Asymmetric Heavy Tails and Implicit Bias in Gaussian Noise Injections

SUPPLEMENTARY DOCUMENT

The supplementary document is organised as follows.

1. The supplementary document begins first with a presentation of additional experiments that are referenced directly in the main text (Section A).
2. We then cover the cost-functions used to train neural networks in Section B; and give an overview in Section C of the other potential sources of the implicit effect gradient noise skew which we explored.
3. In Section D, we provide an overview of the assumptions we will be making in our analysis. We then describe in Section E the numerical method we use to approximate the drift term $b(\mathbf{w}, \alpha, \theta)$ defined in (4.7).
4. We end with metastability analysis of asymmetric stable processes (Section F); followed by the technical proofs of the lemmas, theorems, and corollaries that we present in the main body and the supplementary document of the paper (Section G).

Before beginning the supplementary document we make a quick note of network architectures and training hyper-parameters.

Network Architectures Networks were trained using stochastic gradient descent with a learning rate of 0.0003 and batch sizes specified in text. MLP network architectures are specified in text. Convolutional (CONV) networks are 2 hidden layer networks. The first layer has 32 filters, a kernel size of 4, and a stride length of 2. The second layer has 128 filters, a kernel size of 4, and a stride length of 2. The final output layer is a dense layer.

A. Additional Experimental Results

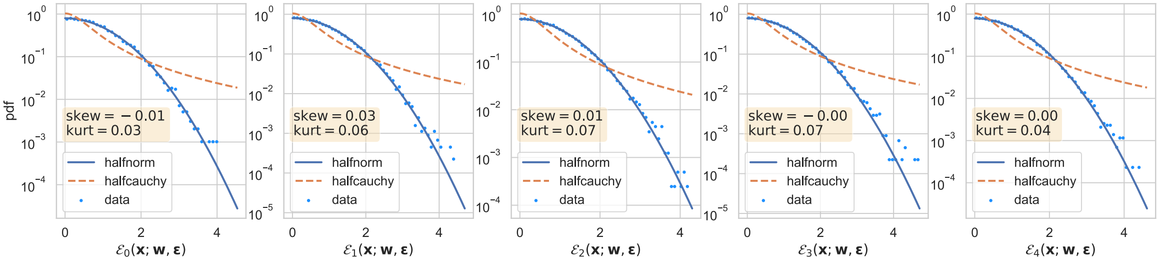


Figure A.1. We measure the skewness and kurtosis *at initialisation* of the noise accumulated on network activations at each layer i for a 4-layer 256-unit wide MLP trained to regress the function $\lambda(\mathbf{x}) = \sum_i \sin(2\pi q_i \mathbf{x} + \phi(i))$ with $q_i \in (5, 10, \ldots, 45, 50),\ \mathbf{x} \in \mathbb{R}$ and experiencing additive-GNIs. We plot the probability density function of positive samples, comparing against half-normal (non-heavy-tailed) and half-Cauchy (heavy-tailed) distributions, where $\mathcal{E}_i(\mathbf{x}; \mathbf{w}, \boldsymbol{\epsilon})$ is defined in (2.9). Each blue point represents the noise on an individual activation in a layer i for a point $\mathbf{x}$. This noise is Gaussian (low skew and kurtosis) with a p.d.f. that tracks that of a half-normal.

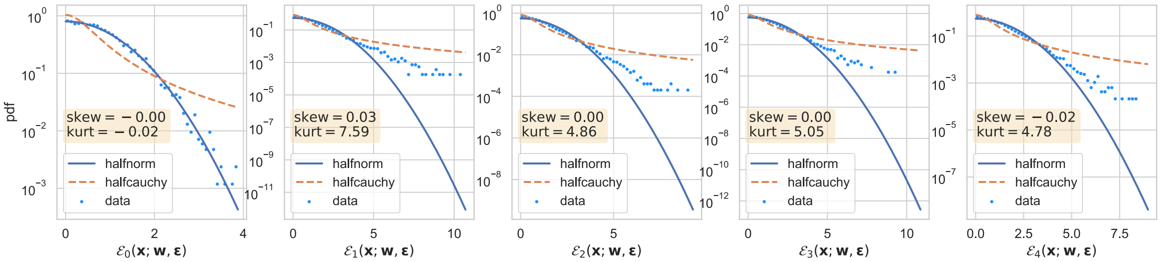


Figure A.2. Here we show the same plots as in Figure A.1 but for multiplicative-GNIs. The forward pass here experiences *symmetric heavy-tailed noise* for all layers past the data layer.

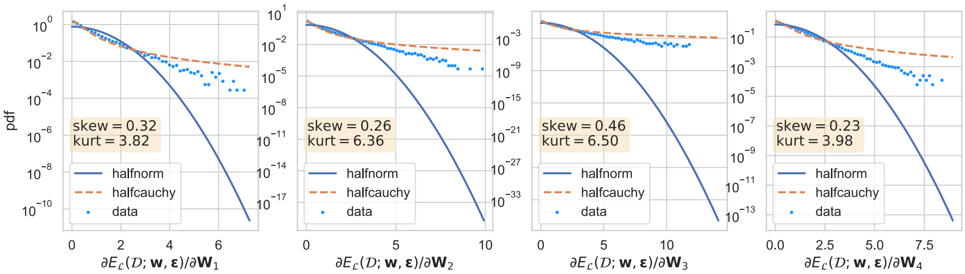
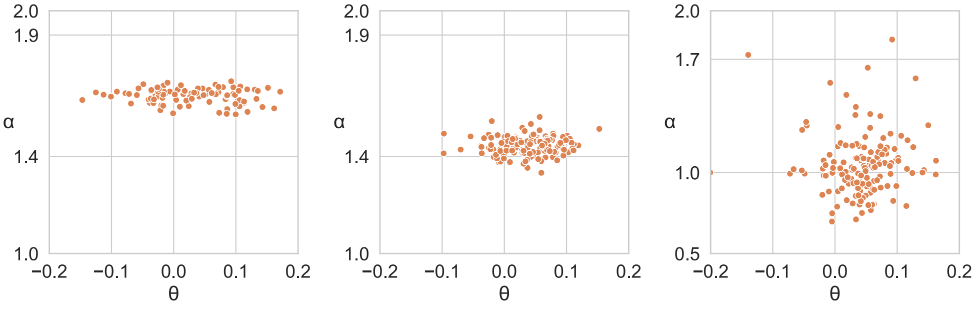


Figure A.3. Here we show the same plots as in Figure 2 but for multiplicative-GNIs. The gradient noise is *skewed and heavy-tailed*, with a p.d.f. that is more Cauchy-like than Gaussian. The kurtosis decays as the gradients approach the input layer, as predicted by Theorem 3.1.



(a) CIFAR10 (×)

(b) SVHN (×)

(c) BHP (+)

Figure A.4. We model $\nabla E_{\mathcal{L}}(\cdot)$ as being drawn from some α-stable distribution $\mathcal{S}_\alpha$ and estimate the tail-index α and skewness θ using maximum likelihood estimation as in Nolan (2001). We plot the results as a scatter for a batch of size $B = 512$ for CIFAR10 and SVHN and $B = 32$ for Boston House Prices. Additive (+) and multiplicative (×) GNIs have $\sigma^2 = 0.1$.

B. Cost Functions

B.1. Mean Square Error

In the case of regression the most commonly used loss is the mean-square error.

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = (\mathbf{y} - \mathbf{h}_L(\mathbf{x}))^2 \ .$$

B.2. Cross Entropy Loss

In the case of classification, we use the cross-entropy loss. If we consider our network outputs $\mathbf{h}_L$ to be the pre-$\mathrm{softmax}$ of logits of the final layer then the loss is for a data-label pair $(\mathbf{x}, \mathbf{y})$

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\sum_{c=0}^{C} \mathbf{y}_c \log\left(\mathrm{softmax}(\mathbf{h}_L(\mathbf{x}))_c\right) \ , \tag{B.1}$$

where c indexes over the C possible classes of the classification problem.

C. Other Potential Sources of the Skewness in the Gradient Noise

The product of correlated random variables can be skewed (Oliveira et al., 2016; Nadarajah & Pogány, 2016). Our first hypothesis was that the skew came from the correlation of $(\partial E_{\mathcal{L}}(\cdot)/\partial h_i^m)$ and $(\partial h_i^m/\partial W_{i,l,j})$. As a test, Figure C.5 of the Appendix reproduces Figure 2 with linear κ, isolating gradient correlation as a potential source of skew. Gradients are not skewed, demonstrating that the asymmetry stems from non-linear κ.

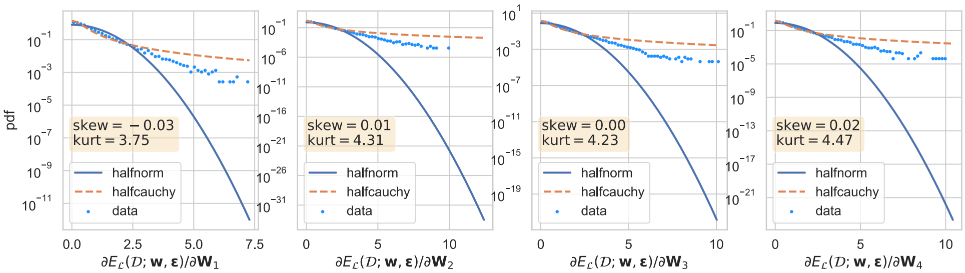


Figure C.5. Here we reproduce Figure 2 but with no non-linearities. This gradient noise is clearly *heavy-tailed* but *not* skewed.

D. Overview of the Assumptions

Due to space limitations and to avoid obscuring the main take home messages of our theoretical results, we did not present the two assumption required for Theorem 4.2. These two assumptions are properly presented in their respective sections (Sections E and G.6), where we first provide the required technical context for defining them in each section. In this section, we will shortly discuss the semantics of these assumptions from a higher-level perspective for the convenience of the reader.

- **Assumption E.1.** This assumption is essentially an assumption of the tails of the function $\partial\varphi$, with $\varphi(w) = e^{-\varepsilon^{-\alpha} f(w)}$. In particular, in order to make our approximation scheme (to the fractional derivatives) convergent, this assumption makes sure that outside of a compact region, the function $\partial\varphi$ exponentially decays.
- **Assumption G.1.** This assumption enforces a certain structure on the Euler-Maruyama discretisation given in Section 4:

  $$\tilde{\mathbf{w}}_{n+1} = \tilde{\mathbf{w}}_n + \eta_{n+1} b_{h,K}(\tilde{\mathbf{w}}_n, \alpha, \theta) + \varepsilon \eta_{n+1}^{1/\alpha} \Delta \mathbf{L}_{n+1}^{\alpha,\theta}.$$

  As a first condition, we make sure that the step-sizes are decreasing while their sum is diverging, which is a standard assumption. The second condition is essentially a Lyapunov condition that requires the modified drift b behaves well, so that we can control the weak error of the sample averages by using (Panloup, 2008). The final condition is similar to the second condition in nature, and requires ergodicity of an SDE defined through the approximate drift $b_{h,K}$, in order to enable us link the weak error to the error induced by the approximation scheme used for the fractional derivatives.

E. Fractional Differentiation and the Approximation Scheme

In this section, we provide the details of the Riesz-Feller type fractional derivative $\mathcal{D}^{\alpha-2,-\theta}$, whose definition was omitted in the main document for clarity. We then present the details of the approximation method for the drift term $b(\mathbf{w},\alpha,\theta)$ defined in (4.7).

The building block of our analysis is a first-order approximation of $\mathcal{D}^{\alpha-2,-\theta}\partial_w\varphi$ for any $\partial_w\varphi \in L^1(\mathbb{R})\cap\mathcal{C}^4(\mathbb{R})$. We consider the one-dimensional case for simplicity since the Lévy motion we consider has independent coordinates, and the multi-dimensional numerical approximation can be reduced to the one-dimensional case. Assume the tail index $1<\alpha<2$ and the skewness parameter satisfies $-1<\theta<1$.

