Fractional Underdamped Langevin Dynamics: Retargeting SGD with Momentum under Heavy-Tailed Gradient Noise

SUPPLEMENTARY DOCUMENT

Umut Şimşekli [*1 2] **Lingjiong Zhu** [*3] **Yee Whye Teh** [2] **Mert Gurbuzbalaban** [4]

S1. Proof of Theorem 1

Proof. Let $q(\mathbf{x}, \mathbf{v}, t)$ denote the probability density of $(\mathbf{x}_t, \mathbf{v}_t)$. Then it satisfies the fractional Fokker-Planck equation (see Proposition 1 and Section 7 in (Schertzer et al., 2001)):

$$\partial_t q(\mathbf{x}, \mathbf{v}, t) = \gamma \sum_{i=1}^{d} \frac{\partial[(c(\mathbf{v}, \alpha))_i q(\mathbf{x}, \mathbf{v}, t)]}{\partial v_i} + \sum_{i=1}^{d} \frac{\partial[\partial_{x_i} f(\mathbf{x}) q(\mathbf{x}, \mathbf{v}, t)]}{\partial v_i} - \frac{\gamma}{\beta} \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} q(\mathbf{x}, \mathbf{v}, t) - \sum_{i=1}^{d} \frac{\partial[(\partial_{v_i} g(\mathbf{v}) q(\mathbf{x}, \mathbf{v}, t)]}{\partial x_i}.$$

We can compute that

$$\gamma \sum_{i=1}^{d} \frac{\partial[(c(\mathbf{v}, \alpha))_i \phi(\mathbf{x}) \psi(\mathbf{v})]}{\partial v_i} + \sum_{i=1}^{d} \frac{\partial[\partial_{x_i} f(\mathbf{x}) \phi(\mathbf{x}) \psi(\mathbf{v})]}{\partial v_i} - \frac{\gamma}{\beta} \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \phi(\mathbf{x}) \psi(\mathbf{v}) - \sum_{i=1}^{d} \frac{\partial[(\partial_{v_i} g(\mathbf{v}) \phi(\mathbf{x}) \psi(\mathbf{v})]}{\partial x_i}$$
$$= \frac{\gamma}{\beta} \phi(\mathbf{x}) \left[\beta \sum_{i=1}^{d} \frac{\partial[(c(\mathbf{v}, \alpha))_i \psi(\mathbf{v})]}{\partial v_i} - \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \psi(\mathbf{v}) \right] + \sum_{i=1}^{d} \frac{\partial[\partial_{x_i} f(\mathbf{x}) \phi(\mathbf{x}) \psi(\mathbf{v})]}{\partial v_i} - \sum_{i=1}^{d} \frac{\partial[\partial_{v_i} g(\mathbf{v}) \phi(\mathbf{x}) \psi(\mathbf{v})]}{\partial x_i}. \tag{S1}$$

Furthermore, we can compute that

$$\beta \sum_{i=1}^{d} \frac{\partial[(c(\mathbf{v}, \alpha))_i \psi(\mathbf{v})]}{\partial v_i} - \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \psi(\mathbf{v}) = \sum_{i=1}^{d} \frac{\partial}{\partial v_i} \left[\frac{\mathcal{D}_{v_i}^{\alpha-2}(\psi(\mathbf{v}) \partial_{v_i} \beta g(\mathbf{v}))}{\psi(\mathbf{v})} \psi(\mathbf{v}) \right] - \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \psi(\mathbf{v})$$
$$= -\sum_{i=1}^{d} \frac{\partial}{\partial v_i} \left[\frac{\mathcal{D}_{v_i}^{\alpha-2}(\partial_{v_i} \psi(\mathbf{v}))}{\psi(\mathbf{v})} \psi(\mathbf{v}) \right] - \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \psi(\mathbf{v})$$
$$= -\sum_{i=1}^{d} \frac{\partial^2}{\partial v_i^2} \mathcal{D}_{v_i}^{\alpha-2} \psi(\mathbf{v}) - \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \psi(\mathbf{v})$$
$$= \sum_{i=1}^{d} \mathcal{D}_{v_i}^{2} \mathcal{D}_{v_i}^{\alpha-2} \psi(\mathbf{v}) - \sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha} \psi(\mathbf{v}) = 0, \tag{S2}$$

[*] Equal contribution [1] LTCI, Télécom Paris, Institut Polytechnique de Paris, Paris, France [2] Department of Statistics, University of Oxford, Oxford, UK [3] Department of Mathematics, Florida State University, Tallahassee, USA [4] Department of Management Science and Information Systems, Rutgers Business School, Piscataway, USA. Correspondence to: Umut Şimşekli <umut.simsekli@telecom-paris.fr>.

