(\eta_n,\Delta_n)}\in G\Big)\le\mathbb{P}(Geom(p^*)\ge u)+\epsilon\ \forall u=1,2,\cdots;\tag{C.34}$$

- *For any $l \in [k']$,*

$$\lim_{n\to\infty} \mathbb{P}_x\Big(X^{*,\eta_n}_{t_l} \notin \bigcup_{j:\ m_j\in G^{large}} B(m_j,\Delta_n),\ X^{*,\eta_n}_s \in \bigcup_{j:\ m_j\in G} \Omega_j\ \forall s\in[0,t_{k'}]\Big) = 0 \tag{C.35}$$

where $p^* > 0$ is a constant that does not vary with our choices of η_n or Δ_n.

Proof. Let

$$\nu_j \triangleq q_j = \mu_j(E_j)$$
$$\nu_{j,k} \triangleq \mu_j(E_{j,k})$$

so from the definition of $q_{j,k}$ we have $q_{j,k} = \mathbb{1}\{j\neq k\}\nu_{j,k} + \sum_{l:\ m_l\in G^{\text{small}}} \nu_{j,l}p_{l,k}$.

In order to specify our choice of $(\Delta_n)_{n\geq 1}$, we consider a construction of sequences $(\bar{\boldsymbol{\Delta}}(j))_{j\geq 0}, (\bar{\boldsymbol{\eta}}(j))_{j\geq 0}$ as follows. Fix some $\theta \in (0,\alpha-1)/2)$. Let $\bar{\boldsymbol{\Delta}}(0) = \bar{\boldsymbol{\eta}}(0) = 1$. One can see the existence of some $(\bar{\boldsymbol{\Delta}}(j))_{j\geq 1}, (\bar{\boldsymbol{\eta}}(j))_{j\geq 1}$ such that

- $\bar{\boldsymbol{\Delta}}(j) \in \big(0, \bar{\boldsymbol{\Delta}}(j-1)/2\big],\ \bar{\boldsymbol{\eta}}(j) \in \big(0, \bar{\boldsymbol{\eta}}(j-1)/2\big]$ for all $j \geq 1$;
- (Due to Lemma C.2) for any $j\geq 1$, $\eta \in (0, \bar{\boldsymbol{\eta}}(j)]$, (remember that x and i are the fixed constants prescribed in the description of the lemma)

$$\mathbb{P}_x\Big(\sigma^*_0\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) < \eta^\theta,\ \widetilde{I}^G_0\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i\Big) > 1 - \frac{1}{2^j}.$$

  For definitions of $\sigma^G_k, \tau^G_k, I^G_k, \widetilde{I}^G_k$, see (24)-(29).

- (Due to Lemma C.6) for any $j\geq 1$, $\eta\in(0,\bar{\boldsymbol{\eta}}(j)]$,

$$\Big|\mathbb{P}_x\Big(\tau^*_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) - \sigma^*_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) < \eta^\theta,\ I^G_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_2 \ \Big|\ \widetilde{I}^G_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big) - p_{i_1,i_2}\Big| < 1/2^j$$

  uniformly for all $k\geq 0$ and all $m_{i_1}\in G^{\text{small}}, m_{i_2}\in G^{\text{large}}$. Also, by definition of σ^* and τ^*, we must have

$$\mathbb{P}_x\Big(\tau^*_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) - \sigma^*_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = 0,\ I^G_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1 \ \Big|\ \widetilde{I}^G_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big) = 1$$

  for all $k\geq 0$ and $m_{i_1}\in G^{\text{large}}$.

- (Due to Proposition B.24) for any $j\geq 1$, $\eta\in(0,\bar{\boldsymbol{\eta}}(j)]$,

$$\begin{aligned}
&-\frac{1}{2^j} + \exp\big(-(1+\frac{1}{2^j})q_{i_1}u\big)\frac{\nu_{i_1,i_2} - \frac{1}{2^j}}{q_{i_1}}\\
\leq&\mathbb{P}_x\Big(\sigma^*_{k+1}\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) - \tau^*_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) > u,\ \widetilde{I}^G_{k+1} = i_2 \ \Big|\ I^G_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big)\\
\leq&\frac{1}{2^j} + \exp\big(-(1-\frac{1}{2^j})q_{i_1}u\big)\frac{\nu_{i_1,i_2} + \frac{1}{2^j}}{q_{i_1}}
\end{aligned}$$

  uniformly for all $k\geq 1$, all $u > 1/2^j$, and all $m_{i_1}\in G^{\text{large}}, m_{i_2}\in G$.

- (Due to G being absorbing and, again, Proposition B.24) for any $j\geq 1$, for any $j\geq 1$, $\eta\in(0,\bar{\boldsymbol{\eta}}(j)]$, (Recall the definition of T_k, I_k in (C.11)-(C.14))

$$\mathbb{P}_x\Big(I_{k+1}\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_2 \mid I_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big) < \frac{1}{2^j} \tag{C.36}$$

  uniformly for all $k\geq 0$, $m_{i_1}\in G, m_{i_2}\notin G$.

- (Due to Lemma C.4) There exists some $p^* > 0$ such that for any $j \geq 1$, for any $j \geq 1$, $\eta \in (0, \bar{\boldsymbol{\eta}}(j)]$,

$$\mathbb{P}_x\Big(m_{v+I_k\left(\eta,\bar{\boldsymbol{\Delta}}(j)\right)} \notin G^{\text{large}}\ \forall v \in [un_{\min}] \ \Big|\ I_k\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big) \leq \mathbb{P}(Geom(p^*) \geq u) + 1/2^j \tag{C.37}$$

uniformly for all $k \geq 0, u \geq 1$ and $m_{i_1} \in G$.

- (Due to Lemma C.7) for any $j \geq 1$, for any $j \geq 1$, $\eta \in (0, \bar{\boldsymbol{\eta}}(j)]$,

$$\mathbb{P}_x\Big(X^{*,\eta_n}_{t_k} \notin \bigcup_{j:\ m_j \in G^{\text{large}}} B(m_j, \bar{\boldsymbol{\Delta}}(j)),\ X^{*,\eta_n}_s \in \bigcup_{j:\ m_j \in G} \Omega_j\ \forall s \in [0, t_{k'}]\Big) < 1/2^j \tag{C.38}$$

uniformly for all $k \in [k']$.

Fix such $(\bar{\boldsymbol{\Delta}}(j))_{j\geq 0}, (\bar{\boldsymbol{\eta}}(j))_{j\geq 0}$. Define a function $\mathbf{J}(\cdot) : \mathbb{N} \mapsto \mathbb{N}$ as

$$\mathbf{J}(n) = 0 \vee \max\{j \geq 0 :\ \bar{\boldsymbol{\eta}}(j) \geq \eta_n\}$$

with the convention that $\max \emptyset = -\infty$. Lastly, let

$$\Delta_n = \bar{\boldsymbol{\Delta}}(\mathbf{J}(n))\ \forall n \geq 1.$$

Note that, due to $\lim_n \eta_n = 0$, we have $\lim_n \mathbf{J}(n) = \infty$, hence $\lim_n \Delta_n = 0$. Besides, the definition of $\mathbf{J}(\cdot)$ tells us that in case that $\mathbf{J}(n) \geq 1$ (which will hold for all n sufficiently large), the claims above holds with $\eta = \eta_n$ and $j = \mathbf{J}(n)$. In particular, by combining $\lim_n \mathbf{J}(n) = \infty$ with (C.36)(C.37)(C.38) respectively, we have (C.33)(C.34)(C.35).

Now it remains to prove (C.32). To this end, it suffices to show that, for any positive integer K, we have $(S^{(n)}_0, W^{(n)}_0, \cdots, S^{(n)}_K, W^{(n)}_K)$ converges in distribution $(S_0, W_0, \cdots, S_K, W_K)$ as n tends to infinity. In particular, note that $S_0 = 0, W_0 = \pi_G(m_i)$, so $W_0 = m_j$ with probability $p_{i,j}$ if $m_i \in G^{\text{small}}$, and $W_0 \equiv m_i$ if $m_i \in G^{\text{large}}$.