When $\theta = 0$, Şimşekli (2017) developed the numerical approximation method for the drift term $b(\mathbf{w},\alpha,0)$ by approximately computing the Riesz potential[3] $\mathcal{D}^{\gamma}$ via the fractional centred difference method provided by Ortigueira (2006); Ortiguera (2006b). It is shown that for any $-1<\gamma<0$, we have the following numerical error,

$$\left|\mathcal{D}^{\gamma}\partial_w\varphi(w) - \Delta^{\gamma}_{h,K}\partial_w\varphi(w)\right| = \mathcal{O}\left(h^2 + 1/(hK)\right)\,,$$

as $h\to 0$, where $K\in\mathbb{N}\cup\{0\}$ is the truncation parameter and $\partial_w\varphi(w)$ satisfies some regularity conditions, and the operator $\Delta^{\gamma}_{h,K}$ is given by

$$\Delta^{\gamma}_{h,K}f(w) = \frac{1}{h^{\gamma}}\sum_{k=-K}^{K} g_{\gamma,k}f(w-kh),$$

for any test function f satisfying some regularity conditions, where

$$g_{\gamma,k} := \frac{(-1)^k\Gamma(\gamma+1)}{\Gamma\left(\frac{\gamma}{2}-k+1\right)\Gamma\left(\frac{\gamma}{2}+k+1\right)}.$$

We study the numerical error when approximating the drift term $b(\mathbf{w},\alpha,\theta)$ with the skewness parameter $-1<\theta<1$ and provide the truncation error with a truncation parameter K in Corollary E.1. Based on this result, Theorem 4.2 follows, which quantifies the bias induced by α-stable noise on gradient updates using the Euler-Maruyama scheme.

Instead of using the centred difference method to implement the approximation for the $\theta=0$ case, we tackle the more general $\theta\neq 0$ case by using shifted Grünwald-Letnikov difference operators to approach the left and right fractional derivative respectively. Let us define the parameter $-1<\gamma:=\alpha-2<0$. Then, we can now formally define the Riesz-Feller type fractional derivative operator as follows:

$$\mathcal{D}^{\gamma,-\theta}f(w) := \frac{1}{2\cos(\gamma\pi/2)}\left[(1-\theta)\mathcal{I}_+^{-\gamma}f(w) + (1+\theta)\mathcal{I}_-^{-\gamma}f(w)\right]\,, \tag{E.1}$$

with

$$\mathcal{I}_{\pm}^{-\gamma}f(w) := \frac{1}{\Gamma(-\gamma)}\int_0^{\infty}\frac{f(w\pm\xi)}{\xi^{\gamma+1}}d\xi\,. \tag{E.2}$$

Before we proceed, we first introduce difference operators $\mathcal{A}^{\gamma}_{h,p}$ and $\mathcal{B}^{\gamma}_{h,q}$, where p and q are two non-negative integers chosen to be the shifted parameters,

$$\mathcal{A}^{\gamma}_{h,p}f(w) = \frac{1}{h^{\gamma}}\sum_{k=0}^{\infty}\tilde{g}_{\gamma,k}f(w-(k-p)h), \tag{E.3}$$

$$\mathcal{B}^{\gamma}_{h,q}f(w) = \frac{1}{h^{\gamma}}\sum_{k=0}^{\infty}\tilde{g}_{\gamma,k}f\left(w+(k-q)h\right). \tag{E.4}$$

Essentially, we defined a forward shifted difference operator $\mathcal{A}^{\gamma}_{h,p}$ to approximate the left fractional derivatives, and a backward shifted difference operator $\mathcal{B}^{\gamma}_{h,p}$ to approximate the right one. The coefficients $\tilde{g}_{\gamma,k} := \frac{(-1)^k\Gamma(-\gamma+k)}{\Gamma(k+1)\Gamma(-\gamma)}$ are from the

[3] Note that when $\theta=0$, $\mathcal{D}^{\alpha-2} = \mathcal{D}^{\alpha-2,0}$ recovers the Riesz potential.

coefficients of of the power series $(1-z)^\gamma$ with $-1<\gamma<0$ and $|z|\leq 1$. For any negative fractional number $-1<\gamma<0$ and $|z|\leq 1$, we have

$$(1-z)^\gamma=\sum_{k=0}^{\infty}(-1)^k\binom{-\gamma+k-1}{k}z^k, \tag{E.5}$$

where the binomial coefficient $\binom{-\gamma+k-1}{k}$ is well-defined and the binomial series converges for any complex number $|z|\leq 1$; see e.g. Kroneburg (2011). Indeed, when $-1<\gamma<0$, we get

$$\binom{-\gamma+k-1}{k}=\frac{\Gamma(-\gamma+k)}{\Gamma(k+1)\Gamma(-\gamma)}. \tag{E.6}$$

We first present the following first-order approximation result of the fractional derivative $\mathcal{D}^{\gamma,-\theta}$.

Theorem E.1. *Let $\mathcal{D}^{\gamma,-\theta}$ denote the fractional derivative for $-1<\gamma<0$ and $-1<\theta<1$ as in (E.1). Suppose the function $f\in L^1(\mathbb{R})\cap\mathcal{C}^4(\mathbb{R})$. Define*

$$\Delta_{h,p,q}^{\gamma,-\theta}f(w)=\frac{1}{2\cos(\gamma\pi/2)}\left[(1+\theta)\mathcal{A}_{h,p}^{\gamma}f(w)+(1-\theta)\mathcal{B}_{h,q}^{\gamma}f(w)\right]. \tag{E.7}$$

Then $\Delta_{h,p,q}^{\gamma,-\theta}f(w)$ is an approximation of $\mathcal{D}^{\gamma,-\theta}f(w)$ with the first-order accuracy:

$$\begin{aligned}&\left|\mathcal{D}^{\gamma,-\theta}f(w)-\Delta_{h,p,q}^{\gamma,-\theta}f(w)\right|\\&\quad\leq\left[|p-q|+|\theta|(p+q-\gamma)+\left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]\frac{C}{4\pi(|\gamma|+2)}h+\mathcal{O}\left(h^2\right),\end{aligned} \tag{E.8}$$

as $h\to 0$, uniformly for all $w\in\mathbb{R}$, where $C>0$ is a constant that may depend on f and $\mathcal{O}(\cdot)$ hides the dependence on p, q and γ.

Next, we provide an error bound for numerically computing the drift term $b(\mathbf{w},\alpha,\theta)$ by truncating the approximation series in Theorem E.1 as follows. Let us first define the operators $\mathcal{A}_{h,p,K}^{\gamma}$ and $\mathcal{B}_{h,q,K}^{\gamma}$:

$$\mathcal{A}_{h,p,K}^{\gamma}f(w):=\frac{1}{h^\gamma}\sum_{k=0}^{K}\tilde{g}_{\gamma,k}f\left(w-(k-p)h\right), \tag{E.9}$$

$$\mathcal{B}_{h,q,K}^{\gamma}f(w):=\frac{1}{h^\gamma}\sum_{k=0}^{K}\tilde{g}_{\gamma,k}f\left(w+(k-q)h\right), \tag{E.10}$$

with $\tilde{g}_{\gamma,k}:=\frac{(-1)^k\Gamma(-\gamma+k)}{\Gamma(k+1)\Gamma(-\gamma)}$, and $K\in\mathbb{N}\cup\{0\}$.

Before we state the next result, let us first introduce the following assumption.

Assumption E.1. *Suppose the function $\partial_w\varphi\in L^1(\mathbb{R})\cap\mathcal{C}^4(\mathbb{R})$. In addition, there exist constants $C_p,C_q>0$ satisfying*

$$|\partial_w\varphi(w-|k-p|h)|\leq C_pe^{-|k-p|h},\quad|\partial_w\varphi(w+|k-q|h)|\leq C_qe^{-|k-q|h}, \tag{E.11}$$

and $\min\{|k-p|,\,|k-q|\}>K$ for the constant $K\in\mathbb{N}\cup\{0\}$.

We have the following result.

Corollary E.1. *Suppose Assumption E.1 holds for $\partial_w\varphi$, and recall the truncated series $\mathcal{A}_{h,p,K}^{\gamma}$ and $\mathcal{B}_{h,q,K}^{\gamma}$ with $K\in\mathbb{N}\cup\{0\}$ defined in (E.9) and (E.10). Let us also define the operator:*

$$\Delta_{h,p,q,K}^{\gamma,-\theta}=\frac{1}{2\cos(\gamma\pi/2)}\left[(1-\theta)\mathcal{B}_{h,q,K}^{\gamma}+(1+\theta)\mathcal{A}_{h,p,K}^{\gamma}\right]. \tag{E.12}$$

Then the truncation error is bounded in first-order accuracy as follows,

$$\begin{aligned}&\left|\mathcal{D}^{\gamma,-\theta}\partial_w\varphi(w)-\Delta^{\gamma,-\theta}_{h,p,q,K}\partial_w\varphi(w)\right|\\&\quad\leq\frac{C}{4\pi(|\gamma|+2)}\left[|p-q|+|\theta|(p+q-\gamma)+\left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]h\\&\qquad\qquad+((1+\theta)C_p+(1-\theta)C_q)\frac{1}{hK}+\mathcal{O}(h^2),\end{aligned}\tag{E.13}$$

where $C, C_p, C_q > 0$ are constants that may depend on $\partial_w\varphi$ and $\mathcal{O}(\cdot)$ hides the dependence on p and q.

In particular, by taking $p = q = 0$, Corollary E.1 implies that

$$\begin{aligned}&\left|\mathcal{D}^{\gamma,-\theta}\partial_w\varphi(w)-\Delta^{\gamma,-\theta}_{h,p=0,q=0,K}\partial_w\varphi(w)\right|\\&\quad\leq\frac{C}{4\pi}\left(|\theta|+\left|\tan\left(\frac{\gamma\pi}{2}\right)\right|\right)h+((1+\theta)C_{p=0}+(1-\theta)C_{q=0})\frac{1}{hK}+\mathcal{O}(h^2),\end{aligned}\tag{E.14}$$

where $\Delta^{\gamma,-\theta}_{h,K}=\Delta^{\gamma,-\theta}_{h,p=0,q=0,K}$.

Corollary E.1 implies that one can approximate $\mathcal{D}^{\gamma,-\theta}$ by the truncated $\Delta^{\gamma,-\theta}_{h,K}$ instead of $\Delta^{\gamma,-\theta}_{h}$. Based on this result, we are able to quantify in Theorem 4.2 the bias induced when implementing Euler-Maruyama scheme to approximate the expectation of a test function g with respect to the target distribution π, where $\nu(g)=\int g(\mathbf{w})\pi(d\mathbf{w})$.

F. Metastability Analysis

In this section, we will focus on the metastability properties of the process

$$dw_t=-\nabla_w f(w_t)dt+\varepsilon dL^{\alpha,\theta}_t.\tag{F.1}$$

We will be interested in the first exit time, which is, roughly speaking, the expected time required for the process to exit a neighborhood of a local minimum. We will summarise the related theoretical results, which show that the first exit time behaviour of systems driven by asymmetric stable processes are similar to the ones of driven by symmetric stable processes. This informally implies that the process will quickly escape from narrow minima regions and will spend more time (in fact will get stuck) in wide minima regions. In this section, we make this argument rigorous.

For simplicity of the presentation, we consider the one-dimensional case where $L^{\alpha,\theta}_t$ is an asymmetric α-stable Lévy process with Lévy measure

$$\nu(dy)=\left(\frac{1-\theta}{2}c_\alpha 1_{y<0}+\frac{1+\theta}{2}c_\alpha 1_{y>0}\right)\frac{dy}{|y|^{1+\alpha}},\tag{F.2}$$

where $\theta\in(-1,1)$ and $\alpha\in(0,2)$ and $c_\alpha:=\frac{\alpha}{\Gamma(1-\alpha)\cos(\pi\alpha/2)}$. Then, the left and right tails of the Lévy measure are given by

$$\begin{aligned}H_-(-u)&:=\int_{(-\infty,-u)}\nu(dy)=\frac{1-\theta}{2}C_\alpha u^{-\alpha},\\H_+(u)&:=\int_{(u,+\infty)}\nu(dy)=\frac{1+\theta}{2}C_\alpha u^{-\alpha},\end{aligned}$$

where $C_\alpha:=\frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)}$, and

$$H(u):=H_-(-u)+H_+(u)=C_\alpha u^{-\alpha},\qquad\text{for any } u>0.$$

Let us assume that the function $w\mapsto f(w)$ satisfies the following conditions:

Assumption F.1. *(i) $f\in\mathcal{C}^1(\mathbb{R})\cap\mathcal{C}^3([-K,K])$ for some $K>0$;*

(ii) f has exactly n local minima m_i, $1 \leq i \leq n$ and $n-1$ local maxima s_i, $1 \leq i \leq n-1$, enumerated in increasing order with $s_0 = -\infty$ and $s_n = +\infty$:

$$-\infty < m_1 < s_1 < m_2 < \cdots < s_{n-1} < m_n < +\infty. \tag{F.3}$$

All extrema of f are non-degenerate, i.e. $\partial_w^2 f(m_i) > 0$, $1 \leq i \leq n$, and $\partial_w^2 f(s_i) < 0$, $1 \leq i \leq n-1$.