Proceedings of the 37^{th} *International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020.

where we used the property $\mathcal{D}^2 u(x) = -\frac{\partial^2}{\partial x^2} u(x)$ (Proposition 1 in (Şimşekli, 2017)) and the semi-group property of the Riesz derivative $\mathcal{D}^a \mathcal{D}^b u(x) = \mathcal{D}^{a+b} u(x)$.

Finally, we can compute that

$$\begin{aligned}
&\sum_{i=1}^{d} \frac{\partial[\partial_{x_i} f(\mathbf{x})\phi(\mathbf{x})\psi(\mathbf{v})]}{\partial v_i} - \sum_{i=1}^{d} \frac{\partial[\partial_{v_i} g(\mathbf{v})\phi(\mathbf{x})\psi(\mathbf{v})]}{\partial x_i} \\
&= \sum_{i=1}^{d} \partial_{x_i} f(\mathbf{x})\phi(\mathbf{x})\partial_{v_i}\psi(\mathbf{v}) - \sum_{i=1}^{d} \partial_{v_i} g(\mathbf{v})\psi(\mathbf{v})\partial_{x_i}\phi(\mathbf{x}) \\
&= -\beta \sum_{i=1}^{d} \frac{\partial f(\mathbf{x})}{\partial_{x_i}}\phi(\mathbf{x})\frac{\partial g(\mathbf{v})}{\partial v_i}\psi(\mathbf{v}) + \beta \sum_{i=1}^{d} \frac{\partial g(\mathbf{v})}{\partial_{v_i}}\psi(\mathbf{v})\frac{\partial f(\mathbf{x})}{\partial x_i}\phi(\mathbf{x}) = 0. \qquad \text{(S3)}
\end{aligned}$$

Therefore, it follows from (S1), (S2) and (S3) that we have

$$\gamma \sum_{i=1}^{d} \frac{\partial[(c(\mathbf{v},\alpha))_i \phi(\mathbf{x})\psi(\mathbf{v})]}{\partial v_i} + \sum_{i=1}^{d} \frac{\partial[\partial_{x_i} f(\mathbf{x})\phi(\mathbf{x})\psi(\mathbf{v})]}{\partial v_i} - \frac{\gamma}{\beta}\sum_{i=1}^{d} \mathcal{D}_{v_i}^{\alpha}\phi(\mathbf{x})\psi(\mathbf{v}) - \sum_{i=1}^{d} \frac{\partial[(\partial_{v_i} g(\mathbf{v})\phi(\mathbf{x})\psi(\mathbf{v})]}{\partial x_i} = 0.$$

Hence we conclude that $\pi(d\mathbf{x}, d\mathbf{v}) = \frac{e^{-\beta(f(\mathbf{x})+g(\mathbf{v}))} d\mathbf{x} d\mathbf{v}}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\beta(f(\mathbf{x}')+g(\mathbf{v}'))} d\mathbf{x}' d\mathbf{v}'}$ is an invariant probability measure. The proof is complete. □

S2. Proof of Theorem 2

Proof. We can compute that

$$(c(\mathbf{v},\alpha))_i = \frac{\mathcal{D}_{v_i}^{\alpha-2}(v_i e^{-\frac{1}{2}\|\mathbf{v}\|^2})}{e^{-\frac{1}{2}\|\mathbf{v}\|^2}} = e^{\frac{1}{2}v_i^2} \mathcal{D}_{v_i}^{\alpha-2}\left(v_i e^{-\frac{1}{2}v_i^2}\right), \qquad \text{(S4)}$$

for every $1 \leq i \leq d$.

Recall the definition of Fourier transform and its inverse:

$$\mathcal{F}\{f(x)\}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ix\omega} f(x) dx, \qquad \mathcal{F}^{-1}\{f(\omega)\}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ix\omega} f(\omega) d\omega. \qquad \text{(S5)}$$

Notice that the Fourier transform of $e^{-\frac{1}{2}x^2}$ is itself, i.e. $\mathcal{F}\{e^{-\frac{1}{2}x^2}\}(\omega) = e^{-\frac{1}{2}\omega^2}$, and moreover, $\mathcal{F}\{x^n f(x)\}(\omega) = i^n \frac{d^n}{d\omega^n}\{\mathcal{F}\{f(x)\}(\omega)\}$, and therefore,

$$\mathcal{F}\left\{x e^{-\frac{1}{2}x^2}\right\}(\omega) = -i\omega e^{-\frac{1}{2}\omega^2}. \qquad \text{(S6)}$$