For clarity of the exposition, we restate some important claims above under the new notational system with $\widetilde{S}^{(n)}_k, \widetilde{W}^{(n)}_k, S^{(n)}_k, W^{(n)}_k$ we introduced right above this lemma. Given any $\epsilon > 0$, the following claims hold for all n sufficiently large:

- First of all,

$$\mathbb{P}_x\Big(\widetilde{S}^{(n)}_0 < \eta^\theta_n,\ \widetilde{W}^{(n)}_0 = m_i\Big) > 1 - \epsilon. \tag{C.39}$$

- For all $k \geq 0$ and all $m_{i_1} \in G^{\text{small}}, m_{i_2} \in G^{\text{large}}$,

$$\Big|\mathbb{P}_x\Big(S^{(n)}_k - \widetilde{S}^{(n)}_k < \eta^\theta_n,\ W^{(n)}_k = m_{i_2} \ \Big|\ \widetilde{W}^{(n)}_k = m_{i_1}\Big) - p_{i_1,i_2}\Big| < \epsilon. \tag{C.40}$$

- For all $k \geq 0$ and all $m_{i_1} \in G^{\text{large}}$,

$$\mathbb{P}_x\Big(S^{(n)}_k - \widetilde{S}^{(n)}_k = 0,\ W^{(n)}_k = m_{i_1} \ \Big|\ \widetilde{W}^{(n)}_k = m_{i_1}\Big) = 1. \tag{C.41}$$

- For all $k \geq 0$, all $m_{i_1} \in G^{\text{large}}, m_{i_2} \in G$ and all $u > \epsilon$,

$$\begin{aligned}
&-\epsilon + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_2} - \epsilon}{q_{i_1}} \\
\leq&\mathbb{P}_x\Big(\widetilde{S}^{(n)}_{k+1} - S^{(n)}_k > u,\ \widetilde{W}^{(n)}_{k+1} = m_{i_2} \ \Big|\ W^{(n)}_k = m_{i_1}\Big) \\
\leq&\mathbb{P}_x\Big(\widetilde{S}^{(n)}_{k+1} - S^{(n)}_k > u - \eta^\theta_n,\ \widetilde{W}^{(n)}_{k+1} = m_{i_2} \ \Big|\ W^{(n)}_k = m_{i_1}\Big) \\
\leq&\epsilon + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_2} + \epsilon}{q_{i_1}}
\end{aligned} \tag{C.42}$$

- Here is one implication of (C.40). Since $|G| \le n_{\min}$, we have

$$\mathbb{P}_x\Big(S_k^{(n)} - \widetilde{S}_k^{(n)} \ge \eta_n^\theta \;\Big|\; \widetilde{W}_k^{(n)} = m_{i_1}\Big) < n_{\min} \cdot \epsilon \tag{C.43}$$

for all $k \ge 0$ and $m_{i_1} \in G^{\text{small}}$.

- Note that for any $m_{i_1}, m_{i_2} \in G^{\text{large}}$ and any $k \ge 0$

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{(n)} - S_k^{(n)} > u,\ W_{k+1}^{(n)} = m_{i_2} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
=&\mathbb{1}\{i_2 \ne i_1\}\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{(n)} - S_k^{(n)} > u,\ \widetilde{W}_{k+1}^{(n)} = m_{i_2} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
+&\sum_{i_3:\ m_{i_3} \in G^{\text{small}}} \int_{s>0} \mathbb{P}_x\Big(S_{k+1}^{(n)} - \widetilde{S}_{k+1}^{(n)} \ge (u-s)\vee 0,\ \widetilde{W}_{k+1}^{(n)} = m_{i_2} \;\Big|\; \widetilde{W}_k^{(n)} = m_{i_3}\Big) \\
&\qquad \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{(n)} - S_k^{(n)} = ds,\ \widetilde{W}_{k+1}^{(n)} = m_{i_3} \;\Big|\; W_k^{(n)} = m_{i_1}\Big).
\end{aligned}$$

Fix some i_3 with $m_{i_3} \in G^{\text{small}}$. On the one hand, due to (C.43),

$$\begin{aligned}
&\int_{s\in(0,u-\eta_n^\theta]} \mathbb{P}_x\Big(S_{k+1}^{(n)} - \widetilde{S}_{k+1}^{(n)} \ge u-s,\ \widetilde{W}_{k+1}^{(n)} = m_{i_2} \;\Big|\; \widetilde{W}_k^{(n)} = m_{i_3}\Big) \\
&\qquad \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{(n)} - S_k^{(n)} = ds,\ \widetilde{W}_{k+1}^{(n)} = m_{i_3} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
\le& n_{\min}\epsilon.
\end{aligned}$$

On the other hand, by considering the integral on $(u-\eta_n^\theta, \infty)$, we get

$$\begin{aligned}
&\int_{s\in(u-\eta_n^\theta,\infty)} \mathbb{P}_x\Big(S_{k+1}^{(n)} - \widetilde{S}_{k+1}^{(n)} \ge (u-s)\vee 0,\ \widetilde{W}_{k+1}^{(n)} = m_{i_2} \;\Big|\; \widetilde{W}_k^{(n)} = m_{i_3}\Big) \\
&\qquad \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{(n)} - S_k^{(n)} = ds,\ \widetilde{W}_{k+1}^{(n)} = m_{i_3} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
\ge& \int_{s\in(u,\infty)} \mathbb{P}_x\Big(\widetilde{W}_{k+1}^{(n)} = m_{i_2} \;\Big|\; \widetilde{W}_k^{(n)} = m_{i_3}\Big) \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{(n)} - S_k^{(n)} = ds,\ \widetilde{W}_{k+1}^{(n)} = m_{i_3} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
\ge& (p_{i_3,i_2} - \epsilon)\Big(-\epsilon + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_3} - \epsilon}{q_{i_1}}\Big)
\end{aligned}$$

due to (C.40) and (C.42). Meanwhile,

$$\begin{aligned}
&\int_{s\in(u-\eta_n^\theta,\infty)} \mathbb{P}_x\Big(S_{k+1}^{(n)} - \widetilde{S}_{k+1}^{(n)} \ge (u-s)\vee 0,\ \widetilde{W}_{k+1}^{(n)} = m_{i_2} \;\Big|\; \widetilde{W}_k^{(n)} = m_{i_3}\Big) \\
&\qquad \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{(n)} - S_k^{(n)} = ds,\ \widetilde{W}_{k+1}^{(n)} = m_{i_3} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
\le& (n_{\min}\epsilon + p_{i_3,i_2} + \epsilon)\Big(\epsilon + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_3} + \epsilon}{q_{i_1}}\Big)
\end{aligned}$$

due to (C.40), (C.42) and (C.43).

- Therefore, for any $m_{i_1}, m_{i_2} \in G^{\text{large}}$ and any $k \ge 0$,

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{(n)} - S_k^{(n)} > u,\ W_{k+1}^{(n)} = m_{i_2} \;\Big|\; W_k^{(n)} = m_{i_1}\Big) \\
\le& g(\epsilon) + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\mathbb{1}\{i_2 \ne i_1\}\nu_{i_1,i_2} + \sum_{i_3:\ m_{i_3}\in G^{\text{small}}} \nu_{i_1,i_3}p_{i_3,i_2}}{q_{i_1}}
\end{aligned}$$

$$\leq g(\epsilon) + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{q_{i_1,i_2}}{q_{i_1}} \tag{C.44}$$

and

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{(n)} - S_k^{(n)} > u,\ W_{k+1}^{(n)} = m_{i_2} \ \Big|\ W_k^{(n)} = m_{i_1}\Big)\\
\geq& -g(\epsilon) + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\mathbb{1}\{i_2 \neq i_1\}\nu_{i_1,i_2} + \sum_{i_3:\ m_{i_3}\in G^{\text{small}}}\nu_{i_1,i_3}p_{i_3,i_2}}{q_{i_1}}\\
\geq& -g(\epsilon) + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{q_{i_1,i_2}}{q_{i_1}} \qquad\qquad (C.45)
\end{aligned}$$

where $q^* = \max_i q_i$ and

$$g(\epsilon) \triangleq \epsilon + \frac{\epsilon}{q^*} + n_{\min}(1+\epsilon)\epsilon + \epsilon\frac{1+\epsilon}{q^*}n_{\min} + n_{\min}(\epsilon + \frac{\epsilon}{q^*}) + n_{\min}(n_{\min}\epsilon + \epsilon + 1)(1 + \frac{1}{q^*})\epsilon.$$

Note that $\lim_{\epsilon\downarrow 0} g(\epsilon) = 0$.