(iii) $|\partial_w f(w)| > c_1 |w|^{1+c_2}$ as $w \to \pm\infty$ for some $c_1, c_2 > 0$.

First, we consider the first exit time from a single well. For $\varepsilon > 0$ and $\gamma > 0$, define

$$\Omega_\varepsilon^i := \left[s_{i-1} + 2\varepsilon^\gamma, s_i - 2\varepsilon^\gamma\right], \tag{F.4}$$

with the convention that $\Omega_\varepsilon^1 := (-\infty, s_1 - 2\varepsilon^\gamma]$ and $\Omega_\varepsilon^n := [s_{n-1} + 2\varepsilon^\gamma, +\infty)$. The first exit time from the i-th well is defined as

$$\sigma^i(\varepsilon; \theta) := \inf\{t \geq 0 : w_t \notin [s_{i-1} + \varepsilon^\gamma, s_i - \varepsilon^\gamma]\}, \tag{F.5}$$

for $i = 1, \ldots, n$. Let us also define

$$\lambda^i(\varepsilon; \theta) := \frac{1-\theta}{2} C_\alpha \left|\frac{s_{i-1} - m_i}{\varepsilon}\right|^{-\alpha} + \frac{1+\theta}{2} C_\alpha \left|\frac{s_i - m_i}{\varepsilon}\right|^{-\alpha}, \tag{F.6}$$

for $i = 1, \ldots, n$. We have the following first exit time result from Imkeller & Pavlyukevich (2008).

Proposition F.1 (Proposition 3.1. in (Imkeller & Pavlyukevich, 2008))**.** *There exists $\gamma_0 > 0$ such that for any $0 < \gamma \leq \gamma_0$, $i = 1, 2, \ldots, n$,*

$$\lambda^i(\varepsilon; \theta)\sigma^i(\varepsilon; \theta) \to \exp(1), \qquad \textit{in distribution as } \varepsilon \to 0, \tag{F.7}$$

where $\exp(1)$ denotes the exponential distribution with mean 1, and

$$\lim_{\varepsilon \to 0} \mathbb{E}_w \left[\lambda^i(\varepsilon; \theta)\sigma^i(\varepsilon; \theta)\right] = 1, \tag{F.8}$$

where the limit holds uniformly over $w \in \Omega_\varepsilon^i$.

The above result implies that as $\varepsilon \to 0$,

$$\mathbb{E}_w \left[\sigma^i(\varepsilon; \theta)\right] \sim \left(\frac{1-\theta}{2} C_\alpha \left|s_{i-1} - m_i\right|^{-\alpha} + \frac{1+\theta}{2} C_\alpha \left|s_i - m_i\right|^{-\alpha}\right)^{-1} \varepsilon^{-\alpha}. \tag{F.9}$$

If $|s_i - m_i| > |s_{i-1} - m_i|$, i.e. the i-th well is asymmetric and the local minimum m_i is closer to the saddle point on the left s_{i-1} than the saddle point on the right s_i, then $\lambda^i(\varepsilon; \theta) < \lambda^i(\varepsilon; 0)$ and $\mathbb{E}_w[\sigma^i(\varepsilon; \theta)] > \mathbb{E}_w[\sigma^i(\varepsilon; 0)]$ for positive θ and $\lambda^i(\varepsilon; \theta) > \lambda^i(\varepsilon; 0)$ and $\mathbb{E}_w[\sigma^i(\varepsilon; \theta)] < \mathbb{E}_w[\sigma^i(\varepsilon; 0)]$ for negative θ. Similarly, if $|s_i - m_i| < |s_{i-1} - m_i|$, i.e. the i-th well is asymmetric and the local minimum m_i is closer to the saddle point on the right s_i than the saddle point on the left s_{i-1}, then $\lambda^i(\varepsilon; \theta) > \lambda^i(\varepsilon; 0)$ and $\mathbb{E}_w[\sigma^i(\varepsilon; \theta)] < \mathbb{E}_w[\sigma^i(\varepsilon; 0)]$ for positive θ and $\lambda^i(\varepsilon; \theta) < \lambda^i(\varepsilon; 0)$ and $\mathbb{E}_w[\sigma^i(\varepsilon; \theta)] > \mathbb{E}_w[\sigma^i(\varepsilon; 0)]$ for negative θ. The intuition is that when the well is asymmetric, the dynamics can exit the well faster when there is a skewness θ towards the the saddle point closer to the minimum of the well.

Next, we consider transitions between the wells. For any $0 < \Delta < \Delta_0 := \min_{1 \leq i \leq n}\{|m_i - s_{i-1}|, |m_i - s_i|\}$ and $w \in \mathbb{R}$ denote $B_\Delta(w) := \{v : |w - v| \leq \Delta\}$. Define

$$\tau^i(\varepsilon; \theta) := \inf\left\{t \geq 0 : w_t \in \cup_{k \neq i} B_\Delta(m_k)\right\}. \tag{F.10}$$

Then, we have the following result about transitions between the wells from Imkeller & Pavlyukevich (2008).

Proposition F.2 (Proposition 4.3. in (Imkeller & Pavlyukevich, 2008))**.** *For any $0 < \Delta < \Delta_0$ and $j \neq i$*

$$\lim_{\varepsilon \to 0} \mathbb{P}_w \left(w_{\tau^i(\varepsilon; \theta)} \in B_\Delta(m_j)\right) = \frac{q_{ij}}{q_i}, \tag{F.11}$$

uniformly for $w \in B_\Delta(m_i)$, $i = 1, \dots, n$, *where*

$$q_{ij} = \left(\frac{1-\theta}{2} 1_{j<i} + \frac{1+\theta}{2} 1_{j>i}\right) \cdot \left| |s_{j-1} - m_i|^{-\alpha} - |s_j - m_i|^{-\alpha} \right|, \quad i \neq j, \tag{F.12}$$

$$-q_{ii} = q_i = \sum_{j \neq i} q_{ij} = \frac{1-\theta}{2} |s_{i-1} - m_i|^{-\alpha} + \frac{1+\theta}{2} |s_i - m_i|^{-\alpha}. \tag{F.13}$$

From (F.11)-(F.13), we can compute that

$$\frac{q_{ij}}{q_i} = \begin{cases} \frac{\left||s_{j-1}-m_i|^{-\alpha}-|s_j-m_i|^{-\alpha}\right|}{\frac{1-\theta}{1+\theta}|s_{i-1}-m_i|^{-\alpha}+|s_i-m_i|^{-\alpha}} & \text{if } j > i, \\ \frac{\left||s_{j-1}-m_i|^{-\alpha}-|s_j-m_i|^{-\alpha}\right|}{|s_{i-1}-m_i|^{-\alpha}+\frac{1+\theta}{1-\theta}|s_i-m_i|^{-\alpha}} & \text{if } j < i. \end{cases} \tag{F.14}$$

Therefore, q_{ij}/q_i is increasing in θ for $j > i$ and decreasing in θ for $j < i$. This is consistent with the intuition that when $\theta > 0$, it is more likely for the dynamics to transit to a well on the right side, and when $\theta < 0$, it is more likely for the dynamics to transit to a well on the left side.

Next, we consider the following metastability result due to Theorem 1.1. in Imkeller & Pavlyukevich (2008). It describes the metastability phenomenon, which basically says that there exists a time scale under which the system behaves like a continuous time Markov process with a finite state space consisting of values in the set of stable attractors.

Theorem F.1 (Theorem 1.1. in (Imkeller & Pavlyukevich, 2008))**.** *If* $w_0 = w \in (s_{i-1}, s_i)$ *for some* $i = 1, 2, \dots, n$, *then for any* $t > 0$, *in the sense of finite-dimensional distributions,*

$$w_{t/H(1/\varepsilon)} \to Y_t(m_i), \qquad \text{as } \varepsilon \to 0, \tag{F.15}$$

where w_t *is defined in* (F.1) *and* $H(1/\varepsilon) = C_\alpha \varepsilon^\alpha$, *where* $Y_t(m_i)$ *that starts at* m_i *is a continuous-time Markov process on a finite states space* $\{m_1, \dots, m_n\}$ *with the infinitesimal generator* $Q = (q_{ij})_{i,j=1}^n$, *where* q_{ij} *is defined in* (F.12).

The Markov process $Y_t(m_i)$ admits a unique invariant distribution π satisfying $Q^T \pi = 0$. In the case of double well, i.e. $n = 2$ and $m_1 < s_1 = 0 < m_2$ separated by a local maximum at $s_1 = 0$, where without loss of generality we assume that $m_2 > |m_1|$, i.e. the second local minimum lies in a wider valley. A simple calculation yields that

$$q_{12} = \frac{1+\theta}{2} \frac{1}{|m_1|^\alpha} = -q_{11}, \quad \text{and} \quad q_{21} = \frac{1-\theta}{2} \frac{1}{m_2^\alpha} = -q_{22}, \tag{F.16}$$

so that it follows from $Q^T \pi = 0$ and $\pi_1 + \pi_2 = 1$ that

$$\pi_1 = \frac{(1+\theta)^{-1}|m_1|^\alpha}{(1+\theta)^{-1}|m_1|^\alpha + (1-\theta)^{-1} m_2^\alpha}, \tag{F.17}$$

$$\pi_2 = \frac{(1-\theta)^{-1} m_2^\alpha}{(1+\theta)^{-1}|m_1|^\alpha + (1-\theta)^{-1} m_2^\alpha}. \tag{F.18}$$

In particular, the ratio $\frac{\pi_2}{\pi_1} = \frac{1+\theta}{1-\theta}\left(\frac{m_2}{|m_1|}\right)^\alpha$ is increasing in $\frac{m_2}{|m_1|}$ and θ. That reveals that in the equilibrium the process will spend more time in the second valley if the second valley is wide and there is a drift towards to the right. In the symmetric case, i.e. $\theta = 0$, $\pi_2 > \pi_1$ since $m_2 > |m_1|$ so that in the equilibrium the process spends more time in the wider valley. In the asymmetric case, i.e. $\theta \neq 0$, if there is a strong skewness towards the left, i.e. $\theta < 0$ and $|\theta|$ is large, then in the equilibrium the process may spend more time in the narrower valley. Indeed $\pi_2 > \pi_1$ if and only if $\theta > \frac{|m_1|^\alpha - m_2^\alpha}{|m_1|^\alpha + m_2^\alpha}$.

G. Postponed Proofs

G.1. Proof of Lemma 3.1

Before we proceed to the proof of Lemma 3.1 we present some intermediary results that are required for the proof.

Definition G.1. *(Asymptotic order of magnitude) A positive sequence* a_m *is of the same order of magnitude as another positive sequence* b_m *(*$a_m \asymp b_m$*, i.e. 'asymptotically equivalent') if there exist some* $c, C > 0$ *such that:* $c \leq \frac{a_m}{b_m} \leq C$ *for any* $m \in \mathbb{N}$.

Lemma G.1 (Lemma A.1 in (Vladimirova et al., 2019))**.** *Let X be a normal random variable such that $X \sim \mathcal{N}(0, \sigma^2)$. Then the following asymptotic equivalence holds*

$$\|X\|_m \asymp \sqrt{m}.$$

We know that the centering of variables does not change their tail properties (Vershynin, 2018; Kuchibhotla & Chakrabortty, 2018). As such Lemma G.1 also applies to $X \sim \mathcal{N}(\mu, \sigma^2)$, as $\|X\| \asymp \|X - \mu\| \asymp \sqrt{m}$.