Hence,

$$\mathcal{D}_x^{\alpha-2}\left(x e^{-\frac{1}{2}x^2}\right) = \mathcal{F}^{-1}\left\{-i\omega|\omega|^{\alpha-2} e^{-\frac{1}{2}\omega^2}\right\}(x) = \frac{-i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega|\omega|^{\alpha-2} e^{-\frac{1}{2}\omega^2 + i\omega x} d\omega. \qquad \text{(S7)}$$

Furthermore, we can compute that

$$\begin{aligned}
\frac{-i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega|\omega|^{\alpha-2} e^{-\frac{1}{2}\omega^2 + i\omega x} d\omega &= \frac{-i}{\sqrt{2\pi}} \int_{0}^{\infty} \omega^{\alpha-1} e^{-\frac{1}{2}\omega^2 + i\omega x} d\omega + \frac{-i}{\sqrt{2\pi}} \int_{-\infty}^{0} \omega(-\omega)^{\alpha-2} e^{-\frac{1}{2}\omega^2 + i\omega x} d\omega \\
&= \frac{-i}{\sqrt{2\pi}} \int_{0}^{\infty} \omega^{\alpha-1} e^{-\frac{1}{2}\omega^2 + i\omega x} d\omega + \frac{i}{\sqrt{2\pi}} \int_{0}^{\infty} \omega^{\alpha-1} e^{-\frac{1}{2}\omega^2 - i\omega x} d\omega \\
&= \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} \omega^{\alpha-1} \sin(\omega x) e^{-\frac{1}{2}\omega^2} d\omega.
\end{aligned}$$

By the Taylor expansion of sine function, we get

$$\begin{aligned}\sqrt{\frac{2}{\pi}}\int_0^\infty \omega^{\alpha-1}\sin(\omega x)e^{-\frac{1}{2}\omega^2}d\omega &= \sqrt{\frac{2}{\pi}}\int_0^\infty \omega^{\alpha-1}\sum_{k=0}^\infty \frac{(-1)^k(\omega x)^{2k+1}}{(2k+1)!}e^{-\frac{1}{2}\omega^2}d\omega \\ &= \sqrt{\frac{2}{\pi}}\sum_{k=0}^\infty \frac{(-1)^k x^{2k+1}}{(2k+1)!}\int_0^\infty \omega^{2k+\alpha}e^{-\frac{1}{2}\omega^2}d\omega \\ &= \sqrt{\frac{2}{\pi}}\sum_{k=0}^\infty \frac{(-1)^k x^{2k+1}}{(2k+1)!}2^{\frac{2k+\alpha-1}{2}}\Gamma\left(\frac{2k+\alpha+1}{2}\right),\end{aligned}$$

where we used the identity $\int_0^\infty x^a e^{-\frac{1}{2}x^2}dx = 2^{\frac{a-1}{2}}\Gamma(\frac{a+1}{2})$, for any given $a > -1$. Moreover, for any given $x, y > 0$, we have the identity:

$$\sum_{k=0}^\infty \frac{(-1)^k x^k}{(2k+1)!}\Gamma(k+y) = \Gamma(y)_1F_1\left(y;\frac{3}{2};-\frac{x}{4}\right), \tag{S8}$$

where ${}_1F_1$ is the Kummer confluent hypergeometric function. Therefore, we conclude that

$$\begin{aligned}\sqrt{\frac{2}{\pi}}\sum_{k=0}^\infty \frac{(-1)^k x^{2k+1}}{(2k+1)!}2^{\frac{2k+\alpha-1}{2}}\Gamma\left(\frac{2k+\alpha+1}{2}\right) &= \sqrt{\frac{2}{\pi}}2^{\frac{\alpha-1}{2}}x\sum_{k=0}^\infty \frac{(-1)^k(2x^2)^k}{(2k+1)!}\Gamma\left(k+\frac{\alpha+1}{2}\right) \\ &= \frac{2^{\frac{\alpha}{2}}x}{\sqrt{\pi}}\Gamma\left(\frac{\alpha+1}{2}\right)\cdot {}_1F_1\left(\frac{\alpha+1}{2};\frac{3}{2};-\frac{x^2}{2}\right).\end{aligned}$$

Hence, we get for every $1 \le i \le d$,

$$(c(\mathbf{v},\alpha))_i = \frac{2^{\frac{\alpha}{2}}v_i e^{\frac{1}{2}v_i^2}}{\sqrt{\pi}}\Gamma\left(\frac{\alpha+1}{2}\right)\cdot {}_1F_1\left(\frac{\alpha+1}{2};\frac{3}{2};-\frac{v_i^2}{2}\right). \tag{S9}$$