Now we apply the bounds in (C.39)(C.44)(C.45) to establish the weak convergence claim regarding $(S_0^{(n)}, W_0^{(n)}, \cdots, S_K^{(n)}, W_K^{(n)})$. Fix some positive integer K, some strictly positive real numbers $(s_k)_{k=0}^K$, a sequence $(w_k)_{k=0}^K \in \big(G^{\text{large}}\big)^{K+1}$ with $w_k = m_{i_k}$ for each k, and some $\epsilon > 0$ such that $\epsilon < \min_{k=0,1,\cdots,K}\{s_k\}$. On the one hand, the definition of the CTMC Y implies that

$$\begin{aligned}
&\mathbb{P}\Big(S_0 < t_0, W_0 = w_0;\ S_k > s_k \text{ and } W_k = w_k\ \forall k \in [K]\Big)\\
=&\mathbb{P}(\pi_G(m_i) = w_0)\prod_{k=1}^K\Big(S_k > s_k \text{ and } W_k = w_k \ \Big|\ W_{k-1} = w_{k-1}\Big)\\
=&\Big(\mathbb{1}\{m_i \in G^{\text{large}},\ i_0 = i\} + \mathbb{1}\{m_i \in G^{\text{small}}\}p_{i_0,i_1}\Big)\cdot\prod_{k=1}^K \exp(-q_{i_{k-1}}s_k)\frac{q_{i_{k-1},i_k}}{q_{i_{k-1}}}.
\end{aligned}$$

On the other hand, using (C.39)(C.44)(C.45), we know that for all n sufficiently large,

$$\begin{aligned}
&\mathbb{P}_x\Big(S_0^{(n)} < s_0, W_0^{(n)} = w_0;\ S_k^{(n)} > s_k \text{ and } W_k^{(n)} = w_k\ \forall k \in [K]\Big)\\
\geq&(1-\epsilon)\Big(\mathbb{1}\{m_i \in G^{\text{large}},\ i_0 = i\} + \mathbb{1}\{m_i \in G^{\text{small}}\}(p_{i_0,i_1} - \epsilon)\Big)\cdot\prod_{k=1}^K\Big(-g(\epsilon) + \exp(-(1+\epsilon)q_{i_{k-1}}s_k)\frac{q_{i_{k-1},i_k}}{q_{i_{k-1}}}\Big)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_x\Big(S_0^{(n)} < s_0, W_0^{(n)} = w_0;\ S_k^{(n)} > s_k \text{ and } W_k^{(n)} = w_k\ \forall k \in [K]\Big)\\
\leq&\Big(\mathbb{1}\{m_i \in G^{\text{large}},\ i_0 = i\} + \mathbb{1}\{m_i \in G^{\text{small}}\}(p_{i_0,i_1} + \epsilon)\Big)\cdot\prod_{k=1}^K\Big(g(\epsilon) + \exp(-(1-\epsilon)q_{i_{k-1}}s_k)\frac{q_{i_{k-1},i_k}}{q_{i_{k-1}}}\Big).
\end{aligned}$$

Since $\epsilon > 0$ can be arbitrarily small, we now obtain

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{P}_x\Big(S_0^{(n)} < t_0, W_0^{(n)} = w_0;\ S_k^{(n)} > s_k \text{ and } W_k^{(n)} = m_k\ \forall k \in [K]\Big)\\
=&\mathbb{P}\Big(S_0 < s_0, W_0 = w_0;\ S_k > s_k \text{ and } W_k = w_k\ \forall k \in [K]\Big),
\end{aligned}$$

and the arbitrariness of the integer K, the strictly positive real numbers $(s_k)_{k=0}^K$, and the sequence $(w_k)_{k=0}^K \in \big(G^{\text{large}}\big)^{K+1}$ allows us to conclude the proof. □

To extend the result above to the case where the communication class G is transient, we revisit the definition of the $Y^\dagger$ in (44). Let $\bar{G} = G^{\text{large}} \cup \{\dagger\}$ and let $m_0 = \dagger$ (remember that all the local minimizers of f on $[-L, L]$ are $m_1, \cdots, m_{\min}$). Meanwhile, using q_i and $q_{i,j}$ in (38)(39), we can define

$$q^\dagger_{i,j} = \begin{cases} q_{i,j} & \text{if } i \geq 1,\ j \geq 1 \\ \mathbb{1}\{j = 0\} & \text{if } i = 0 \\ \sum_{j \in [n_{\min}], m_j \notin G} q_{i,j} & \text{if } i \geq 1,\ j = 0. \end{cases}$$

and $q^\dagger_0 = 1,\ q^\dagger_i = q_i\ \forall i \geq 1$. Next, fix some i with $m_i \in G$ and $x \in \Omega_i$. Define a sequence of random variables $(S^\dagger_k)_{k\geq 0}, (W^\dagger_k)_{k\geq 0}$ such that $S^\dagger_k = 0$ and $W_0 = \pi_G(m_i), W^\dagger_0 = \dagger \mathbb{1}\{W_0 \notin G^{\text{large}}\} + W_0 \mathbb{1}\{W_0 \in G^{\text{large}}\}$ (see the definition of random mapping π_G in (40)) and (for all $k \geq 0$ and i, j with $m_j, m_l \in \bar{G}$)

$$\mathbb{P}\Big(W^\dagger_{k+1} = m_l,\ S^\dagger_{k+1} > t \ \Big|\ W^\dagger_k = m_j,\ (W^\dagger_l)^{k-1}_{l=0},\ (S^\dagger_l)^k_{l=0}\Big) \tag{C.46}$$

$$= \mathbb{P}\Big(W^\dagger_{k+1} = m_l,\ S^\dagger_{k+1} > t \ \Big|\ W^\dagger_k = m_j\Big) = \exp(-q^\dagger_j t)\frac{q^\dagger_{j,l}}{q^\dagger_j}\ \forall t > 0 \tag{C.47}$$

Then it is easy to see that $Y^\dagger(\pi_G(m_i))$ defined in (44) is a $\big((S^\dagger_k)_{k\geq 0}, (W^\dagger_k)_{k\geq 0}\big)$ jump process. In particular, from at any state that is not $\dagger$ (namely, any m_j with $m_j \in G^{\text{large}}$), the probability that $Y^\dagger$ moves to $\dagger$ in the next transition is equal to the chance that, starting from the same state, Y moves to a state that is not in G. Once entering $m_0 = \dagger$, the process $Y^\dagger$ will only make dummy jumps (with interarrival times being iid Exp(1)): indeed, we have $q^\dagger_0 = q^\dagger_{0,0} = 1$ and $q^\dagger_{0,j} = 0$ for any $j \geq 1$, implying that, given $W^\dagger_k = m_0 = \dagger$, we must have $W^\dagger_{k+1} = m_0 = \dagger$. These dummy jumps ensure that $Y^\dagger$ is stuck at the cemetery state $\dagger$ after visiting it.

Similarly, we can characterize the jump times and locations of the jump process $\hat{X}^{\dagger,*,\eta,\Delta}$ (for the definition, see (43)). When there is no ambiguity about the sequences $(\eta_n)_{n\geq 1}, (\Delta_n)_{n\geq 1}$, let $\hat{X}^{\dagger,(n)} = \hat{X}^{\dagger,*,\eta_n,\Delta_n}$ and $X^{\dagger,(n)} = X^{*,\eta_n,\Delta_n}$. Also, recall that τ_G defined in (13) is the step n when X^η_n exits the communication class G. Now let $(E_k)_{k\geq 0}$ be a sequence of iid Exp(1) random variables that is also independent of the noises $(Z_k)_{k\geq 1}$ (so they are independent from the SGD iterates X^η_n). For all $n \geq 1, k \geq 0$, define (see (24)-(29) and (33) for definitions of the quantities involved)

$$\widetilde{S}^{\dagger,(n)}_k = \begin{cases} \sigma^*_k(\eta_n, \Delta_n) \wedge \mathbf{T}^*(\tau_G(\eta_n), \eta_n) - \tau^*_{k-1}(\eta_n, \Delta_n) & \text{if } \tau^*_{k-1}(\eta_n, \Delta_n) < \mathbf{T}^*(\tau_G(\eta_n), \eta_n) \\ 0 & \text{otherwise} \end{cases}$$

$$S^{\dagger,(n)}_k = \begin{cases} \tau^*_k(\eta_n, \Delta_n) \wedge \mathbf{T}^*(\tau_G(\eta_n), \eta_n) - \tau^*_{k-1}(\eta_n, \Delta_n) & \text{if } \tau^*_{k-1}(\eta_n, \Delta_n) < \mathbf{T}^*(\tau_G(\eta_n), \eta_n) \\ E_k & \text{otherwise} \end{cases}$$

$$\widetilde{W}^{\dagger,(n)}_k = \begin{cases} m_{\widetilde{I}^G_k(\eta_n,\Delta_n)} & \text{if } \sigma^*_k(\eta_n, \Delta_n) < \mathbf{T}^*(\tau_G(\eta_n), \eta_n) \\ \dagger & \text{otherwise} \end{cases}$$

$$W^{\dagger,(n)}_k = \begin{cases} m_{I^G_k(\eta_n,\Delta_n)} & \text{if } \tau^*_k(\eta_n, \Delta_n) < \mathbf{T}^*(\tau_G(\eta_n), \eta_n) \\ \dagger & \text{otherwise} \end{cases}$$

with the convention that $\tau^*_{-1} = 0$. Note that $\mathbf{T}^*(\tau_G(\eta_n), \eta_n)$ is the scaled timestamp for $X^{(n)} = X^{*,\eta}$ corresponding to $\tau_G(\eta_n)$, hence $\mathbf{T}^*(\tau_G(\eta_n), \eta_n) = \min\{t \geq 0 : X^{(n)} \notin \bigcup_{j:\ m_j \in G} \Omega_j\}$. One can see that $\hat{X}^{\dagger,(n)}$ is a $\big((S^{\dagger,(n)}_k)_{k\geq 0}, (W^{\dagger,(n)}_k)_{k\geq 0}\big)$ jump process. The next lemma is similar to Lemma C.9 and discusses the convergence of the jump times and locations of $\hat{X}^{\dagger,(n)}$ on a communication class G in the transient case.