Lemma G.2 (Lemma 3.1 of Vladimirova et al. (2019))**.** *Let $\kappa : \mathbb{R} \rightarrow \mathbb{R}$ be a non-linearity that obeys the extended envelope property. And let X be a variable for which $\|X_+\|_m \asymp \|X_-\|_m$ where X_- and X_+ denote the left and right tail of the variable respectively* [4]. *Then we have:*

$$\|\kappa(X)\|_m \asymp \|X\|_m \, , \qquad \textit{for any } m \geq 1 \, . \tag{G.1}$$

Lemma G.3. *Let $X_1, \dots, X_N$ be variables that each obeys $\|X_i\|_m \lesssim m^r, p \in R, i = 1, \dots, N$ and $(W_i, \dots, W_N) \in \mathbb{R}^N$.*

$$\left\| \sum_{i=1}^{N} W_i X_i \right\|_m \lesssim m^r \, .$$

Proof of Lemma G.3. By Minkowski's inequality we have that

$$\begin{aligned} &\left\| \sum_{i=1}^{N} W_i X_i \right\|_m \leq \sum_{i=1}^{N} \|W_i X_i\|_m \leq \sum_{l=1}^{N} |W_i A_i| m^r, \; (A_1, \dots, A_N) \in \mathbb{R}^N \\ \Leftrightarrow &\left\| \sum_{i=1}^{N} W_i X_i \right\|_m \lesssim m^r \, . \end{aligned}$$

The A_i here are constants that upper bound the asymptotics of each norm $\|X_i\|_m$ in the sum. □

Proof of Lemma 3.1. **Additive Noise.** Consider first the noised data, $\widetilde{\mathbf{h}}_0(\mathbf{x}) = \mathbf{x} + \boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_0 \sim \mathcal{N}(0, \sigma_0^2)$. As Lemma G.1 shows, Gaussian random variables have an m^{th} norm that is asymptotically equivalent to $\sqrt{m}$,

$$\left\| \widetilde{h}_{0,l}(\mathbf{x}) \right\|_m \asymp \sqrt{m}, \qquad \text{for any } l = 1, \dots, n_0 \, ,$$

where n_0 is the dimensionality of data.

Let us now assume that $\left\| \widetilde{h}_{i,l}(\mathbf{x}) \right\|_m \lesssim \sqrt{m}$, for any $l = 1, \dots, n_i$, for some layer i. The pre-non-linearity at this layer is given by $\tilde{\mathbf{g}}_i = \mathbf{W}_{i+1} \widetilde{\mathbf{h}}_i$. The j^{th} element of $\tilde{\mathbf{g}}_i$ is defined as a sum,

$$g_{i,j}(\mathbf{x}) = \sum_{l=1}^{n_i} W_{i+1,l,j} \widetilde{h}_{i,l}(\mathbf{x}) \, ,$$

where $W_{i+1,l,j}$ is the weight that maps from the l^{th} neuron in layer i to the j^{th} in layer $i+1$. By Lemma G.3,

$$\|g_{i,j}(\mathbf{x})\|_m \lesssim \sqrt{m}, \qquad m = 1, \dots, n_i \, .$$

As such if we assume the non-linearities ϕ at each layer obey the extended envelope property, then we have by Lemma G.2:

$$\begin{aligned} \|\phi(g_{i,j}(\mathbf{x}))\|_m = \left\| \widehat{h}_{i+1,j}(\mathbf{x}) \right\|_m &\asymp \|\kappa(g_{i,j}(\mathbf{x}))\|_m \\ \Leftrightarrow \left\| \widehat{h}_{i+1,j}(\mathbf{x}) \right\|_m &\lesssim \sqrt{m}, \qquad j = 1, \dots, n_{i+1} \, . \end{aligned}$$

[4] We weaken Vladimirova et al. (2019)'s requirement for X to be symmetric as the proof they give still holds here.

Note that $\widetilde{\mathbf{h}}_{i+1} = \widehat{\mathbf{h}}_{i+1} + \boldsymbol{\epsilon}_{i+1}, \boldsymbol{\epsilon}_{i+1} \sim \mathcal{N}(0, \sigma^2_{i+1})$. By Lemma G.3, once again $\|\widetilde{h}_{i+1,j}(\mathbf{x})\|_m \lesssim \sqrt{m}$, is Gaussian in its tails. By recursion, with $\widetilde{\mathbf{h}}_0$ as the base case, we have that

$$\left\|\widetilde{h}_{i,l}(\mathbf{x})\right\|_m \lesssim \sqrt{m}, \qquad \text{for any } m \geq 1;\ i = 1, \dots, L-1;\ l = 1, \dots, n_i\,.$$

Multiplicative Noise. Consider first the noised data, $\widetilde{\mathbf{h}}_0(\mathbf{x}) = \mathbf{x} \circ \boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_0 \sim \mathcal{N}(1, \sigma^2_0)$. As Lemma G.1 shows, Gaussian random variables have an m^{th} norm that is asymptotically equivalent to $\sqrt{m}$,

$$\left\|\widetilde{h}_{0,l}(\mathbf{x})\right\|_m \asymp \sqrt{m}, \qquad \text{for any } l = 1, \dots, n_0\,,$$

where n_0 is the dimensionality of data.

Let us now assume that $\left\|\widetilde{h}_{i,l}(\mathbf{x})\right\|_m \lesssim m^r$, $\forall l = 1, \dots, n_i$, for some layer i. The pre-non-linearity at this layer is given by $\tilde{\mathbf{g}}_i = \mathbf{W}_{i+1}\widetilde{\mathbf{h}}_i$. The j^{th} element of $\tilde{\mathbf{g}}_i$ is defined as a sum,

$$g_{i,j}(\mathbf{x}) = \sum_{l=1}^{n_i} W_{i+1,l,j}\widetilde{h}_{i,l}(\mathbf{x})\,,$$

where $W_{i+1,l,j}$ is the weight that maps from the l^{th} neuron in layer i to the j^{th} in layer $i+1$. By Lemma G.3,

$$\|g_{i,j}(\mathbf{x})\|_m \lesssim m^r, \qquad m = 1, \dots, n_i\,.$$

As such if we assume the non-linearities ϕ at each layer obey the extended envelope property, then we have by Lemma G.2:

$$\begin{aligned}\|\phi(g_{i,j}(\mathbf{x}))\|_m = \left\|\widehat{h}_{i+1,j}(\mathbf{x})\right\|_m &\asymp \|\kappa(g_{i,j}(\mathbf{x}))\|_m \\ \Leftrightarrow \left\|\widehat{h}_{i+1,j}(\mathbf{x})\right\|_m &\lesssim m^r, \qquad j = 1, \dots, n_{i+1}\,.\end{aligned}$$

Note that $\widetilde{\mathbf{h}}_{i+1} = \widehat{\mathbf{h}}_{i+1} \circ \boldsymbol{\epsilon}_{i+1}, \boldsymbol{\epsilon}_{i+1} \sim \mathcal{N}(1, \sigma^2_{i+1})$. By Hölder's inequality we have that,

$$\left\|\widetilde{\mathbf{h}}_{i+1,j}(\mathbf{x})\right\|_m \lesssim m^{r+\frac{1}{2}}, \qquad j = 1, \dots, n_{i+1}\,.$$

By recursion, with $\widetilde{\mathbf{h}}_0$ as the base case, we have that

$$\left\|\widetilde{h}_{i,l}(\mathbf{x})\right\|_m \lesssim m^{\frac{i+1}{2}}, \qquad \text{for any } m \geq 1;\ i = 1, \dots, L-1;\ l = 1, \dots, n_i\,.$$

□

G.2. Proof of Theorem 3.1

Before we proceed to the proof of Theorem 3.1 we present some intermediary results that are required for the proof.

Lemma G.4. *Let X be a bounded random variable such that $|X| \leq C$, then X is sub-Weibull with parameter $\theta = 0$,*

$$\|X\|_m \lesssim m^0, \qquad \textit{for every } m \geq 1. \tag{G.2}$$

Proof of Lemma G.4. The moments of X obey

$$\mathbb{E}\left[|X|^m\right] \leq C^k\,.$$

Taking the root of this we find that $\|X\|_m := \mathbb{E}\left[|X|^m\right]^{\frac{1}{m}}$ is not dependent on m and scales as a constant, $\|X\|_m \asymp m^0$. □

Proof of Theorem 3.1. We first consider the gradient for a single data-label pair of $W_{i,l,j}$ the weight that maps from neuron l in layer $i-1$ to neuron j in layer i. We study this gradient using the chain rule, where we decompose $\partial E_{\mathcal{L}}(\cdot)/\partial W_{i,l,j}$ as

$$\frac{\partial E_{\mathcal{L}}(\cdot)}{\partial \widetilde{h}_{i,j}} \cdot \frac{\partial \widetilde{h}_{i,j}}{\partial W_{i,l,j}},$$

where $\widetilde{h}_{i,j}$ is the (noised) activation of the j^{th} neuron in the i^{th} layer. Thus, the gradient noise on the weights can be described as the product of two random variables.

Additive Noise. Let us first consider the properties of $\partial E_{\mathcal{L}}(\cdot)/\partial \widetilde{h}_{i,j}$ for the additive case.

Regression. In the case of regression we use a mean-square-error (MSE) we have that:

$$\Delta \mathcal{L}(\mathbf{x}, \mathbf{y}) = 2(\mathbf{y} - \mathbf{h}_L(\mathbf{x}))\mathcal{E}_L(\mathbf{x}) + (\mathcal{E}_L(\mathbf{x}))^2 \ ,$$

where we imply all terms' dependence on $\mathbf{w}, \boldsymbol{\epsilon}$ for brevity of notation. One can already see that the derivative of this object with respect to each element of $\widetilde{\mathbf{h}}_L$ will have tail properties that are asymptotically equivalent to those of $\mathcal{E}_L$, which we know by Lemma 3.1.

$$\left\| \frac{\partial \Delta \mathcal{L}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{L,j}} \right\|_m \lesssim \sqrt{m}, \qquad m = 1, \ldots, n_L \ .$$

If we center this distribution the tail properties of this variable are unchanged (Vershynin, 2018; Kuchibhotla & Chakrabortty, 2018). In particular the asymptotic behaviour of $\|\cdot\|_m$ is unchanged and we have that:

$$\left\| \frac{\partial \Delta \mathcal{L}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{L,j}} - \mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\partial \Delta \mathcal{L}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{L,j}} \right] \right\|_m \lesssim \sqrt{m},$$

$$\Leftrightarrow \left\| \frac{\partial E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{L,j}} \right\|_m \lesssim \sqrt{m}, \qquad m = 1, \ldots, n_L \ .$$

Classification. In the case of classification we use a cross-entropy (CE) error. There is no easy closed-form for $\Delta\mathcal{L}(\mathbf{x}, \mathbf{y})$ here, but we can infer the properties of $\nabla_{\widetilde{\mathbf{h}}_L} \Delta \mathcal{L}(\mathbf{x}, \mathbf{y})$ by studying the properties of the gradient $\nabla_{\mathbf{h}_L} \mathcal{L}(\mathbf{x}, \mathbf{y})$. For CE we know that:

$$\nabla_{\mathbf{h}_L} \mathcal{L}(\mathbf{x}, \mathbf{y}) = \text{sigmoid}(\mathbf{h}_L(\mathbf{x})) - \mathbf{y}$$

in the binary label case. In the multi-label classification case we typically use a $\mathrm{softmax}$, which is also a bounded function. We can already see that any noise $\mathcal{E}_L$ added to $\mathbf{h}_L(\mathbf{x})$ will induce a change in the gradient that is inherently bounded by the $\mathrm{sigmoid}$ non-linearity, meaning that $\Delta\mathcal{L}(\mathbf{x}, \mathbf{y})$ will be bounded. As such the centered variable $E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})$ will also be bounded *and* zero mean,. By Lemma G.4 any bounded and zero mean distribution will be sub-Weibull with parameter $\theta = 0$, and will thus *also* be sub-Gaussian

$$\left\| \frac{\partial E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{L,j}} \right\|_m \lesssim \sqrt{m}, \qquad m = 1, \ldots, n_L \ .$$

Synthesizing the regression and classification settings we can conclude that each constitutive element of $\nabla_{\widetilde{\mathbf{h}}_L} E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})$ will be sub-Gaussian and will have zero mean.