By the identity $e^x \cdot {}_1F_1(a;b;-x) = {}_1F_1(b-a;b;x)$, we get

$$(c(\mathbf{v},\alpha))_i = \frac{2^{\frac{\alpha}{2}}v_i}{\sqrt{\pi}}\Gamma\left(\frac{\alpha+1}{2}\right)\cdot {}_1F_1\left(\frac{2-\alpha}{2};\frac{3}{2};\frac{v_i^2}{2}\right). \tag{S10}$$

In particular, when $\alpha = 2$, by applying the identity

$$\sum_{k=0}^\infty \frac{(-1)^k x^k}{(2k+1)!}\Gamma\left(k+\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}e^{-x/4}, \tag{S11}$$

we get

$$\sqrt{\frac{2}{\pi}}\sum_{k=0}^\infty \frac{(-1)^k x^{2k+1}}{(2k+1)!}2^{\frac{2k+1}{2}}\Gamma\left(\frac{2k+3}{2}\right) = \sqrt{\frac{2}{\pi}}2^{\frac{1}{2}}x\sum_{k=0}^\infty \frac{(-1)^k(2x^2)^k}{(2k+1)!}\Gamma\left(k+\frac{3}{2}\right) = xe^{-\frac{x^2}{2}}.$$

The proof is complete. □

S3. Proof of Theorem 3

Proof. Let $\psi_\alpha(x) = e^{-g_\alpha(x)}$ be the probability density function of the symmetric α-stable distribution $\mathcal{S}\alpha\mathcal{S}(\frac{1}{\alpha^{1/\alpha}})$ such that

$$\mathcal{F}\{\psi_\alpha(x)\}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-i\omega x}\psi_\alpha(x)dx = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{\alpha}|\omega|^\alpha}. \tag{S12}$$

Therefore, we get

$$\begin{aligned}
\mathcal{D}_x^{\alpha-2}(\psi_\alpha(x)\partial_x g_\alpha(x)) &= -\mathcal{D}_x^{\alpha-2}(\partial_x \psi_\alpha(x)) \\
&= -\mathcal{F}^{-1}\left\{|\omega|^{\alpha-2}\mathcal{F}\left\{\partial_x\psi_\alpha(x)\right\}(\omega)\right\}(x) \\
&= -\mathcal{F}^{-1}\left\{|\omega|^{\alpha-2}(i\omega)\mathcal{F}\left\{\psi_\alpha(x)\right\}(\omega)\right\}(x) \\
&= \frac{-i}{\sqrt{2\pi}}\mathcal{F}^{-1}\left\{|\omega|^{\alpha-2}\omega e^{-\frac{1}{\alpha}|\omega|^\alpha}\right\}(x) \\
&= \frac{i}{\sqrt{2\pi}}\mathcal{F}^{-1}\left\{\partial_\omega e^{-\frac{1}{\alpha}|\omega|^\alpha}\right\}(x) \\
&= \frac{i}{\sqrt{2\pi}}(-ix)\mathcal{F}^{-1}\left\{e^{-\frac{1}{\alpha}|\omega|^\alpha}\right\}(x) \\
&= x\psi_\alpha(x).
\end{aligned}$$

Hence, we conclude that

$$\frac{\mathcal{D}_x^{\alpha-2}(\psi_\alpha(x)\partial_x g_\alpha(x))}{\psi_\alpha(x)} = x, \tag{S13}$$

and it follows that

$$(c(\mathbf{v},\alpha))_i = v_i, \qquad 1 \leq i \leq d. \tag{S14}$$

The proof is complete. □

S4. Proof of Proposition 1

Proof. It is straightforward to verify that the result holds for the cases $\alpha = 1$ and $\alpha = 2$. Assume $\alpha \in (0,1)$ or $\alpha \in (1,2)$. Let X be the unit symmetric α-stable random variable defined by its characteristic function

$$\phi_X(t) := \mathbb{E}(e^{itX}) = e^{-|t|^\alpha}.$$

By taking inverse Fourier transformation, its density $\psi_\alpha(x) = e^{-g_\alpha(x)}$ can be expressed as

$$\psi_\alpha(x) := \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_X(t)e^{-itx}dt.$$