Lemma C.10. *Assume that the communication class G is transient. Given any $m_i \in G$, $x \in \Omega_i$, finitely many real numbers $(t_l)^{k'}_{l=1}$ such that $0 < t_1 < t_2 < \cdots < t_{k'}$, and a sequence of strictly positive*

real numbers $(\eta_n)_{n\geq 1}$ *with* $\lim_{n\to 0}\eta_n = 0$*, there exists a sequence of strictly positive real numbers* $(\Delta_n)_{n\geq 1}$ *with* $\lim_n \Delta_n = 0$ *such that*

- *Under* $\mathbb{P}_x$ *(so* $X_0^\eta = x$*), as* n *tends to* ∞*,*

$$(S_0^{\dagger,(n)}, W_0^{\dagger,(n)}, S_1^{\dagger,(n)}, W_1^{\dagger,(n)}, S_2^{\dagger,(n)}, W_2^{\dagger,(n)}, \cdots) \Rightarrow (S_0^{\dagger}, W_0^{\dagger}, S_1^{\dagger}, W_1^{\dagger}, S_2^{\dagger}, W_2^{\dagger}, \cdots) \tag{C.48}$$

- *For any* $l \in [k']$*,*

$$\lim_{n\to\infty} \mathbb{P}_x\Big(X_{t_l}^{\dagger,(n)} \notin \bigcup_{j:\ m_j\in G^{large}} B(m_j,\Delta_n),\ X_s^{\dagger,(n)} \in \bigcup_{j:\ m_j\in G} \Omega_j\ \forall s\in[0,t_l]\Big) = 0 \tag{C.49}$$

Proof. Let

$$\begin{aligned}
&\nu_{j,k} \triangleq \mu_j(E_{j,k})\ \forall j,k\geq 1,\ j\neq k\\
&p_{j,\dagger} \triangleq \sum_{\widetilde{j}:\ m_{\widetilde{j}}\notin G} p_{j,\widetilde{j}} \quad \forall m_j \in G^{\text{small}}\\
&q_{j,\dagger} \triangleq \sum_{k:m_k\notin G} \nu_{j,k} + \sum_{k:\ m_k\in G^{\text{small}}} \nu_{j,k}p_{k,\dagger} \quad \forall m_j\in G^{\text{large}}.
\end{aligned}$$

In order to specify our choice of $(\Delta_n)_{n\geq 1}$, we consider a construction of sequences $(\bar{\boldsymbol{\Delta}}(j))_{j\geq 0}, (\bar{\boldsymbol{\eta}}(j))_{j\geq 0}$ as follows. Fix some $\theta\in(0,\alpha-1)/2)$. Let $\bar{\boldsymbol{\Delta}}(0) = \bar{\boldsymbol{\eta}}(0) = 1$. One can see the existence of some $(\bar{\boldsymbol{\Delta}}(j))_{j\geq 1}, (\bar{\boldsymbol{\eta}}(j))_{j\geq 1}$ such that

- $\bar{\boldsymbol{\Delta}}(j)\in\big(0,\bar{\boldsymbol{\Delta}}(j-1)/2\big]$, $\bar{\boldsymbol{\eta}}(j)\in\big(0,\bar{\boldsymbol{\eta}}(j-1)/2\big]$ for all $j\geq 1$;
- (Due to Lemma C.2) for any $j\geq 1$, $\eta\in(0,\bar{\boldsymbol{\eta}}(j)]$, (remember that x and i are the fixed constants prescribed in the description of the lemma)

$$\mathbb{P}_x\Big(\sigma_0^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) < \eta^\theta,\ \widetilde{I}_0^G\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i\Big) > 1-\frac{1}{2^j}.$$

  For definitions of $\sigma_k^G, \tau_k^G, I_k^G, \widetilde{I}_k^G$, see (24)-(29).

- (Due to Lemma C.6) for any $j\geq 1$, $\eta\in(0,\bar{\boldsymbol{\eta}}(j)]$,

$$\Big|\mathbb{P}_x\Big(\tau_k^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) - \sigma_k^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) < \eta^\theta,\ I_k^G\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_2 \ \Big|\ \widetilde{I}_k^G\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big) - p_{i_1,i_2}\Big| < 1/2^j$$

  uniformly for all $k\geq 0$ and all $m_{i_1}\in G^{\text{small}}, m_{i_2}\notin G^{\text{small}}$. Also, by definition of σ^* and τ^*, we must have

$$\mathbb{P}_x\Big(\tau_k^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) - \sigma_k^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = 0,\ I_k^G\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1 \ \Big|\ \widetilde{I}_k^G\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big) = 1$$

  for all $k\geq 0$ and $m_{i_1}\in G^{\text{large}}$.

- (Due to Proposition B.24) for any $j\geq 1$, $\eta\in(0,\bar{\boldsymbol{\eta}}(j)]$,

$$\begin{aligned}
&-\frac{1}{2^j} + \exp\big(-(1+\frac{1}{2^j})q_{i_1}u\big)\frac{\nu_{i_1,i_2}-\frac{1}{2^j}}{q_{i_1}}\\
\leq&\mathbb{P}_x\Big(\sigma_{k+1}^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) - \tau_k^*\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) > u,\ \widetilde{I}_{k+1}^G = i_2 \ \Big|\ I_k^G\big(\eta,\bar{\boldsymbol{\Delta}}(j)\big) = i_1\Big)\\
\leq&\frac{1}{2^j} + \exp\big(-(1-\frac{1}{2^j})q_{i_1}u\big)\frac{\nu_{i_1,i_2}+\frac{1}{2^j}}{q_{i_1}}
\end{aligned}$$

  uniformly for all $k\geq 1$, all $u>1/2^j$, and all $m_{i_1}\in G^{\text{large}}, m_{i_2}\in G$.

- (Due to Lemma C.7) for any $j \geq 1$, for any $j \geq 1$, $\eta \in (0, \bar{\boldsymbol{\eta}}(j)]$,

$$\mathbb{P}_x\Big(X_{t_k}^{(n)} \notin \bigcup_{j:\ m_j \in G^{\mathrm{large}}} B(m_j, \bar{\boldsymbol{\Delta}}(j)),\ X_s^{(n)} \in \bigcup_{j:\ m_j \in G} \Omega_j\ \forall s \in [0, t_k]\Big) < 1/2^j \tag{C.50}$$

  uniformly for all $k \in [k']$.

Fix such $(\bar{\boldsymbol{\Delta}}(j))_{j\geq 0}, (\bar{\boldsymbol{\eta}}(j))_{j\geq 0}$. Define a function $\mathbf{J}(\cdot) : \mathbb{N} \mapsto \mathbb{N}$ as

$$\mathbf{J}(n) = 0 \vee \max\{j \geq 0 :\ \bar{\boldsymbol{\eta}}(j) \geq \eta_n\}$$

with the convention that $\max \emptyset = -\infty$. Lastly, let

$$\Delta_n = \bar{\boldsymbol{\Delta}}(\mathbf{J}(n))\ \forall n \geq 1.$$

Note that, due to $\lim_n \eta_n = 0$, we have $\lim_n \mathbf{J}(n) = \infty$, hence $\lim_n \Delta_n = 0$. Besides, since $X_t^{\dagger,(n)} = X_t^{(n)}$ given $X_s^{(n)} \in \bigcup_{j:\ m_j \in G} \Omega_j$ for all $s \in [0, t]$, by combining $\lim_n \mathbf{J}(n) = \infty$ with (C.50) we obtain (C.49).

Now it remains to prove (C.48). To this end, it suffices to show that, for any positive integer K, we have $(S_0^{\dagger,(n)}, W_0^{\dagger,(n)}, \cdots, S_K^{\dagger,(n)}, W_K^{\dagger,(n)})$ converges in distribution $(S_0^{\dagger}, W_0^{\dagger}, \cdots, S_K^{\dagger}, W_K^{\dagger})$ as n tends to infinity. In particular, due to introduction of the dummy jumps, we know that for any k with $\tau_k^*(\eta_n, \Delta_n) \geq \tau_G(\eta_n)$ (in other words, $\hat{X}^{\dagger,(n)}$ has reached state $\dagger$ within the first k jumps) we have $S_{k+1}^{\dagger,(n)} \sim \mathrm{Exp}(1)$ and $W_{k+1}^{\dagger,(n)} \equiv \dagger$. Similarly, for any k with $S_0^{\dagger} + \cdots + S_k^{\dagger} \leq \tau_G^Y$, we have $S_{k+1}^{\dagger} \sim \mathrm{Exp}(1)$ and $W_{k+1}^{\dagger} \equiv \dagger$. Therefore, it suffices to show that, for any positive integer K, any series of strictly positive real numbers $(s_k)_{k=0}^K$, any sequence $(w_k)_{k=0}^K \in \big(\bar{G}\big)^{K+1}$ such that $w_j \neq \dagger$ for any $j < K$, indices i_k such that $w_k = m_{i_k}$ for each k, we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_x\Big(S_0^{\dagger,(n)} < t_0, W_0^{\dagger,(n)} = w_0;\ S_k^{\dagger,(n)} > s_k \text{ and } W_k^{\dagger,(n)} = w_k\ \forall k \in [K]\Big) \\
=&\mathbb{P}\Big(S_0^{\dagger} < s_0, W_0^{\dagger} = w_0;\ S_k^{\dagger} > s_k \text{ and } W_k^{\dagger} = w_k\ \forall k \in [K]\Big)
\end{aligned} \tag{C.51}$$

Fix some $(s_k)_{k=0}^K$, $(w_k)_{k=0}^K \in \big(\bar{G}\big)^{K+1}$, and indices $(i_k)_{k=1}^K$ satisfying the conditions above. Besides, arbitrarily choose some $\epsilon > 0$ so that $\epsilon < \min_{k=0,\cdots,K} s_k$. To proceed, we start by translating the inequalities established above under the new system of notations.