We can now turn to the partial derivatives of the form $\partial E_{\mathcal{L}}(\cdot)/\partial \widetilde{h}_i^m$. Assume $\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})$ is of order m^r, with,

$$\left\| \frac{\partial E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{i,j}} \right\|_m \lesssim m^r, \qquad m = 1, \ldots, n_i \ ,$$

which entails for gradients at the previous $(i-1)^{\text{th}}$ layer we have

$$\frac{\partial E_{\mathcal{L}}(\cdot)}{\partial \widetilde{h}_{i-1,l}} = \nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) \nabla_{\widetilde{h}_{i-1,l}} \widetilde{\mathbf{h}}_i(\mathbf{x}) \ ,$$

where

$$\nabla_{\widetilde{h}_{i-1,l}}\widetilde{\mathbf{h}}_i(\mathbf{x}) = \kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right) \circ \left(\mathbf{W}_{i,l}\right),$$

where $\circ$ denotes the element wise product and $\mathbf{W}_{i,l}$ is the l^{th} column of the weight matrix $\mathbf{W}_i$. By definition, activation functions that obey the extended envelope property will have gradients that are bounded in norm, by some constant d_2. As such, by Lemma G.4 κ' will be sub-Weibull with $r = 0$. By Hölder's inequality we have that

$$\left\|\left(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y})\right)_z \left(\kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right)\right)_j\right\|_m \leq \left\|\left(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y})\right)_j\right\|_{2m} \left\|\left(\kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right)\right)_j\right\|_{2m},$$

where j indexes over the elements of both Jacobians. By definition, we know that there exists $A > 0$ and $B > 0$ such that $\|(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y}))_z\|_{2m} \leq A(2m)^p$ and $\|(\kappa'(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})))_j\|_{2m} \leq B$. As such

$$\left\|\left(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y})\right)_z \left(\kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right)\right)_j\right\|_m \leq \left\|\left(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y})\right)_j\right\|_{2m} \left\|\left(\kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right)\right)_j\right\|_{2m} \leq AB2^p m^r.$$

Thus we know that the product of these two variables will be asymptotically upper-bounded by

$$\left\|\left(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y})\right)_j \left(\kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right)\right)_j\right\|_m \lesssim m^r.$$

We now need to take into account the weighted sum across rows and columns (i.e. over indices j) that occurs. By Lemma G.3 we know that

$$\left\|\sum_{j=1}^{n_i}\left(\nabla_{\widetilde{\mathbf{h}}_i} E_{\mathcal{L}}(\mathbf{x},\mathbf{y})\right)_j \left(\kappa'\left(\mathbf{W}_{i,l}\widetilde{h}_{i-1,l}(\mathbf{x})\right) \circ \left(\mathbf{W}_{i,l}\right)\right)_j\right\|_m \lesssim m^r,$$

$$\Leftrightarrow \left\|\frac{\partial E_{\mathcal{L}}(\cdot)}{\partial \widetilde{h}_{i-1,l}}\right\|_m \lesssim m^r, \qquad m = 1,\ldots,n_i.$$

By recursion, with $\widetilde{\mathbf{h}}_L$ as the base case, gradients at layer i bounded in norm by m^r induce gradients at layer $i-1$, also bounded in norm by m^r. By recursion with the L^{th} layer as the base case we can claim that,

$$\left\|\frac{\partial E_{\mathcal{L}}(\cdot)}{\partial \widetilde{h}_{i,j}}\right\|_m \lesssim \sqrt{m}.$$

We have now defined the first constitutive term of $\partial E_{\mathcal{L}}(\cdot)/\partial W_{i,l,j}$. Defining $\partial\widetilde{h}_{i,j}/\partial W_{i,l,j}$ is much simpler:

$$\frac{\partial \widetilde{h}_{i,j}}{\partial W_{i,l,j}} = \kappa'\left(W_{i,l,j}\widetilde{h}_{i-1,l}(\mathbf{x})\right)\left(\widetilde{h}_{i-1,l}(\mathbf{x})\right).$$

Here $\widetilde{h}_{i-1,l}(\mathbf{x})$, which we know is sub-Gaussian by Lemma 3.1, is once again multiplied to a bounded variable, κ'. Thus reapplying Hölder's inequality we obtain that

$$\left\|\frac{\partial \widetilde{h}_{i,j}}{\partial W_{i,l,j}}\right\|_m \lesssim \sqrt{m}.$$

We can now bring together the characterisations of the gradients that constitute $\partial E_{\mathcal{L}}(\cdot)/\partial W_{i,l,j}$. We can re-use Hölder's inequality to show that the product of these variables will be sub-exponential

$$\left\|\frac{\partial E_{\mathcal{L}}((\mathbf{x},\mathbf{y});\mathbf{w},\boldsymbol{\epsilon})}{\partial W_{i,l,j}}\right\|_m \lesssim m, \qquad \text{for every } m \geq 1.$$

Multiplicative Noise. In the case of multiplicative noise we know that by Lemma 3.1, the accumulated noise at layer L will be of the same order as that at layer $L-1$, because we are not multiplying noise to the final layer, thus

$$\left\| \frac{\partial E_{\mathcal{L}}(\mathbf{x}, \mathbf{y})}{\partial \widetilde{h}_{L,j}} \right\|_m \lesssim m^{\frac{L}{2}}, \qquad m = 1, \ldots, n_L .$$

Repeating the analysis done for the additive case we can claim that,

$$\left\| \frac{\partial E_{\mathcal{L}}(\cdot)}{\partial \widetilde{h}_i^m} \right\|_m \lesssim m^{\frac{L}{2}}.$$

We have now defined the first constitutive term of $\partial E_{\mathcal{L}}(\cdot)/\partial W_{i,l,j}$. We now define $\partial \widetilde{h}_{i,j}/\partial W_{i,l,j}$.

$$\frac{\partial \widetilde{h}_i^m}{\partial W_{i,l,j}} = \kappa' \left(W_{i,l,j} \widetilde{h}_{i-1,l}(\mathbf{x}) \right) \left(\widetilde{h}_{i-1,l}(\mathbf{x}) \right) .$$

Here $\widetilde{h}_{i-1,l}(\mathbf{x})$, which we know is sub-Weibull with a parameter $p = \frac{i}{2}$ by Lemma 3.1, is once again multiplied to a bounded variable, κ'. Thus reapplying Hölder's inequality we obtain that

$$\left\| \frac{\partial \widetilde{h}_i^m}{\partial W_{i,l,j}} \right\|_m \lesssim m^{\frac{i}{2}}.$$

We can re-use Hölder's inequality to show that the product of these variables will be sub-Weibull with $p = \frac{L+i}{2}$

$$\left\| \frac{\partial E_{\mathcal{L}}((\mathbf{x}, \mathbf{y}); \mathbf{w}, \boldsymbol{\epsilon})}{\partial W_{i,l,j}} \right\|_m \lesssim m^{\frac{L+i}{2}}, \qquad \text{for every } m \geq 1 .$$

Mean of Noise. Finally, by definition these gradients are zero mean,

$$\frac{\partial E_{\mathcal{L}}((\mathbf{x}, \mathbf{y}); \mathbf{w}, \boldsymbol{\epsilon})}{\partial W_{i,l,j}} = \frac{\partial \Delta \mathcal{L}(\mathbf{x}, \mathbf{y})}{\partial W_{i,l,j}} - \mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\partial \Delta \mathcal{L}(\mathbf{x}, \mathbf{y})}{\partial W_{i,l,j}} \right] .$$

□

G.3. Proof of Theorem 4.1

Before we proceed to the proof of Theorem 4.1, we present some technical results that will be used in the proof of Theorem 4.1 later.

For the d-dimensional asymmetric fractional Langevin dynamics $\mathbf{w}_t$, its infinitesimal generator is given in the following proposition.

Proposition G.1. *The asymmetric fractional Langevin dynamics $\mathbf{w}_t$ has the infinitesimal generator:*

$$\mathcal{L} f(\mathbf{w}) = \sum_{i=1}^{d} \left((b(\mathbf{w}, \alpha, \theta))_i - \varepsilon^\alpha \frac{\theta_i}{\cos(\alpha\pi/2)} \frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)} \right) \frac{\partial f(\mathbf{w})}{\partial w_i} + \varepsilon^\alpha \sum_{i=1}^{d} \mathcal{H}_{w_i}^{\alpha,\theta_i} f(\mathbf{w}), \tag{G.3}$$

where

$$\begin{aligned} \mathcal{H}_{w_i}^{\alpha,\theta_i} f(\mathbf{w}) := & \left(\frac{1}{2} + \frac{\theta_i}{2} \right) \frac{1}{\cos(\alpha\pi/2)} \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty \frac{f(\mathbf{w} + \xi \mathbf{e}_i) - f(\mathbf{w}) - \partial_{w_i} f(\mathbf{w}) \xi}{\xi^{\alpha+1}} d\xi \\ & + \left(\frac{1}{2} - \frac{\theta_i}{2} \right) \frac{1}{\cos(\alpha\pi/2)} \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty \frac{f(\mathbf{w} - \xi \mathbf{e}_i) - f(\mathbf{w}) + \partial_{w_i} f(\mathbf{w}) \xi}{\xi^{\alpha+1}} d\xi, \end{aligned} \tag{G.4}$$

where $\mathbf{e}_i$ is the i-th basis vector in $\mathbb{R}^d$, i.e. a d-dimensional unit vector with i-th coordinate being 1 and all the other coordinates being 0.

Proof of Proposition G.1. Since the asymmetric fractional Langevin dynamics is driven by the d-dimensional $\mathbf{L}_t^{\alpha,\theta}$, it suffices to show that the infinitesimal generator of d-dimensional $\mathbf{L}_t^{\alpha,\theta}$ is given by

$$\sum_{i=1}^{d}\mathcal{G}_{w_i}^{\alpha,\theta_i}f(\mathbf{w})=\sum_{i=1}^{d}\mathcal{H}_{w_i}^{\alpha,\theta_i}f(\mathbf{w})-\sum_{i=1}^{d}\frac{\theta_i}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}\frac{\partial}{\partial w_i}f(\mathbf{w}). \tag{G.5}$$

We start the proof by considering the dimension $d=1$ first. The one-dimensional α-stable Lévy motion with tail-index $1<\alpha<2$ and skewness $\theta\in(-1,1)$ has the infinitesimal generator given by:

$$\begin{aligned}\mathcal{G}^{\alpha,\theta}f(w):=&\frac{1+\theta}{2}\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_{z>0}[f(w+z)-f(w)-1_{|z|\leq 1}f'(w)z]\frac{dz}{|z|^{1+\alpha}}\\&+\frac{1-\theta}{2}\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_{z<0}[f(w+z)-f(w)-1_{|z|\leq 1}f'(w)z]\frac{dz}{|z|^{1+\alpha}}+af'(w),\end{aligned} \tag{G.6}$$

where $a\in\mathbb{R}$ is chosen so that $\mathcal{G}^{\alpha,\theta}w=0$ to be consistent with $\mu=0$ in $\mathbf{L}_t^{\alpha,\theta}-\mathbf{L}_s^{\alpha,\theta}\sim\mathcal{S}_\alpha((t-s)^{1/\alpha},\theta,\mu)$ for any $t>s$. Thus, we can compute that

$$\mathcal{G}^{\alpha,\theta}w=\frac{1+\theta}{2}\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_{z>1}\frac{dz}{z^\alpha}-\frac{1-\theta}{2}\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_{z>1}\frac{dz}{z^\alpha}+a=0, \tag{G.7}$$

which yields that

$$a=-\theta\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}. \tag{G.8}$$

Therefore, with $1<\alpha<2$,

$$\mathcal{G}^{\alpha,\theta}f(w)=\mathcal{H}^{\alpha,\theta}f(w)-\theta\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}f'(w), \tag{G.9}$$

where

$$\begin{aligned}\mathcal{H}^{\alpha,\theta}f(w):=&\left(\frac{1}{2}+\frac{\theta}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty\frac{f(w+\xi)-f(w)-f'(w)\xi}{\xi^{\alpha+1}}d\xi\\&+\left(\frac{1}{2}-\frac{\theta}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty\frac{f(w-\xi)-f(w)+f'(w)\xi}{\xi^{\alpha+1}}d\xi.\end{aligned} \tag{G.10}$$