Writing $e^{-itx} = \cos(tx) - i\sin(tx)$, we compute

$$\psi_\alpha(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-|t|^\alpha}\left[\cos(tx) - i\sin(tx)\right]dt = \frac{1}{\pi}\int_0^\infty e^{-t^\alpha}\cos(tx)dt\,, \tag{S15}$$

where we used the fact that $\phi_X(t)$ and $\cos(tx)$ are even functions of t, whereas $\sin(tx)$ is an odd function of t. If we define,

$$g_\alpha(x) = -\log(\psi_\alpha(x)),$$

then

$$g'_\alpha(x) = \frac{\psi'_\alpha(x)}{\psi_\alpha(x)}\,, \tag{S16}$$

where the superscript $'$ denotes derivative with respect to x. Similarly,

$$g''_\alpha(x) = \frac{\psi''_\alpha(x)}{\psi_\alpha(x)} - \left(\frac{\psi'_\alpha(x)}{\psi_\alpha(x)}\right)^2. \tag{S17}$$

If $g''_\alpha(x)$ is uniformly bounded over $x \in \mathbb{R}$, it can be seen that the map $v \mapsto g'_\alpha(v)$ will be Lipschitz. Therefore, it suffices to show that $x \mapsto g''_\alpha(x)$ is a bounded function on the real line. Note that the function $\psi_\alpha(x)$ is infinitely many differentiable, and the integral (S15) is absolutely convergent. Therefore, we can differentiate both sides of (S15) with respect to x to obtain

$$\psi'_\alpha(x) := \frac{1}{\pi}\int_0^\infty -te^{-t^\alpha}\sin(tx)dt\,,$$
$$\psi''_\alpha(x) := \frac{1}{\pi}\int_0^\infty -t^2e^{-t^\alpha}\cos(tx)dt\,.$$

In particular, since $|\cos(tx)| \leq 1$ and $|\sin(tx)| \leq 1$ this implies that

$$\psi'_\alpha(x) \leq M_1(\alpha) := \frac{1}{\pi}\int_0^\infty te^{-t^\alpha}dt < \infty\,,$$
$$\psi''_\alpha(x) \leq M_2(\alpha) := \frac{1}{\pi}\int_0^\infty t^2e^{-t^\alpha}dt < \infty\,.$$

It is also well-known that a symmetric α stable random variable has a decay in its density satisfying $\psi_\alpha(x) \sim \frac{1}{|x|^{1+\alpha}}$ when $|x|$ is large. In fact, Wintner (1941) derived a large-x expansion for $\psi_\alpha(x)$ when $0 < \alpha < 1$ and $x > 0$. This expansion is equivalent to

$$\begin{aligned}\psi_\alpha(x) &= \frac{1}{\pi}\sum_{n=1}^\infty \frac{(-1)^{n+1}}{n!}\frac{\Gamma(1+\alpha n)}{x^{\alpha n+1}}\sin\left(\frac{\pi\alpha n}{2}\right)\\ &= \frac{1}{\pi}\left(\frac{\Gamma(1+\alpha)}{x^{\alpha+1}}\sin(\pi\alpha/2) - \frac{\Gamma(1+2\alpha)}{2x^{2\alpha+1}}\sin(\pi\alpha) + \frac{\Gamma(1+3\alpha)}{6x^{3\alpha+1}}\sin(\pi3\alpha/2) + \cdots\right),\end{aligned}$$

(see eqn. (11) from (Montroll & Bendler, 1984)) where it can be seen from the Stirling's approximation of the gamma function and the ratio test that the series converges absolutely. A similar absolutely convergent series sum (with exactly the same leading term) is also available in the literature for $\alpha \in (1, 2)$ which says that

$$\psi_\alpha(x) = \frac{1}{\pi}\frac{\Gamma(1+\alpha)}{x^{\alpha+1}}\sin\left(\frac{\pi\alpha}{2}\right) + O\left(\frac{1}{x^{2\alpha}}\right)$$

(see eqn. (3.58) from (Montroll & West, 1979)). By differentiating the series sum for $\psi_\alpha(x)$ with respect to x, we can express $\psi'_\alpha(x)$ and $\psi''_\alpha(x)$ as a series sum. After a straightforward computation, we obtain

$$\frac{\psi''_\alpha(x)}{\psi_\alpha(x)} = O\left(\frac{1}{x^2}\right), \quad \left(\frac{\psi'_\alpha(x)}{\psi_\alpha(x)}\right)^2 = O\left(\frac{1}{x^2}\right),$$

which implies from (S17) that $g''_\alpha(x) \to 0$ as $x \to \infty$. This shows that $g''_\alpha(x)$ is bounded on the interval $[0, \infty)$. On the other hand, $\psi_\alpha(x)$ is an even function and therefore $g''_\alpha(x)$ is an even function satisfying $g''_\alpha(x) = g''_\alpha(-x)$. We conclude that $g''_\alpha(x)$ is bounded on the real line. This completes the proof. □