- First of all, for all n sufficiently large, (remember that x and i are prescribed constants in the description of this lemma)

$$\mathbb{P}_x\Big(\widetilde{S}_0^{\dagger,(n)} < \eta_n^{\theta},\ \widetilde{W}_0^{\dagger,(n)} = m_i\Big) > 1 - \epsilon. \tag{C.52}$$

- For all $k \geq 0$ and all $m_{i_1} \in G^{\mathrm{small}}, m_{i_2} \in G^{\mathrm{large}}$, it holds for all n sufficiently large that

$$\Big|\mathbb{P}_x\Big(S_k^{\dagger,(n)} - \widetilde{S}_k^{\dagger,(n)} < \eta_n^{\theta},\ W_k^{\dagger,(n)} = m_{i_2} \ \Big|\ \widetilde{W}_k^{\dagger,(n)} = m_{i_1}\Big) - p_{i_1,i_2}\Big| < \epsilon. \tag{C.53}$$

- On the other hand, for all $k \geq 0$ and all $m_{i_1} \in G^{\mathrm{small}}$, it holds for all n sufficiently large that

$$\begin{aligned}
&\Big|\mathbb{P}_x\Big(S_k^{\dagger,(n)} - \widetilde{S}_k^{\dagger,(n)} < \eta_n^{\theta},\ W_k^{\dagger,(n)} = \dagger \ \Big|\ \widetilde{W}_k^{\dagger,(n)} = m_{i_1}\Big) - \sum_{i_2:\ m_{i_2} \notin G} p_{i_1,i_2}\Big| \\
=&\Big|\mathbb{P}_x\Big(S_k^{\dagger,(n)} - \widetilde{S}_k^{\dagger,(n)} < \eta_n^{\theta},\ W_k^{\dagger,(n)} = \dagger \ \Big|\ \widetilde{W}_k^{\dagger,(n)} = m_{i_1}\Big) - p_{i_1,\dagger}\Big| < \epsilon.
\end{aligned} \tag{C.54}$$

- For all $k \geq 0$ and all $m_{i_1} \in G^{\text{large}}$, it holds for all n that

$$\mathbb{P}_x\Big(S_k^{\dagger,(n)} - \widetilde{S}_k^{\dagger,(n)} = 0,\ W_k^{\dagger,(n)} = m_{i_1} \ \Big|\ \widetilde{W}_k^{\dagger,(n)} = m_{i_1}\Big) = 1. \tag{C.55}$$

- For all $k \geq 0$, all $m_{i_1} \in G^{\text{large}}, m_{i_2} \in G$ and all $u > \epsilon$, the following claim holds for all n sufficiently large:

$$\begin{aligned}
&-\epsilon + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_2} - \epsilon}{q_{i_1}} \\
\leq &\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} > u,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_2} \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
\leq &\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} > u - \eta_n^\theta,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_2} \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
\leq &\epsilon + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_2} + \epsilon}{q_{i_1}}
\end{aligned} \tag{C.56}$$

- On the other hand, for all $k \geq 0$, all $m_{i_1} \in G^{\text{large}}$, the following claim holds for all n sufficiently large:

$$\begin{aligned}
&-\epsilon + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{-\epsilon + \sum_{i_2:\ m_{i_2} \notin G} \nu_{i_1,i_2}}{q_{i_1}} \\
\leq &\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} > u,\ \widetilde{W}_{k+1}^{\dagger,(n)} = \dagger \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
\leq &\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} > u - \eta_n^\theta,\ \widetilde{W}_{k+1}^{\dagger,(n)} = \dagger \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
\leq &\epsilon + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\epsilon + \sum_{i_2:\ m_{i_2} \notin G} \nu_{i_1,i_2}}{q_{i_1}}
\end{aligned} \tag{C.57}$$

- Here is one implication of (C.53)(C.54). Since $|G| \leq n_{\min}$, we have (when n is sufficiently large)

$$\mathbb{P}_x\Big(S_k^{(n)} - \widetilde{S}_k^{(n)} \geq \eta_n^\theta \ \Big|\ \widetilde{W}_k^{(n)} = m_{i_1}\Big) < n_{\min} \cdot \epsilon \tag{C.58}$$

  for all $k \geq 0$ and $m_{i_1} \in G^{\text{small}}$.

- Note that for any $m_{i_1}, m_{i_2} \in G^{\text{large}}$ and any $k \geq 0$

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} > u,\ W_{k+1}^{\dagger,(n)} = m_{i_2} \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
= &\mathbb{1}\{i_2 \neq i_1\}\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} > u,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_2} \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
+ &\sum_{i_3:\ m_{i_3} \in G^{\text{small}}} \int_{s>0} \mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)} - \widetilde{S}_{k+1}^{\dagger,(n)} \geq (u-s) \vee 0,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_2} \ \Big|\ \widetilde{W}_k^{\dagger,(n)} = m_{i_3}\Big) \\
&\qquad \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} = ds,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_3} \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big).
\end{aligned}$$

  Fix some i_3 with $m_{i_3} \in G^{\text{small}}$. Due to (C.58)

$$\begin{aligned}
&\int_{s \in (0, u-\eta_n^\theta]} \mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)} - \widetilde{S}_{k+1}^{\dagger,(n)} \geq u - s,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_2} \ \Big|\ \widetilde{W}_k^{\dagger,(n)} = m_{i_3}\Big) \\
&\qquad \cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)} - S_k^{\dagger,(n)} = ds,\ \widetilde{W}_{k+1}^{\dagger,(n)} = m_{i_3} \ \Big|\ W_k^{\dagger,(n)} = m_{i_1}\Big) \\
\leq &n_{\min}\epsilon.
\end{aligned}$$

Meanwhile, by considering the integral on $(u-\eta_n^\theta,\infty)$, we get

$$\begin{aligned}
&\int_{s\in(u-\eta_n^\theta,\infty)} \mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)}-\widetilde{S}_{k+1}^{\dagger,(n)}\ge (u-s)\vee 0,\ \widetilde{W}_{k+1}^{\dagger,(n)}=m_{i_2}\ \Big|\ \widetilde{W}_k^{\dagger,(n)}=m_{i_3}\Big)\\
&\qquad\cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}=ds,\ \widetilde{W}_{k+1}^{\dagger,(n)}=m_{i_3}\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
\ge&\int_{s\in(u,\infty)} \mathbb{P}_x\Big(\widetilde{W}_{k+1}^{\dagger,(n)}=m_{i_2}\ \Big|\ \widetilde{W}_k^{\dagger,(n)}=m_{i_3}\Big)\cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}=ds,\ \widetilde{W}_{k+1}^{\dagger,(n)}=m_{i_3}\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
\ge&(p_{i_3,i_2}-\epsilon)\Big(-\epsilon+\exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_3}-\epsilon}{q_{i_1}}\Big)
\end{aligned}$$

due to (C.53) and (C.56). As for the upper bound,

$$\begin{aligned}
&\int_{s\in(u-\eta_n^\theta,\infty)} \mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)}-\widetilde{S}_{k+1}^{\dagger,(n)}\ge (u-s)\vee 0,\ \widetilde{W}_{k+1}^{\dagger,(n)}=m_{i_2}\ \Big|\ \widetilde{W}_k^{\dagger,(n)}=m_{i_3}\Big)\\
&\qquad\cdot \mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}=ds,\ \widetilde{W}_{k+1}^{\dagger,(n)}=m_{i_3}\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
\le&(n_{\min}\epsilon+p_{i_3,i_2}+\epsilon)\Big(\epsilon+\exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\nu_{i_1,i_3}+\epsilon}{q_{i_1}}\Big)
\end{aligned}$$

due to (C.53), (C.56) and (C.58).