Similarly, for the multi-dimensional case, we can show that the infinitesimal generator for $\mathbf{L}_t^{\alpha,\theta}$ is given by:

$$\sum_{i=1}^{d}\mathcal{G}_{w_i}^{\alpha,\theta_i}f(\mathbf{w})=\sum_{i=1}^{d}\mathcal{H}_{w_i}^{\alpha,\theta_i}f(\mathbf{w})-\sum_{i=1}^{d}\frac{\theta_i}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}\frac{\partial}{\partial w_i}f(\mathbf{w}), \tag{G.11}$$

where

$$\begin{aligned}\mathcal{H}_{w_i}^{\alpha,\theta_i}f(\mathbf{w}):=&\left(\frac{1}{2}+\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty\frac{f(\mathbf{w}+\xi\mathbf{e}_i)-f(\mathbf{w})-\partial_{w_i}f(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi\\&+\left(\frac{1}{2}-\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty\frac{f(\mathbf{w}-\xi\mathbf{e}_i)-f(\mathbf{w})+\partial_{w_i}f(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi,\end{aligned} \tag{G.12}$$

where $\mathbf{e}_i$ is the i-th basis vector in $\mathbb{R}^d$, i.e. a d-dimensional unit vector with i-th coordinate being 1 and all the other coordinates being 0. □

We recall that for any $\theta_i\in(-1,1)$, $1\leq i\leq d$, and $1<\alpha<2$, we have

$$(b(\mathbf{w},\alpha,\theta))_i=\frac{\varepsilon^\alpha}{\varphi(\mathbf{w})}\mathcal{D}_{w_i}^{\alpha-2,-\theta_i}(\partial_{w_i}\varphi(\mathbf{w})),\qquad\varphi(\mathbf{w})=e^{-\varepsilon^{-\alpha}f(\mathbf{w})}, \tag{G.13}$$

where

$$\mathcal{D}^{\alpha-2,-\theta_i}_{w_i}(\partial_{w_i}\varphi(\mathbf{w})) := \frac{-1}{2\cos(\alpha\pi/2)}\left[(1-\theta_i)\mathcal{I}^{2-\alpha}_{+,w_i}(\partial_{w_i}\varphi(\mathbf{w})) + (1+\theta_i)\mathcal{I}^{2-\alpha}_{-,w_i}(\partial_{w_i}\varphi(\mathbf{w}))\right], \tag{G.14}$$

and

$$\mathcal{I}^{2-\alpha}_{\pm,w_i}(\partial_{w_i}\varphi(\mathbf{w})) := \frac{1}{\Gamma(2-\alpha)}\int_0^\infty \frac{\partial_{w_i}\varphi(\mathbf{w}\pm\xi\mathbf{e}_i)}{\xi^{\alpha-1}}d\xi. \tag{G.15}$$

In the next result, we provide an alternative formula for $b(\mathbf{w},\alpha,\theta) = ((b(\mathbf{w},\alpha,\theta))_i, 1 \leq i \leq d)$ that is defined in (4.7).

Proposition G.2. *For any $\theta_i \in (-1,1)$, $1 \leq i \leq d$, and $1 < \alpha < 2$, we have*

$$\begin{aligned}(b(\mathbf{w},\alpha,\theta))_i := &\frac{\varepsilon^\alpha}{\varphi(\mathbf{w})}\left(\frac{1}{2}-\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty \frac{\int_{w_i}^{w_i+\xi}\varphi(\mathbf{w}+(y-w_i)\mathbf{e}_i)dy - \varphi(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi \\ &+ \frac{\varepsilon^\alpha}{\varphi(\mathbf{w})}\left(\frac{1}{2}+\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty \frac{\int_{w_i}^{w_i-\xi}\varphi(\mathbf{w}+(y-w_i)\mathbf{e}_i)dy + \varphi(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi \\ &+ \varepsilon^\alpha\theta_i\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)},\end{aligned} \tag{G.16}$$

with $\varphi(\mathbf{w}) := \exp(-\varepsilon^{-\alpha}f(\mathbf{w}))$.

Proof of Proposition G.2. Let us first consider the case when $\theta_i = 0$, $1 \leq i \leq d$. We have

$$\begin{aligned}\mathcal{D}^{\alpha-2}_{w_i}(\partial_{w_i}\varphi(\mathbf{w})) &:= -\mathcal{I}^{2-\alpha}_{w_i}(\partial_{w_i}\varphi(\mathbf{w})) \\ &= \frac{-1}{2\cos(\alpha\pi/2)}\left[\mathcal{I}^{2-\alpha}_{+,w_i}(\partial_{w_i}\varphi(\mathbf{w})) + \mathcal{I}^{2-\alpha}_{-,w_i}(\partial_{w_i}\varphi(\mathbf{w}))\right],\end{aligned} \tag{G.17}$$

where

$$\mathcal{I}^{2-\alpha}_{+,w_i}(\partial_{w_i}\varphi(\mathbf{w})) := \frac{1}{\Gamma(2-\alpha)}\int_0^\infty \frac{\partial_{w_i}\varphi(\mathbf{w}+\xi\mathbf{e}_i)}{\xi^{\alpha-1}}d\xi, \tag{G.18}$$

$$\mathcal{I}^{2-\alpha}_{-,w_i}(\partial_{w_i}\varphi(\mathbf{w})) := \frac{1}{\Gamma(2-\alpha)}\int_0^\infty \frac{\partial_{w_i}\varphi(\mathbf{w}-\xi\mathbf{e}_i)}{\xi^{\alpha-1}}d\xi. \tag{G.19}$$

Similarly, when $\theta_i \in (-1,1)$, $1 \leq i \leq d$, and $1 < \alpha < 2$, we have

$$(b(\mathbf{w},\alpha,\theta))_i = \frac{\varepsilon^\alpha}{\varphi(\mathbf{w})}\mathcal{D}^{\alpha-2,-\theta_i}_{w_i}(\partial_{w_i}\varphi(\mathbf{w})), \qquad \varphi(\mathbf{w}) = e^{-\varepsilon^{-\alpha}f(\mathbf{w})}, \tag{G.20}$$

where

$$\mathcal{D}^{\alpha-2,-\theta_i}_{w_i}(\partial_{w_i}\varphi(\mathbf{w})) := \frac{-1}{2\cos(\alpha\pi/2)}\left[(1-\theta_i)\mathcal{I}^{2-\alpha}_{+,w_i}(\partial_{w_i}\varphi(\mathbf{w})) + (1+\theta_i)\mathcal{I}^{2-\alpha}_{-,w_i}(\partial_{w_i}\varphi(\mathbf{w}))\right]. \tag{G.21}$$

□

Now, we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. We recall from Proposition G.1 that the asymmetric fractional Langevin dynamics $\mathbf{w}_t$ has the infinitesimal generator:

$$\mathcal{L}f(\mathbf{w}) = \sum_{i=1}^d\left((b(\mathbf{w},\alpha,\theta))_i - \varepsilon^\alpha\frac{\theta_i}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}\right)\frac{\partial f(\mathbf{w})}{\partial w_i} + \varepsilon^\alpha\sum_{i=1}^d\mathcal{H}^{\alpha,\theta_i}_{w_i}f(\mathbf{w}), \tag{G.22}$$

where $\mathcal{H}_{w_i}^{\alpha,\theta_i} f(\mathbf{w})$ is given in (G.4).

It follows that the adjoint operator $\mathcal{L}^*$ of $\mathcal{L}$ is given by:

$$\mathcal{L}^* f(\mathbf{w}) = -\sum_{i=1}^{d} \frac{\partial}{\partial w_i}\left(\left((b(\mathbf{w},\alpha,\theta))_i - \varepsilon^\alpha \frac{\theta_i}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}\right) f(\mathbf{w})\right) + \varepsilon^\alpha \sum_{i=1}^{d} \mathcal{H}_{w_i}^{\alpha,-\theta_i} f(\mathbf{w}) . \tag{G.23}$$

The probability density function $p(\mathbf{w},t)$ of the Lévy-driven SDE satisfies the Fokker-Planck equation (Schertzer et al., 2001):

$$\begin{aligned}\partial_t p(\mathbf{w},t) &= \mathcal{L}^* p(\mathbf{w},t) \\ &= -\sum_{i=1}^{d} \frac{\partial}{\partial w_i}\left[\left((b(\mathbf{w},\alpha,\theta))_i - \varepsilon^\alpha \theta_i \frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}\right) p(\mathbf{w},t)\right] + \varepsilon^\alpha \sum_{i=1}^{d} \mathcal{H}_{w_i}^{\alpha,-\theta_i} p(\mathbf{w},t).\end{aligned} \tag{G.24}$$

We can compute that

$$\begin{aligned}&\sum_{i=1}^{d} \frac{\partial}{\partial w_i}\left[\left((b(\mathbf{w},\alpha,\theta))_i - \varepsilon^\alpha \theta_i \frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{(\alpha-1)\Gamma(1-\alpha)}\right) \varphi(\mathbf{w})\right] + \varepsilon^\alpha \sum_{i=1}^{d} \mathcal{H}_{w_i}^{\alpha,-\theta_i} \varphi(\mathbf{w}) \\ &= -\varepsilon^\alpha \sum_{i=1}^{d} \frac{\partial}{\partial w_i}\left[\left(\frac{1}{2}-\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty \frac{\int_{w_i}^{w_i+\xi}\varphi(\mathbf{w}+(y-w_i)\mathbf{e}_i)dy - \varphi(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi\right] \\ &\quad - \varepsilon^\alpha \sum_{i=1}^{d} \frac{\partial}{\partial w_i}\left[\left(\frac{1}{2}+\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty \frac{\int_{w_i}^{w_i-\xi}\varphi(\mathbf{w}+(y-w_i)\mathbf{e}_i)dy + \varphi(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi\right] \\ &\quad + \varepsilon^\alpha \sum_{i=1}^{d} \mathcal{H}_{w_i}^{\alpha,-\theta_i} \varphi(\mathbf{w}) \\ &= -\varepsilon^\alpha \sum_{i=1}^{d} \left(\frac{1}{2}-\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty \frac{\varphi(\mathbf{w}+\xi\mathbf{e}_i)-\varphi(\mathbf{w})-\partial_{w_i}\varphi(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi \\ &\quad - \varepsilon^\alpha \sum_{i=1}^{d} \left(\frac{1}{2}+\frac{\theta_i}{2}\right)\frac{1}{\cos(\alpha\pi/2)}\frac{\alpha}{\Gamma(1-\alpha)}\int_0^\infty \frac{\varphi(\mathbf{w}-\xi\mathbf{e}_i)-\varphi(\mathbf{w})+\partial_{w_i}\varphi(\mathbf{w})\xi}{\xi^{\alpha+1}}d\xi \\ &\qquad + \varepsilon^\alpha \sum_{i=1}^{d} \mathcal{H}_{w_i}^{\alpha,-\theta_i} \varphi(\mathbf{w}) = 0.\end{aligned}$$

Hence, we conclude that $\pi(d\mathbf{w}) = \exp(-\varepsilon^{-\alpha} f(\mathbf{w}))d\mathbf{w} / \int_{\mathbb{R}^d} \exp(-\varepsilon^{-\alpha}(\mathbf{w}))d\mathbf{w}$ is an invariant distribution of the asymmetric fractional Langevin dynamics (4.6). Finally, if $b(\mathbf{w},\alpha,\theta)$ is Lipschitz continuous in $\mathbf{w}$, then $\pi(d\mathbf{w})$ is the unique invariant distribution of (4.6), see e.g. Schertzer et al. (2001). □

G.4. Proof of Theorem E.1

Theorem E.1 provides a first-order approximation of the fractional derivative $\mathcal{D}^{\gamma,-\theta}$ when $d = 1$.

Based on the work of Meerschaert & Tadjeran (2004), we will show a first-order approximation for the asymmetric fractional derivative $\mathcal{D}^{-\gamma,-\theta}$ when $-1 < \gamma < 0$ by using the shifted Grünwald-Letnikov difference operators defined in (E.3) and (E.4). Before we proceed to the proof of Theorem E.1, we will first present the Fourier transform property from equations (1) and (12) in Tian et al. (2015).

Property G.1 ((Tian et al., 2015))**.** *Let $-1 < \gamma < 0$ and $f \in L^1(\mathbb{R})$. The Fourier transform of $\mathcal{I}_-^{-\gamma} f$ and $\mathcal{I}_+^{-\gamma} f$ satisfy the following identities:*

$$\mathcal{F}\left[\mathcal{I}_-^{-\gamma} f(w)\right](\zeta) = (i\zeta)^\gamma \hat{f}(\zeta), \quad \mathcal{F}\left[\mathcal{I}_+^{-\gamma} f(w)\right](\zeta) = (-i\zeta)^\gamma \hat{f}(\zeta), \tag{G.25}$$

where $\hat{f}(\zeta)$ denotes the Fourier transform of f, such that $\hat{f}(\zeta) = \int_{-\infty}^{\infty} e^{-i\zeta w} f(w)dw$.