S5. Proof of Corollary 1

Proof. By Proposition 1, we know that ∇G_α is Lipschitz and by our hypthesis ∇f is also Lipschitz and has linear growth. Then the process (19) admits a unique invariant measure (cf. (Schertzer et al., 2001) Section 9), which is given by Theorem 1. The rest of the proof follows from (Panloup, 2008) (Theorem 2). □

S6. Alternative forms of the drift function c with the Gaussian kinetic energy

For some special values of α, we can get alternative formulas for $(c(\mathbf{v}, \alpha))_i$, $1 \leq i \leq d$.

(1) $\alpha = \frac{3}{2}$. Using the identity ${}_1F_1(a; 2a+1; z) = 2^{2a-1}\Gamma(a+\frac{1}{2})e^{\frac{z}{2}}z^{\frac{1}{2}-a}(I_{a-\frac{1}{2}}(\frac{z}{2}) - I_{a+\frac{1}{2}}(\frac{z}{2}))$, where $I_a(x)$ is the modified Bessel function of the first kind, we get

$$(c(\mathbf{v}, \alpha))_i = \frac{2^{\frac{1}{4}}v_i}{\sqrt{\pi}}\Gamma\left(\frac{5}{4}\right)\Gamma\left(\frac{3}{4}\right)e^{\frac{v_i^2}{4}}\left(\frac{v_i^2}{2}\right)^{\frac{1}{4}}\left(I_{-\frac{1}{4}}\left(\frac{v_i^2}{4}\right) - I_{\frac{3}{4}}\left(\frac{v_i^2}{4}\right)\right), \tag{S18}$$

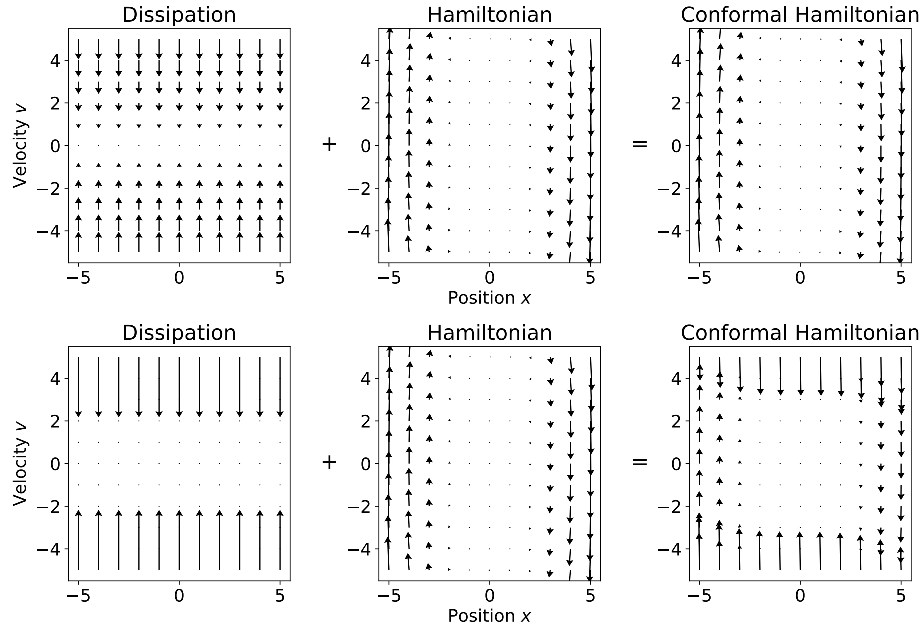

Figure S1. Conformal Hamiltonian fields with the Gaussian kinetic energy for $f(x) = x^4/4$. Top $\alpha = 2$, bottom $\alpha = 1.7$.

for every $1 \leq i \leq d$.

(2) $\alpha = \frac{1}{2}$. Using the identity ${}_1F_1(a; 2a; z) = 2^{2a-1}\Gamma(a + \frac{1}{2})z^{\frac{1}{2}-a}e^{\frac{z}{2}}I_{a-\frac{1}{2}}(\frac{z}{2})$, we get

$$(c(\mathbf{v}, \alpha))_i = \frac{2^{\frac{3}{4}} v_i}{\sqrt{\pi}} \Gamma\left(\frac{3}{4}\right)\Gamma\left(\frac{5}{4}\right)\left(\frac{v_i^2}{2}\right)^{-\frac{1}{4}} e^{\frac{v_i^2}{4}} I_{\frac{1}{4}}\left(\frac{v_i^2}{4}\right), \tag{S19}$$

for every $1 \leq i \leq d$.