- Therefore, for any $m_{i_1},m_{i_2}\in G^{\text{large}}$ and any $k\ge 0$,

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}>u,\ W_{k+1}^{\dagger,(n)}=m_{i_2}\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
\le&g(\epsilon)+\exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\mathbb{1}\{i_2\neq i_1\}\nu_{i_1,i_2}+\sum_{i_3:\ m_{i_3}\in G^{\text{small}}}\nu_{i_1,i_3}p_{i_3,i_2}}{q_{i_1}}\\
\le&g(\epsilon)+\exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{q_{i_1,i_2}}{q_{i_1}}
\end{aligned}\tag{C.59}$$

and

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}>u,\ W_{k+1}^{\dagger,(n)}=m_{i_2}\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
\ge&-g(\epsilon)+\exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\mathbb{1}\{i_2\neq i_1\}\nu_{i_1,i_2}+\sum_{i_3:\ m_{i_3}\in G^{\text{small}}}\nu_{i_1,i_3}p_{i_3,i_2}}{q_{i_1}}\\
\ge&-g(\epsilon)+\exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{q_{i_1,i_2}}{q_{i_1}}
\end{aligned}\tag{C.60}$$

where $q^*=\max_i q_i$ and

$$g(\epsilon)\triangleq 2\epsilon+\frac{\epsilon}{q^*}+n_{\min}(1+\epsilon)\epsilon+\epsilon\frac{1+\epsilon}{q^*}n_{\min}+n_{\min}(\epsilon+\frac{\epsilon}{q^*})+n_{\min}(n_{\min}\epsilon+\epsilon+1)(1+\frac{1}{q^*})\epsilon.$$

Note that $\lim_{\epsilon\downarrow 0}g(\epsilon)=0$.

- On the other hand, for the case where the marker process $\hat{X}^{\dagger,(n)}$ jumps to the cemetery state $\dagger$ from some $m_{i_1}\in G^{\text{large}}$, note that

$$\begin{aligned}
&\mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}>u,\ W_{k+1}^{\dagger,(n)}=\dagger\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
=&\mathbb{P}_x\Big(\widetilde{S}_{k+1}^{\dagger,(n)}-S_k^{\dagger,(n)}>u,\ \widetilde{W}_{k+1}^{\dagger,(n)}=\dagger\ \Big|\ W_k^{\dagger,(n)}=m_{i_1}\Big)\\
+&\sum_{i_3:\ m_{i_3}\in G^{\text{small}}}\int_{s>0}\mathbb{P}_x\Big(S_{k+1}^{\dagger,(n)}-\widetilde{S}_{k+1}^{\dagger,(n)}\ge(u-s)\vee 0,\ \widetilde{W}_{k+1}^{\dagger,(n)}=\dagger\ \Big|\ \widetilde{W}_k^{\dagger,(n)}=m_{i_3}\Big)
\end{aligned}$$

$$\cdot\, \mathbb{P}_x\Big(\widetilde{S}^{\dagger,(n)}_{k+1} - S^{\dagger,(n)}_k = ds,\ \widetilde{W}^{\dagger,(n)}_{k+1} = m_{i_3} \ \Big|\ W^{\dagger,(n)}_k = m_{i_1}\Big).$$

Arguing similarly as we did above by considering the integral on $[0, u-\eta^\theta_n]$ and $(u-\eta^\theta_n, \infty)$ separately, and using (C.54) and (C.57), we then get (for all n sufficiently large)

$$\begin{aligned}
&\mathbb{P}_x\Big(S^{\dagger,(n)}_{k+1} - S^{\dagger,(n)}_k > u,\ W^{\dagger,(n)}_{k+1} = \dagger \ \Big|\ W^{\dagger,(n)}_k = m_{i_1}\Big)\\
\leq& g(\epsilon) + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{\sum_{i_2:m_{i_2}\notin G}\nu_{i_1,i_2} + \sum_{i_2:m_{i_2}\in G^{\rm small}}\nu_{i_1,i_2}p_{i_2,\dagger}}{q_{i_1}}\\
=& g(\epsilon) + \exp\big(-(1-\epsilon)q_{i_1}u\big)\frac{q_{i_1,\dagger}}{q_{i_1}} \qquad \text{(C.61)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_x\Big(S^{\dagger,(n)}_{k+1} - S^{\dagger,(n)}_k > u,\ W^{\dagger,(n)}_{k+1} = \dagger \ \Big|\ W^{\dagger,(n)}_k = m_{i_1}\Big)\\
\geq& -g(\epsilon) + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{\sum_{i_2:m_{i_2}\notin G}\nu_{i_1,i_2} + \sum_{i_2:m_{i_2}\in G^{\rm small}}\nu_{i_1,i_2}p_{i_2,\dagger}}{q_{i_1}}\\
=& -g(\epsilon) + \exp\big(-(1+\epsilon)q_{i_1}u\big)\frac{q_{i_1,\dagger}}{q_{i_1}} \qquad \text{(C.62)}
\end{aligned}$$

For simplicity of presentation, we also let $q_{j,0} = q_{j,\dagger}$ and $p_{j,0} = p_{j,\dagger}$. First of all, remember that we have fixed some series of strictly positive real numbers $(s_k)_{k=0}^K$, some sequence $(w_k)_{k=0}^K \in \big(\bar{G}\big)^{K+1}$ such that $w_j \neq \dagger$ for any $j < K$, and indices i_k such that $w_k = m_{i_k}$ for each k. The definition of the continuous-time Markov chain $Y^\dagger$ implies that

$$\begin{aligned}
&\mathbb{P}\Big(S_0 < t_0, W_0 = w_0;\ S_k > s_k \text{ and } W_k = w_k\ \forall k \in [K]\Big)\\
=&\mathbb{P}(\pi_G(m_i) = w_0)\prod_{k=1}^K \Big(S_k > s_k \text{ and } W_k = m_k \ \Big|\ W_{k-1} = w_{k-1}\Big)\\
=&\Big(\mathbb{1}\{m_i \in G^{\rm large},\ i_0 = i\} + \mathbb{1}\{m_i \in G^{\rm small}\}p_{i_0,i_1}\Big)\cdot\prod_{k=1}^K \exp(-q_{i_{k-1}}s_k)\frac{q_{i_{k-1},i_k}}{q_{i_{k-1}}}.
\end{aligned}$$

On the other hand, using (C.59)-(C.62), we know that for all n sufficiently large,

$$\begin{aligned}
&\mathbb{P}_x\Big(S^{(n)}_0 < s_0, W^{(n)}_0 = w_0;\ S^{(n)}_k > s_k \text{ and } W^{(n)}_k = w_k\ \forall k \in [K]\Big)\\
\geq&(1-\epsilon)\Big(\mathbb{1}\{m_i \in G^{\rm large},\ i_0 = i\} + \mathbb{1}\{m_i \in G^{\rm small}\}(p_{i_0,i_1} - \epsilon)\Big)\cdot\prod_{k=1}^K \Big(-g(\epsilon) + \exp(-(1+\epsilon)q_{i_{k-1}}s_k)\frac{q_{i_{k-1},i_k}}{q_{i_{k-1}}}\Big)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_x\Big(S^{(n)}_0 < s_0, W^{(n)}_0 = w_0;\ S^{(n)}_k > s_k \text{ and } W^{(n)}_k = m_k\ \forall k \in [K]\Big)\\
\leq&\Big(\mathbb{1}\{m_i \in G^{\rm large},\ i_0 = i\} + \mathbb{1}\{m_i \in G^{\rm small}\}(p_{i_0,i_1} + \epsilon)\Big)\cdot\prod_{k=1}^K \Big(g(\epsilon) + \exp(-(1-\epsilon)q_{i_{k-1}}s_k)\frac{q_{i_{k-1},i_k}}{q_{i_{k-1}}}\Big).
\end{aligned}$$

The arbitrariness of $\epsilon > 0$ then allows us to establish (C.51) and conclude the proof. □

Now we are ready to prove Lemma 10 and Lemma 11.