Now, we are ready to prove Theorem E.1.

Proof of Theorem E.1. The main idea for the proof of Theorem E.1 is to use the Fourier transform to estimate the difference between $\mathcal{F}\left[\left((1+\theta)\mathcal{A}^{\gamma}_{h,p}+(1-\theta)\mathcal{B}^{\gamma}_{h,q}\right)f(w)\right](\zeta)$ and $\mathcal{F}\left[\left((1+\theta)\mathcal{I}_{-}^{-\gamma}+(1-\theta)\mathcal{I}_{+}^{-\gamma}\right)f(w)\right](\zeta)$, and then apply the inverse Fourier transform to complete the proof. By the linearity of Fourier transforms, we can apply Fourier transform to (E.3) to obtain

$$\begin{aligned}
\mathcal{F}\left[\mathcal{A}^{\gamma}_{h,p}f(w)\right](\zeta) &= \frac{1}{h^{\gamma}}\sum_{k=0}^{\infty}(-1)^{k}\binom{-\gamma+k-1}{k}e^{-i\zeta(k-p)h}\hat{f}(\zeta)\\
&= \frac{1}{h^{\gamma}}e^{i\zeta ph}\left(1-e^{-i\zeta h}\right)^{\gamma}\hat{f}(\zeta)\\
&= (i\zeta)^{\gamma}\left(\frac{1-e^{-i\zeta h}}{i\zeta h}\right)^{\gamma}e^{i\zeta ph}\hat{f}(\zeta)\\
&= W_{p}(i\zeta h)(i\zeta)^{\gamma}\hat{f}(\zeta). \qquad \text{(G.26)}
\end{aligned}$$

Similarly, we can compute that

$$\begin{aligned}
\mathcal{F}\left[\mathcal{B}^{\gamma}_{h,q}f(w)\right](\zeta) &= \frac{1}{h^{\gamma}}\sum_{k=0}^{\infty}(-1)^{k}\binom{-\gamma+k-1}{k}e^{i\zeta(k-q)h}\hat{f}(\zeta)\\
&= (-i\zeta)^{\gamma}\left(\frac{1-e^{i\zeta h}}{-i\zeta h}\right)^{\gamma}e^{-i\zeta qh}\hat{f}(\zeta)\\
&= W_{-q}(-i\zeta h)(-i\zeta)^{\gamma}\hat{f}(\zeta). \qquad \text{(G.27)}
\end{aligned}$$

In addition, since $W_p(z)$ and $W_{-q}(-z)$ are analytic for any complex number $|z|\leq 1$, there exist series expansions so that by the first-order Taylor expansion we have

$$\begin{aligned}
W_{p}(z) &:= \left(\frac{1-e^{-z}}{z}\right)^{\gamma}e^{pz} = 1+\left(p-\frac{\gamma}{2}\right)z+\mathcal{O}\left(|z|^{2}\right),\\
W_{-q}(-z) &:= \left(\frac{1-e^{z}}{-z}\right)^{\gamma}e^{-qz} = 1-\left(q-\frac{\gamma}{2}\right)z+\mathcal{O}\left(|z|^{2}\right). \qquad \text{(G.28)}
\end{aligned}$$

Next, define a function $\hat{\psi}(h,\zeta)$ as the difference between $\mathcal{F}\left[\left((1+\theta)\mathcal{A}^{\gamma}_{h,p}+(1-\theta)\,\mathcal{B}^{\gamma}_{h,q}\right)f(w)\right](\zeta)$ and $\mathcal{F}\left[\left((1+\theta)\mathcal{I}_{-}^{-\gamma}+(1-\theta)\,\mathcal{I}_{+}^{-\gamma}\right)f(w)\right](\zeta)$. By the linearity of Fourier transform, we have

$$\begin{aligned}
\hat{\psi}(h,\zeta) &= (1+\theta)\left(\mathcal{F}\left[\mathcal{A}^{\gamma}_{h,p}f(w)\right](\zeta)-\mathcal{F}\left[\mathcal{I}_{-}^{-\gamma}f(w)\right](\zeta)\right)+(1-\theta)\left(\mathcal{F}\left[\mathcal{B}^{\gamma}_{h,q}f(w)\right](\zeta)-\mathcal{F}\left[\mathcal{I}_{+}^{-\gamma}f(w)\right](\zeta)\right)\\
&= (1+\theta)(i\zeta)^{\gamma}\hat{f}(\zeta)\left(W_{p}(i\zeta h)-1\right)+(1-\theta)(-i\zeta)^{\gamma}\hat{f}(\zeta)\left(W_{-q}(-i\zeta h)-1\right)\\
&= (1+\theta)(i\zeta)^{\gamma}\hat{f}(\zeta)\left(p-\frac{\gamma}{2}\right)(i\zeta h)-(1-\theta)(-i\zeta)^{\gamma}\hat{f}(\zeta)\left(q-\frac{\gamma}{2}\right)(i\zeta h)\\
&= (1+\theta)(i\zeta)^{\gamma+1}\hat{f}(\zeta)\left(p-\frac{\gamma}{2}\right)h+(1-\theta)(-i\zeta)^{\gamma+1}\hat{f}(\zeta)\left(q-\frac{\gamma}{2}\right)h\\
&\overset{\text{(a)}}{=} \left[(1+\theta)e^{i\pi\gamma/2}\left(p-\frac{\gamma}{2}\right)h-(1-\theta)e^{-i\pi\gamma/2}\left(q-\frac{\gamma}{2}\right)\right]|\zeta|^{1+\gamma}\hat{f}(\zeta)h\\
&\overset{\text{(b)}}{=} \left[\cos\left(\frac{\gamma\pi}{2}\right)\left((p-q)+\theta(p+q-\gamma)\right)+\sin\left(\frac{\gamma\pi}{2}\right)\left((p+q-\gamma)+\theta(p-q)\right)i\right]|\zeta|^{1+\gamma}\hat{f}(\zeta)h, \qquad \text{(G.29)}
\end{aligned}$$

where we used the fact that for any real x, $0<1+\gamma<1$, we have $ix=|x|e^{i\ \mathrm{sign}(x)\pi/2}$ so that $(ix)^{1+\gamma}=|x|^{1+\gamma}e^{i\ \mathrm{sign}(x)\pi(\gamma+1)/2}=\mathrm{sign}(x)|x|^{1+\gamma}e^{i\ \mathrm{sign}(x)\pi\gamma/2}$ which implies equality (a), and we applied Euler's formula with $-1<\gamma<0$ to get equality (b). By our assumption $f\in\mathcal{C}^{4}(\mathbb{R})$, we have

$$|\hat{f}(\zeta)|\leq C(1+|\zeta|)^{-4},$$

for a constant $C > 0$ that may depend on f. Hence, by taking a sufficiently small h, we obtain,

$$\begin{aligned}|\hat{\psi}(h,\zeta)| &\leq \left|\left[\cos^2\left(\frac{\gamma\pi}{2}\right)(p-q+\theta(p+q-\gamma))^2 + \sin^2\left(\frac{\gamma\pi}{2}\right)(p+q-\gamma+\theta(p-q))^2\right]\right|^{\frac{1}{2}} C(1+|\zeta|)^{\gamma-3}h + c_0h^2 \\ &\leq \left[\cos\left(\frac{\gamma\pi}{2}\right)|p-q+\theta(p+q-\gamma)| + \left|\sin\left(\frac{\gamma\pi}{2}\right)\right||p+q-\gamma+\theta(p-q)|\right] C(1+|\zeta|)^{\gamma-3}h + c_0h^2,\end{aligned}$$

where we also used the inequality that $|\zeta|^{\gamma+1} \leq (1+|\zeta|)^{\gamma+1}$ for $-1 < \gamma < 0$, and $c_0 > 0$ is a constant may depend on p, q. When $f \in L^1(\mathbb{R})$, the inverse Fourier transform exists with $-1 < \gamma < 0$, i.e. $\psi(h,w) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} e^{-i\zeta w}\hat{\psi}(h,\zeta)d\zeta$, and it follows that

$$\begin{aligned}|\psi(h,w)| &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\hat{\psi}(h,\zeta)e^{-i\zeta w}d\zeta \\ &\leq \frac{1}{2\pi}\int_{-\infty}^{\infty}|\hat{\psi}(h,\zeta)|d\zeta \\ &\leq \left[\cos\left(\frac{\gamma\pi}{2}\right)|p-q+\theta(p+q-\gamma)| + \left|\sin\left(\frac{\gamma\pi}{2}\right)\right||p+q-\gamma+\theta(p-q)|\right]\frac{C}{2\pi(|\gamma|+2)}h + \mathcal{O}\left(h^2\right),\end{aligned} \tag{G.30}$$

where $\mathcal{O}(\cdot)$ hides the dependence on p, q and γ, and $C > 0$ is a constant that may depend on $f \in L^1(\mathbb{R}) \cap \mathcal{C}^4(\mathbb{R})$.

Hence, we conclude that

$$\begin{aligned}&\left|\mathcal{D}^{\gamma,-\theta}f(w) - \Delta^{\gamma,-\theta}_{h,p,q}f(w)\right| \\ &= \frac{1}{2\cos(\pi\gamma/2)}\left|(1+\theta)\left(\mathcal{A}^{\gamma}_{h,p}f(w) - \mathcal{I}^{\gamma}_{-}f(w)\right) + (1-\theta)\left(\mathcal{B}^{\gamma}_{h,q}f(w) - \mathcal{I}^{\gamma}_{+}f(w)\right)\right| \\ &\leq \left[|p-q| + |\theta|(p+q-\gamma) + \left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]\frac{C}{4\pi(|\gamma|+2)}h + \mathcal{O}\left(h^2\right).\end{aligned} \tag{G.31}$$

The proof is complete. □

G.5. Proof of Corollary E.1

With the definitions of the truncated series $\mathcal{A}^{\gamma}_{h,p,K}$ defined in (E.9) and $\mathcal{B}^{\gamma}_{h,q,K}$ in (E.10), we are now ready to prove Corollary E.1.

Proof of Corollary E.1. We will first control the difference $\left|\Delta^{\gamma,-\theta}_{h,p,q}\partial_w\varphi(w) - \Delta^{\gamma,-\theta}_{h,p,q,K}\partial_w\varphi(w)\right|$. Then the triangular inequality can be applied with the fractional derivative approximation error bound in Theorem E.1 to get the numerical truncation error.