S7. Visual Illustrations

In order to have a better grasp on the dynamics (16) in an optimization context, we also investigate its deterministic part (i.e., (16) without the L_t^α term) as a *conformal Hamiltonian system* (Maddison et al., 2018), where we decompose the overall dynamics into two: the dissipative part $\mathrm{d}(\mathbf{x}_t, \mathbf{v}_t) = (0, -\gamma c(\mathbf{v}_{t-}, \alpha))\mathrm{d}t$ and the Hamiltonian part $\mathrm{d}(\mathbf{x}_t, \mathbf{v}_t) = (\mathbf{v}_t, -\nabla f(\mathbf{x}_t))\mathrm{d}t$, whose combination gives the conformal Hamiltonian. The two parts have different semantics: the Hamiltonian part tries to keep the overall energy of the system ($\nabla f(\mathbf{x}) + \|\mathbf{v}\|^2/2$) constant, while the dissipative part tries to reduce this energy, and this competition determines the behavior of the overall system. In Figure S1, we visualize the conformal Hamiltonians for $f(x) = x^4/4$ for two different values of α. This choice of f is known to be problematic for the classical overdamped dynamics (Maddison et al., 2018; Brosse et al., 2019), which can be clearly observed from Figure S1 (top right) as the conformal Hamiltonian field tends to diverge. On the other hand, for $\alpha = 1.7$, we observe that the strong dissipation, which was introduced due to tolerate heavy-tailed perturbations, can also compensate for fast-growing f.

On the other hand, we visualize the conformal Hamiltonian field generated by this dynamics in Figure S2 for $f(x) = g_1(x) = -\log \frac{1}{\pi}\frac{1}{x^2+1}$. The figure shows that conformal Hamiltonian generated by the dynamics with $\alpha = 2$ has a very slow concentration behavior towards the minimum at the origin, whereas this behavior is alleviated when $\alpha = 1.7$ where the field concentrates faster.

S8. Additional Experimental Results

In this section, we provide the additional experimental results that were mentioned in the main document for width 32, 64, and 512.

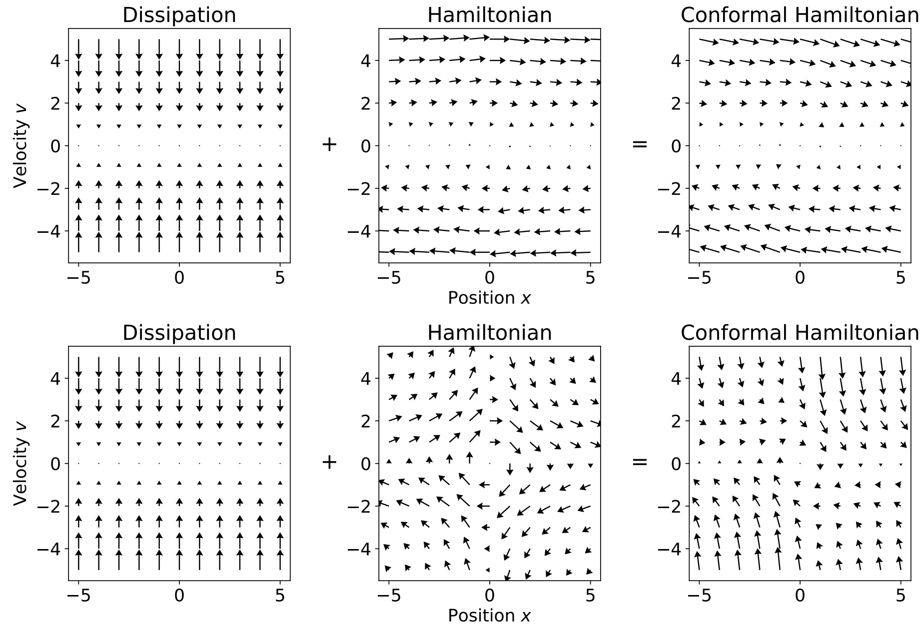

Figure S2. Conformal Hamiltonian fields with the $\mathcal{S}\alpha\mathcal{S}$ kinetic energy for $f(x) = -\log \frac{1}{\pi} \frac{1}{x^2+1}$. Top $\alpha = 2$, bottom $\alpha = 1.7$.