Proof of Lemma 10. From Lemma C.9, one can see the existence of some $(\Delta_n)_{n\geq 1}$ with $\lim_n \Delta_n = 0$ such that (C.32)-(C.35) hold. For simplicity of notations, we let $\hat{X}^{(n)} = \hat{X}^{*,\eta_n,\Delta_n}$, $X^{(n)} = X^{*,\eta_n}$, and let $\bar{t} = t_{k'}$

Combine (C.32) with Lemma C.8, and we immediately get (41). In order to prove (42), it suffices to show that for any $\epsilon > 0$,

$$\limsup_n \mathbb{P}_x\Big(X^{(n)}_{t_k} \notin \bigcup_{j:\ m_j \in G^{\rm large}} B(m_j, \Delta_n)\Big) \leq 4\epsilon\ \forall k \in [k'].$$

Fix $\epsilon > 0$, and observe following bound by decomposing the events

$$\begin{aligned}
&\mathbb{P}_x\Big(X^{(n)}_{t_k} \notin \bigcup_{j:\ m_j \in G^{\rm large}} B(m_j, \Delta_n)\Big)\\
\leq&\mathbb{P}_x\Big(X^{(n)}_{t_k} \notin \bigcup_{j:\ m_j \in G^{\rm large}} B(m_j, \Delta_n),\ X^{(n)}_t \in \bigcup_{j:\ m_j \in G} \Omega_j\ \forall t \in [0, \bar{t}]\Big)\\
+&\mathbb{P}_x\Big(\exists t \in [0,\bar{t}] \text{ such that } X^{(n)}_t \notin \bigcup_{j:\ m_j \in G} \Omega_j\Big)
\end{aligned}$$

Therefore, given (C.35), it suffices to prove

$$\limsup_n \mathbb{P}_x\Big(\exists t \in [0,\bar{t}] \text{ such that } X^{(n)}_t \notin \bigcup_{j:\ m_j \in G} \Omega_j\Big) \leq 3\epsilon. \tag{C.63}$$

Let

$$\begin{aligned}
T^{(n)}_0 &\triangleq \min\{t \geq 0:\ X^{(n)}_t \in \bigcup_{j: m_j \in G} B(m_j, 2\Delta_n)\}\\
I^{(n)}_0 = j &\iff X^{(n)}_{T^{(n)}_0} \in B(m_j, 2\Delta_n)\\
T^{(n)}_k &\triangleq \min\{t > T^{(n)}_{k-1}:\ X^{(n)}_t \in \bigcup_{j: m_j \in G,\ j \neq I^{(n)}_{k-1}} B(m_j, 2\Delta_n)\}\\
I^{(n)}_k = j &\iff X^{(n)}_{T^{(n)}_k} \in B(m_j, 2\Delta_n).
\end{aligned}$$

Building upon this definition, we define the following stopping times and marks that only records the hitting time to minimizer in *large* attraction fields in G (with convention $\mathbf{k}^{(n),\rm large}(-1) = -1, T^{(n),\rm large}_{-1} = 0, T^{(n)}_{-1} = 0$)

$$\begin{aligned}
\mathbf{k}^{(n),\rm large}(k) &\triangleq \min\{l > \mathbf{k}^{(n),\rm large}(k-1):\ m_{I^{(n)}_l} \in G^{\rm large}\}\\
T^{(n),\rm large}_k &\triangleq T^{(n)}_{\mathbf{k}^{(n),\rm large}(k)}, \quad I^{(n),\rm large}_k \triangleq I^{(n)}_{\mathbf{k}^{(n),\rm large}(k)}.
\end{aligned}$$

Now by defining

$$\begin{aligned}
J^{(n)}(t) &\triangleq \#\{k \geq 0:\ T^{(n)}_k \leq t\},\\
J^{(n)}_{\rm large}(t) &\triangleq \#\{k \geq 0:\ T^{(n),\rm large}_k \leq t\},\\
J^{(n)}(s,t) &\triangleq \#\{k \geq 0:\ T^{(n)}_k \in [s,t]\},
\end{aligned}$$

we use $J^{(n)}(t)$ to count the numbers of visits to local minima on G, and $J^{(n)}_{\rm large}(t)$ for the number of visits to minimizers in the *large* attraction fields on G. $J^{(n)}(s,t)$ counts the indices k such that at $T^{(n)}_k$ a minimizer on G is visited and regarding the hitting time we have $T^{(n)}_k \in [s,t]$.

First of all, the weak convergence result in (C.32) implies the existence of some positive integer $N(\epsilon)$ such that

$$\limsup_n \mathbb{P}_x(J^{(n)}_{\text{large}}(\bar{t}) > N(\epsilon)) < \epsilon.$$

Fix such $N(\epsilon)$. Next, from (C.34), we know the existence of some integer $K(\epsilon)$ such that

$$\limsup_n \sup_{k\geq 0} \mathbb{P}_x\Big(J^{(n)}\big(T^{(n),\text{large}}_{k-1}, T^{(n),\text{large}}_k\big) > K(\epsilon)\Big) \leq \epsilon/N(\epsilon).$$

Fix such $K(\epsilon)$ as well. From the results above, we know that for event

$$A_1(n) \triangleq \{J^{(n)}_{\text{large}}(\bar{t}) \leq N(\epsilon)\} \cap \Big\{J^{(n)}\big(T^{(n),\text{large}}_{k-1}, T^{(n),\text{large}}_k\big) \leq K(\epsilon)\ \forall k \in [N(\epsilon)]\Big\},$$

we have $\limsup_n \mathbb{P}_x\Big(\big(A_1(n)\big)^c\Big) \leq 2\epsilon$. On the other hand, on event $A_1(n)$, we must have

$$J^{(n)}(\bar{t}) \leq N(\epsilon)K(\epsilon).$$

Meanwhile, it follows immediately from (C.33) that

$$\limsup_n \sup_{k\geq 0} \mathbb{P}\Big(\exists t \in [T^{(n)}_{k-1}, T^{(n)}_k] \text{ such that } X^{(n)}_t \notin \bigcup_{j:\ m_j\in G} \Omega_j\Big) < \frac{\epsilon}{N(\epsilon)K(\epsilon)},$$

hence for event

$$A_2(n) \triangleq \Big\{X^{(n)}_t \in \bigcup_{j:\ m_j\in G} \Omega_j\ \forall t \in [0, T^{(n)}_{N(\epsilon)K(\epsilon)}]\Big\},$$

we must have $\limsup_n \mathbb{P}_x\Big(\big(A_2(n)\big)^c\Big) \leq \epsilon$. To conclude the proof, note that

$$\begin{aligned}
A_1(n)\cap A_2(n) &\subseteq \{J^{(n)}(\bar{t}) \leq N(\epsilon)K(\epsilon)\} \cap \Big\{X^{(n)}_t \in \bigcup_{j:\ m_j\in G} \Omega_j\ \forall t \in [0, T^{(n)}_{N(\epsilon)K(\epsilon)}]\Big\} \\
&= \{T^{(n)}_{N(\epsilon)K(\epsilon)} \geq \bar{t}\} \cap \Big\{X^{(n)}_t \in \bigcup_{j:\ m_j\in G} \Omega_j\ \forall t \in [0, T^{(n)}_{N(\epsilon)K(\epsilon)}]\Big\} \\
&\subseteq \Big\{X^{(n)}_t \in \bigcup_{j:\ m_j\in G} \Omega_j\ \forall t \in [0, \bar{t}]\Big\}
\end{aligned}$$

so we have established (C.63). □

Proof of Lemma 11. From Lemma C.10, one can see the existence of some $(\Delta_n)_{n\geq 1}$ with $\lim_n \Delta_n = 0$ such that (C.48) and (C.49) hold. For simplicity of notations, we let $\hat{X}^{\dagger,(n)} = \hat{X}^{\dagger,*,\eta_n,\Delta_n}, X^{\dagger,(n)} = X^{\dagger,*,\eta_n}$, and let $\bar{t} = t_{k'}$

Combine (C.48) with Lemma C.8, and we immediately get (45). In order to prove (46), note that

$$\begin{aligned}
&\Big\{X^{\dagger,(n)}_{t_k} \notin \bigcup_{j:\ m_j\in G^{\text{large}}} B(m_j, \Delta_n) \text{ and } X^{\dagger,(n)}_{t_k} \neq \dagger\Big\} \\
=&\Big\{X^{\dagger,(n)}_{t_k} \notin \bigcup_{j:\ m_j\in G^{\text{large}}} B(m_j, \Delta_n) \text{ and } X^{\dagger,(n)}_{s} \in \bigcup_{j:m_j\in G} \Omega_j\ \forall s \in [0, t_k]\Big\}
\end{aligned}$$

so the conclusion of the proof follows directly from (C.49). □

D Proofs of Lemma B.1,B.2,B.3

Proof of Lemma B.1. For any $\epsilon > 0$,

$$\mathbb{P}\Big(U(\epsilon) > \frac{1}{b(\epsilon)}\Big) = \Big(1 - a(\epsilon)\Big)^{\lfloor 1/b(\epsilon)\rfloor}.$$

By taking logarithm on both sides, we have

$$\begin{aligned}\ln \mathbb{P}\Big(U(\epsilon) > \frac{1}{b(\epsilon)}\Big) &= \lfloor 1/b(\epsilon)\rfloor \ln\Big(1 - a(\epsilon)\Big) \\ &= \frac{\lfloor 1/b(\epsilon)\rfloor}{1/b(\epsilon)} \frac{\ln\Big(1 - a(\epsilon)\Big)}{-a(\epsilon)} \frac{-a(\epsilon)}{b(\epsilon)}.\end{aligned}$$

Since $\lim_{x\to 0} \frac{\ln(1+x)}{x} = 1$, we know that for ϵ sufficiently small, we will have

$$-c\frac{a(\epsilon)}{b(\epsilon)} \le \ln \mathbb{P}\Big(U(\epsilon) > \frac{1}{b(\epsilon)}\Big) \le -\frac{a(\epsilon)}{c \cdot b(\epsilon)}. \tag{D.1}$$

By taking exponential on both sides, we conclude the proof. □

Proof of Lemma B.2. To begin with, for any $\epsilon > 0$ we have

$$\mathbb{P}\Big(U(\epsilon) \le \frac{1}{b(\epsilon)}\Big) = 1 - \mathbb{P}\Big(U(\epsilon) > \frac{1}{b(\epsilon)}\Big).$$

Using bound (D.1), we know that for ϵ sufficiently small, $\mathbb{P}(U(\epsilon) > 1/b(\epsilon)) \ge \exp(-c \cdot a(\epsilon)/b(\epsilon))$. The upper bound follows from the generic bound $1 - \exp(-x) \le x,\ \forall x \in \mathbb{R}$ with $x = c \cdot a(\epsilon)/b(\epsilon)$.