By using the definitions of $\mathcal{A}_{h,p}, \mathcal{B}_{h,q}$ and $\mathcal{A}_{h,p,K}, \mathcal{B}_{h,q,K}$, under the Assumption E.1, there exist two universal constants $C_p > 0$ and $C_q > 0$ so that

$$\begin{aligned}&\left|\Delta^{\gamma,-\theta}_{h,p,q}\partial_w\varphi(w) - \Delta^{\gamma,-\theta}_{h,p,q,K}\partial_w\varphi(w)\right| \\ &= \frac{1}{2|\cos(\pi\gamma/2)|}\Big|(1+\theta)\left(\mathcal{A}^{\gamma}_{h,p}\partial_w\varphi(w-(k-p)h) - \mathcal{A}^{\gamma}_{h,p,K}\partial_w\varphi(w-(k-p)h)\right) \\ &\qquad\qquad + (1-\theta)\left(\mathcal{B}^{\gamma}_{h,q}\partial_w\varphi(w+(k-q)h) - \mathcal{B}^{\gamma}_{h,q,K}\partial_w\varphi(w+(k-q)h)\right)\Big| \\ &\leq \frac{1}{2|\cos(\pi\gamma/2)|}\frac{1}{\Gamma(-\gamma)}\frac{1}{h^{\gamma}}\left(\sum_{k=K+p+1}^{\infty}\frac{\Gamma(-\gamma+k)}{\Gamma(k+1)}(1+\theta)|\partial_w\varphi(w-(k-p)h)| \right. \\ &\qquad\qquad \left. + \sum_{k=K+q+1}^{\infty}\frac{\Gamma(-\gamma+k)}{\Gamma(k+1)}(1-\theta)|\partial_w\varphi(w+(k-q)h)|\right) \\ &\leq \frac{(1+\theta)C_p}{h^{\gamma}}\sum_{k=K+p+1}^{\infty}\frac{\Gamma(-\gamma+k)}{\Gamma(k+1)}e^{-(k-p)h} + \frac{(1-\theta)C_q}{h^{\gamma}}\sum_{k=K+q+1}^{\infty}\frac{\Gamma(-\gamma+k)}{\Gamma(k+1)}e^{-(k-q)h}.\end{aligned} \tag{G.32}$$

Next, by applying Stirling's formula, we have as $k \to \infty$:

$$\begin{aligned}\frac{\Gamma(-\gamma+k)}{\Gamma(k+1)} &\sim \frac{\sqrt{2\pi(k-1-\gamma)}\,(k-1-\gamma)^{k-1-\gamma}\,e^{-(k-1-\gamma)}}{\sqrt{2\pi k}\,k^k\,e^{-k}}\\ &= \frac{(k-1-\gamma)^{\,k-1/2-\gamma}}{k^{\,k+1/2}}e^{1+\gamma}\\ &= k^{-\gamma-1}\left(1-\frac{1+\gamma}{k}\right)^k\left(\frac{k-1-\gamma}{k}\right)^{-1/2-\gamma}e^{1+\gamma}\\ &\sim k^{-\gamma-1}.\end{aligned}$$

Therefore, it follows from (G.32) that

$$\begin{aligned}&\left|\Delta^{\gamma,-\theta}_{h,p,q}\partial_w\varphi(w)-\Delta^{\gamma,-\theta}_{h,p,q,K}\partial_w\varphi(w)\right|\\ &\le (1+\theta)C_p h\sum_{k=K+p+1}^{\infty}(hk)^{-\gamma-1}e^{-(k-p)h}+(1-\theta)C_q h\sum_{k=K+q+1}^{\infty}(hk)^{-\gamma-1}e^{-(k-p)h}\\ &\le \left((1+\theta)C_p+(1-\theta)C_q\right)\frac{1}{hK},\end{aligned}\tag{G.33}$$

where we abused the notation such that C_p, C_q in (G.33) may differ from C_p, C_q in (G.32). Finally, the triangular inequality yields that

$$\begin{aligned}&\left|\mathcal{D}^{\gamma,-\theta}\partial_w\varphi(w)-\Delta^{\gamma,-\theta}_{h,p,q,K}\partial_w\varphi(w)\right|\\ &\quad\le\left|\mathcal{D}^{\gamma,-\theta}\partial_w\varphi(w)-\Delta^{\gamma,-\theta}_{h,p,q}\partial_w\varphi(w)\right|+\left|\Delta^{\gamma,-\theta}_{h,p,q}\partial_w\varphi(w)-\Delta^{\gamma,-\theta}_{h,p,q,K}\partial_w\varphi(w)\right|\\ &\quad\le\left[|p-q|+|\theta|(p+q-\gamma)+\left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]\frac{C}{4\pi(|\gamma|+2)}h\\ &\qquad\qquad+\left((1+\theta)C_p+(1-\theta)C_q\right)\frac{1}{hK}+\mathcal{O}\left(h^2\right),\end{aligned}$$

where C_p and C_q are two universal constants following Assumption E.1. The proof is completed. □

G.6. Proof of Theorem 4.2

First, let us recall that $\tilde{\nu}_N(g)=\frac{1}{H_N}\sum_{k=1}^{N}\eta_k g(\mathbf{w}_k)$ is the sample average, where $\mathbf{w}_k$ satisfies the Euler-Maruyama discretisation with the approximated drift $b_{h,K}$:

$$\tilde{\mathbf{w}}_{n+1}=\tilde{\mathbf{w}}_n+\eta_{n+1}b_{h,K}(\tilde{\mathbf{w}}_n,\alpha,\theta)+\varepsilon\eta_{n+1}^{1/\alpha}\Delta\mathbf{L}^{\alpha,\theta}_{n+1},\tag{G.34}$$

The corresponding SDE of (G.34) is given as

$$d\tilde{\mathbf{w}}_t=b_{h,K}(\tilde{\mathbf{w}}_{t-},\alpha,\theta)dt+\varepsilon d\mathbf{L}^{\alpha,\theta}_t,\tag{G.35}$$

and we define $\tilde{\nu}(g)=\int g(\mathbf{w})\tilde{\pi}(d\mathbf{w})$, where $\tilde{\pi}$ is the stationary distribution of (G.35).

Next, let us introduce the following assumption that is needed for Theorem 4.2.

Assumption G.1. *(i) Assume that the step sizes are decreasing and the sum diverges such that* $\lim_{n\to\infty}\eta_n=0\,,\lim_{N\to\infty}H_N=\infty\,.$

(ii) Let $V:\mathbb{R}\to\mathbb{R}_+^*$ be a function in $\mathcal{C}^2$, if $\lim_{|x|\to\infty}V(w)=\infty$, $|\partial_w V|\le C\sqrt{V}$ with some constant $C>0$ and $\partial_x^2 V$ is bounded. Then there exists $a\in(0,1]$, $\delta>0$ and $\beta\in\mathbb{R}$, such that $|b|^2\le CV^a$ and $b(\partial_w V)\le\beta-\delta V^a$ with b defined in (4.7). And the statement also holds for $\tilde{b}$.

(iii) The SDEs defined in (4.6) and (G.35) are geometrically ergodic with their unique invariant measures.

Before we proceed to the proof of Theorem 4.2, let us state a technical lemma bounding the error of $|\mathbb{E}[g(\mathbf{w}_t)]-\mathbb{E}[g(\tilde{\mathbf{w}}_t)]|$, where $(\mathbf{w}_t)_{t\ge0}$ and $(\tilde{\mathbf{w}}_t)_{t\ge0}$ follow SDEs in (4.6) and (G.35).

Lemma G.5. *Let $(\mathbf{w}_t)_{t\geq 0}$ and $(\tilde{\mathbf{w}}_t)_{t\geq 0}$ follow SDEs in (4.6) and (G.35) and g be a given test function with bounded $|\partial_w g|$. Suppose $K \in \mathbb{N} \cup \{0\}$ is a constant satisfying Assumption E.1 with respect to $\partial_x \varphi$ and Assumption G.1 holds, then the following bound holds:*

$$|\mathbb{E}\left[g\left(\mathbf{w}_t\right)\right] - \mathbb{E}\left[g\left(\tilde{\mathbf{w}}_t\right)\right]| \leq \frac{\tilde{C}}{4\pi(|\gamma|+2)}\left[|p-q| + |\theta|(p+q-\gamma) + \left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]h \\ + \left((1+\theta)C_p' + (1-\theta)C_q'\right)\frac{1}{hK} + \mathcal{O}\left(h^2\right), \tag{G.36}$$

where the constants $\tilde{C}, C_p', C_q' > 0$ may depend on the function $\partial_w\varphi$ and the bound for $|\partial_w g|$.

Proof. The proof is inspired by the proof of Lemma 3 in Şimşekli (2017). Let $\{P_t^{\mathbf{w}}\}_{t\geq 0}$ and $\{P_t^{\tilde{\mathbf{w}}}\}_{t\geq 0}$ be the corresponding Markov semigroups, i.e. $P_t^{\mathbf{w}} g(w) = \mathbb{E}_w[g(\mathbf{w}_t)]$, $P_t^{\tilde{\mathbf{w}}} g(w) = \mathbb{E}_w[g(\tilde{\mathbf{w}}_t)]$. Using the Markov semigroup property, following Lemma 3 in Şimşekli (2017), we have

$$|\mathbb{E}\left[g\left(\mathbf{w}_t\right)\right] - \mathbb{E}[g(\tilde{\mathbf{w}}_t)]| = \left|\int_0^t P_s^{\mathbf{w}}\left(\mathcal{L}^{\mathbf{w}} - \mathcal{L}^{\tilde{\mathbf{w}}}\right)P_{t-s}^{\tilde{\mathbf{w}}} g(w)ds\right|,$$

where $\mathcal{L}^{\mathbf{w}}$ and $\mathcal{L}^{\tilde{\mathbf{w}}}$ are the linear generators of $P_t^{\mathbf{w}}$ and $P_t^{\tilde{\mathbf{w}}}$, such that, for $g \in L^2(\pi), \partial_t P_t g = \mathcal{L} P_t g = P_t \mathcal{L} g$. The infinitesimal generators $\mathcal{L}^{\mathbf{w}}$ and $\mathcal{L}^{\tilde{\mathbf{w}}}$ are computed in (G.22). By the interchangeability of integration and differentiation, we have

$$\left|\int_0^t P_s^{\mathbf{w}}\left(\mathcal{L}^{\mathbf{w}} - \mathcal{L}^{\tilde{\mathbf{w}}}\right)P_{t-s}^{\tilde{\mathbf{w}}} g(w)ds\right| = \left|\int_0^t P_s^{\mathbf{w}}\left(b(w,\alpha,\theta) - b_{h,K}(w,\alpha,\theta)\right)P_{t-s}^{\tilde{\mathbf{w}}}\partial_w g(w)ds\right|.$$

By the ergodicity assumptions, for a bounded function f, there exist some constants $c > 0$ and $\lambda_w, \lambda_{\tilde{w}} > 0$ so that

$$|P_s^{\mathbf{w}} f| \leq c\, e^{-\lambda_w s}\|f\|_\infty, \qquad |P_{t-s}^{\tilde{\mathbf{w}}} f| \leq c\, e^{-\lambda_{\tilde{w}}(t-s)}\|f\|_\infty. \tag{G.37}$$

Using the boundedness assumption for $|\partial_w g|$ and Corollary E.1, and the fact that $\int_0^t e^{-\lambda_w s}ds \leq \frac{1}{\lambda_w}$, we have

$$|\mathbb{E}[g(\mathbf{w}_t)] - \mathbb{E}[g(\tilde{\mathbf{w}}_t)]| \leq \frac{\tilde{C}}{4\pi(|\gamma|+2)}\left[|p-q| + |\theta|(p+q-\gamma) + \left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]h \\ + \left((1+\theta)C_p' + (1-\theta)C_q'\right)\frac{1}{hK} + \mathcal{O}\left(h^2\right), \tag{G.38}$$

where the constants $\tilde{C} = \frac{c}{\lambda_w}C$, $C_p' = \frac{c}{\lambda_w}C_p$ and $C_q' = \frac{c}{\lambda_w}C_q$ may depend on $\partial_w\varphi$ and the bound for $|\partial_w g|$. □

Now we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. With the ergodicity assumptions, we have

$$\left|\nu(g) - \lim_{N\to\infty}\tilde{\nu}_N(g)\right| = \left|\nu(g) - \tilde{\nu}(g) + \tilde{\nu}(g) - \lim_{N\to\infty}\tilde{\nu}_N(g)\right| \leq \lim_{t\to\infty}|\mathbb{E}\left[g\left(\mathbf{w}_t\right)\right] - \mathbb{E}\left[g\left(\tilde{\mathbf{w}}_t\right)\right]| + \left|\tilde{\nu}(g) - \lim_{N\to\infty}\tilde{\nu}_N(g)\right|. \tag{G.39}$$

By Assumption G.1(ii), (Panloup, 2008) and similar arguments as in (Şimşekli, 2017), we get,

$$\left|\tilde{\nu}(g) - \lim_{N\to\infty}\tilde{\nu}_N(g)\right| = 0, \quad \text{a.s.}$$

By applying Lemma G.5 as $t \to \infty$, we obtain:

$$\left|\nu(g) - \lim_{N\to\infty}\tilde{\nu}_N(g)\right| \leq \lim_{t\to\infty}|\mathbb{E}[g(\mathbf{w}_t)] - \mathbb{E}[g(\tilde{\mathbf{w}}_t)]| \\ \leq \frac{\tilde{C}}{4\pi(|\gamma|+2)}\left[|p-q| + |\theta|(p+q-\gamma) + \left|\tan\left(\frac{\gamma\pi}{2}\right)\right|(p+q-\gamma+|\theta||p-q|)\right]h \\ + \left((1+\theta)C_p' + (1-\theta)C_q'\right)\frac{1}{hK} + \mathcal{O}\left(h^2\right), \tag{G.40}$$

where $\tilde{C}$, C_p', $C_q' > 0$ are constants that may depend on $\partial_w\varphi$ and the bound of $|\partial_w g|$. Finally, by taking $p = q = 0$, we complete the proof. □