References

Brosse, N., Durmus, A., Moulines, É., and Sabanis, S. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

Şimşekli, U. Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov Chain Monte Carlo. In *International Conference on Machine Learning*, pp. 3200–3209, 2017.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

Montroll, E. W. and Bendler, J. T. On Lévy (or stable) distributions and the Williams-Watts model of dielectric relaxation. *Journal of Statistical Physics*, 34(1):129–162, Jan 1984.

Montroll, E. W. and West, B. J. Chapter 2 - On an enriched collection of stochastic processes. In Montroll, E. and Lebowitz, J. (eds.), *Fluctuation Phenomena*, pp. 61 – 175. Elsevier, 1979.

Panloup, F. Recursive computation of the invariant measure of a stochastic differential equation driven by a Lévy process. *Annals of Applied Probability*, 18(2):379–426, 2008.

Schertzer, D., Larchevêque, M., Duan, J., Yanovsky, V., and Lovejoy, S. Fractional Fokker-Planck equation for nonlinear stochastic differential equations driven by non-Gaussian Lévy stable noises. *Journal of Mathematical Physics*, 42(1): 200–212, 2001.

Wintner, A. The singularities of Cauchy's distributions. *Duke Math. J.*, 8(4):678–681, 12 1941.

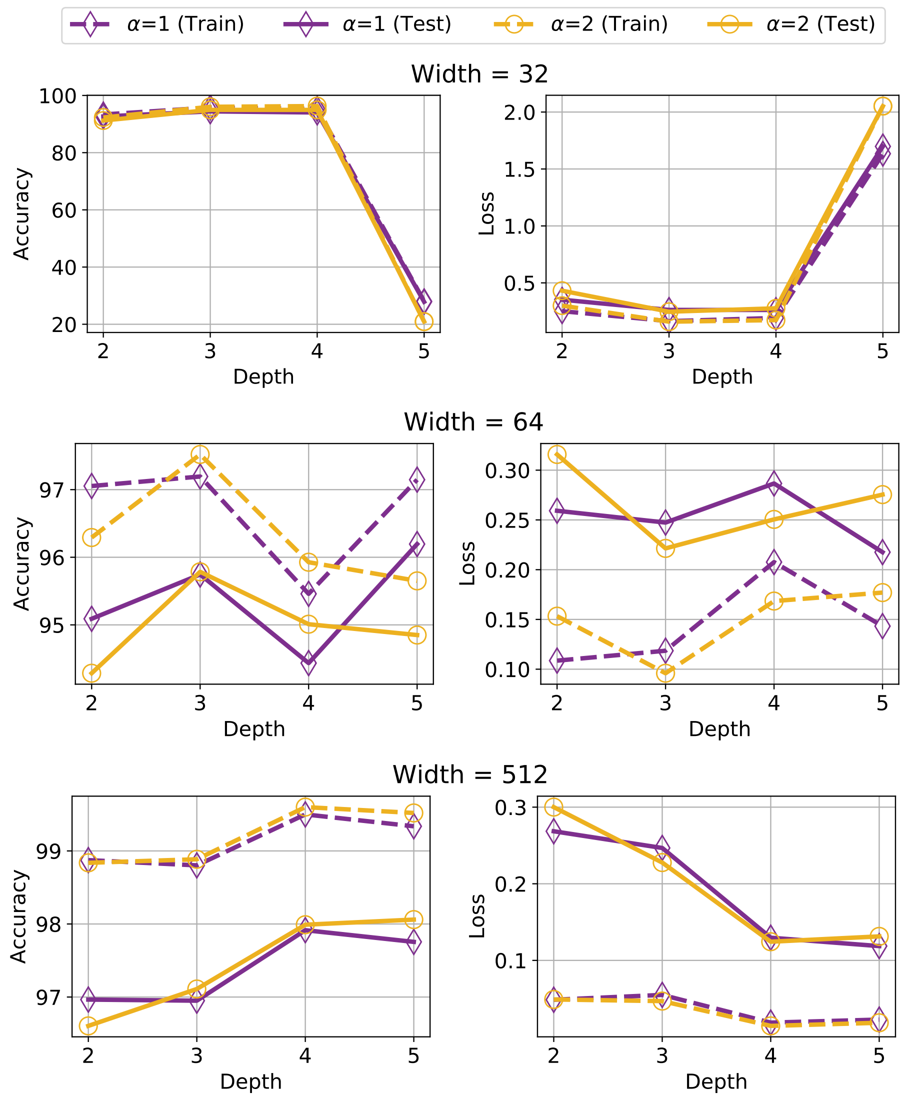

Figure S3. Neural network results on MNIST.