Now we move onto the lower bound. Again, from bound (D.1), we know that for sufficiently small ϵ, we will have

$$\mathbb{P}\Big(U(\epsilon) \le \frac{1}{b(\epsilon)}\Big) \ge 1 - \exp(-\frac{1}{\sqrt{c}} \cdot \frac{a(\epsilon)}{b(\epsilon)}).$$

Due to the assumption that $\lim_{\epsilon\downarrow 0} a(\epsilon)/b(\epsilon) = 0$ and the fact that $1 - \exp(-x) \ge \frac{x}{\sqrt{c}}$ for $x > 0$ sufficiently close to 0, we will have (for ϵ small enough) $\mathbb{P}\Big(U(\epsilon) \le \frac{1}{b(\epsilon)}\Big) \ge \frac{1}{c} \cdot \frac{a(\epsilon)}{b(\epsilon)}$. □

Proof of Lemma B.3. Let $a_k \triangleq x_k - \widetilde{x}_k$. From the fact that $g \in \mathcal{C}^2$ and Taylor expansion of g', one can easily see that

$$\begin{aligned} a_k &= \eta \sum_{j=1}^{k} \Big(g'(\widetilde{x}_{j-1}) - g'(x_{j-1})\Big) + \eta(z_1 + \cdots + z_k) + x - \widetilde{x}; \\ \Rightarrow |a_k| &\le \eta C(|a_0| + \cdots |a_{k-1}|) + \widetilde{c}. \end{aligned}$$

The desired bound then follows immediately from Gronwall's inequality (see Theorem 68, Chapter V of [24], where we let function $\alpha(t)$ be $\alpha(t) = |a_{\lfloor t\rfloor}|$). □

E Notations

Table E.1 lists the notations used in Appendix B.

Table E.1: Summary of notations frequently used in Appendix B

$[k]$	$\{1, 2, \ldots, k\}$	
η	Learning rate (gradient descent step size)	
b	Truncation threshold of stochastic gradient	
ϵ	An accuracy parameter; typically used to denote an $\epsilon-$neighborhood of s_i, m_i	
δ	A threshold parameter used to define *large* noises	
$\bar{\epsilon}$	A constant defined for (B.15)-(B.16). Since $\bar{\epsilon} < \epsilon_0$, in (B.1) the claim holds for $\lvert x-y\rvert < \bar{\epsilon}$. Note that the value of the constant $\bar{\epsilon}$ does not vary with our choice of η, ϵ, δ.	
M	Upper bound of $\lvert f'\rvert$ and $\lvert f''\rvert$	(B.3)
L	Radius of training domain	(B.3)
Ω	The open interval (s_-, s_+); a simplified notation for Ω_i	
φ, φ_c	$\varphi_c(w) \triangleq \varphi(w, c) \triangleq (w \wedge c) \vee (-c)$	truncation operator at level $c > 0$
$Z_n^{\leq\delta,\eta}$	$Z_n \mathbb{1}\{\eta\lvert Z_n\rvert \leq \delta\}$	"small" noise (B.7)
$Z_n^{>\delta,\eta}$	$Z_n \mathbb{1}\{\eta\lvert Z_n\rvert > \delta\}$	"large" noise (B.8)
$T_j^\eta(\delta)$	$\min\{n > T_{j-1}^\eta(\delta) :\ \eta\lvert Z_n\rvert > \delta\}$	arrival time of j-th large noise (B.9)
$W_j^\eta(\delta)$	$Z_{T_j^\eta(\delta)}$	size of j-th large noise (B.10)
$X_n^\eta(x)$	$X_{n+1}^\eta(x) = \varphi_L\Big(X_n^\eta(x) - \varphi_b\big(\eta(f'(X_n^\eta(x)) - Z_{n+1})\big)\Big), \quad X_0^\eta(x) = x$	SGD
$\mathbf{y}_n^\eta(x)$	$\mathbf{y}_n^\eta(x) = \mathbf{y}_{n-1}^\eta(x) - \eta f'(\mathbf{y}_{n-1}^\eta(x)), \quad \mathbf{y}_0^\eta(x) = x$	GD
$Y_n^\eta(x)$	$\mathbf{y}_n^\eta(x)$ perturbed by large noises $(\mathbf{T}^\eta(\delta), \mathbf{W}^\eta(\delta))$	GD + large jump
$\widetilde{\mathbf{y}}_n^\eta(x; \mathbf{t}, \mathbf{w})$	$\mathbf{y}_n^\eta(x)$ perturbed by noise vector $(\mathbf{t}, \mathbf{w})$	perturbed GD
$\mathbf{x}^\eta(t, x)$	$d\mathbf{x}^\eta(t; x) = -\eta f'\Big(\mathbf{x}^\eta(t; x)\Big)dt, \quad \mathbf{x}^\eta(0; x) = x$	ODE
$\mathbf{x}(t, x)$	$\mathbf{x}^1(t, x)$	
$\widetilde{\mathbf{x}}^\eta(t, x; \mathbf{t}, \mathbf{w})$	$\mathbf{x}^\eta(t, x)$ perturbed by noise vector $(\mathbf{t}, \mathbf{w})$	perturbed ODE
$A(n, \eta, \epsilon, \delta)$	$\Big\{\max_{k\in[n\wedge(T_1^\eta(\delta)-1)]} \eta\lvert Z_1 + \cdots + Z_k\rvert \leq \epsilon\Big\}$.	(B.28)
r	$r \triangleq \min\{-s_-, s_+\}$. Effective radius of the attraction field Ω.	
l^*	$l^* \triangleq \lceil r/b\rceil$. The minimum number of jumps required to escape Ω when starting from its local minimum $m = 0$.	
$h(\mathbf{w}, \mathbf{t})$	A mapping defined as $h(\mathbf{w}, \mathbf{t}) = \widetilde{\mathbf{x}}(t_{l^*}, 0; \mathbf{t}, \mathbf{w})$.	
$\bar{t},\ \bar{\delta}$	Necessary conditions for $h(\mathbf{w}, \mathbf{t})$ to be outside of Ω	(B.21)-(B.22)
$\hat{t}(\epsilon)$	$\hat{t}(\epsilon) \triangleq c_1 \log(1/\epsilon)$. The quantity $\hat{t}(\epsilon)/\eta$ provides an upper bound for the time it takes $\mathbf{x}^\eta$ to return to $2\epsilon-$neighborhood of local minimum $m = 0$ when starting from somewhere $\epsilon-$away from s_-, s_+. See (B.23).	
$E(\epsilon)$	$\big\{(\mathbf{w}, \mathbf{t}) \subseteq \mathbb{R}^{l^*} \times \mathbb{R}_+^{l^*-1} : h(\mathbf{w}, \mathbf{t}) \notin [(s_- - \epsilon) \vee (-L), (s_+ + \epsilon) \wedge L]\big\}$	
$p(\epsilon, \delta, \eta)$	The probability that, for $\mathbf{t} = \big(T_j^\eta(\delta) - 1\big)_{j=1}^{l^*}$ and $\mathbf{w} = \big(\eta W_j^\eta(\delta)\big)_{j=1}^{l^*}$, we have $(\mathbf{w}, \mathbf{t}) \in E(\epsilon)$ conditioning on $\{T_1^\eta(\delta) = 1\}$. Intuitively speaking, it characterizes the probability that the first l^* *large* noises alone can drive the ODE out of the attraction field. Defined in (B.53).	

ν_α	The Borel measure on $\mathbb{R}$ with density $$\nu_\alpha(dx) = \mathbb{1}\{x > 0\}\frac{\alpha p_+}{x^{\alpha+1}} + \mathbb{1}\{x < 0\}\frac{\alpha p_-}{|x|^{\alpha+1}}$$ where p_-, p_+ are constants in Assumption 2 in the main paper.	
μ	The product measure $\mu = (\nu_\alpha)^{l^*} \times (\mathbf{Leb}_+)^{l^*-1}$.	
$\sigma(\eta)$	$\min\{n \geq 0 : X_n^\eta \notin \Omega\}$.	first exit time
$H(x)$	$\mathbb{P}(\lvert Z_1\rvert > x) = x^{-\alpha}L(x)$	
$T_{\text{return}}(\epsilon, \eta)$	$\min\{n \geq 0 : X_n^\eta(x) \in [-2\epsilon, 2\epsilon]\}$	