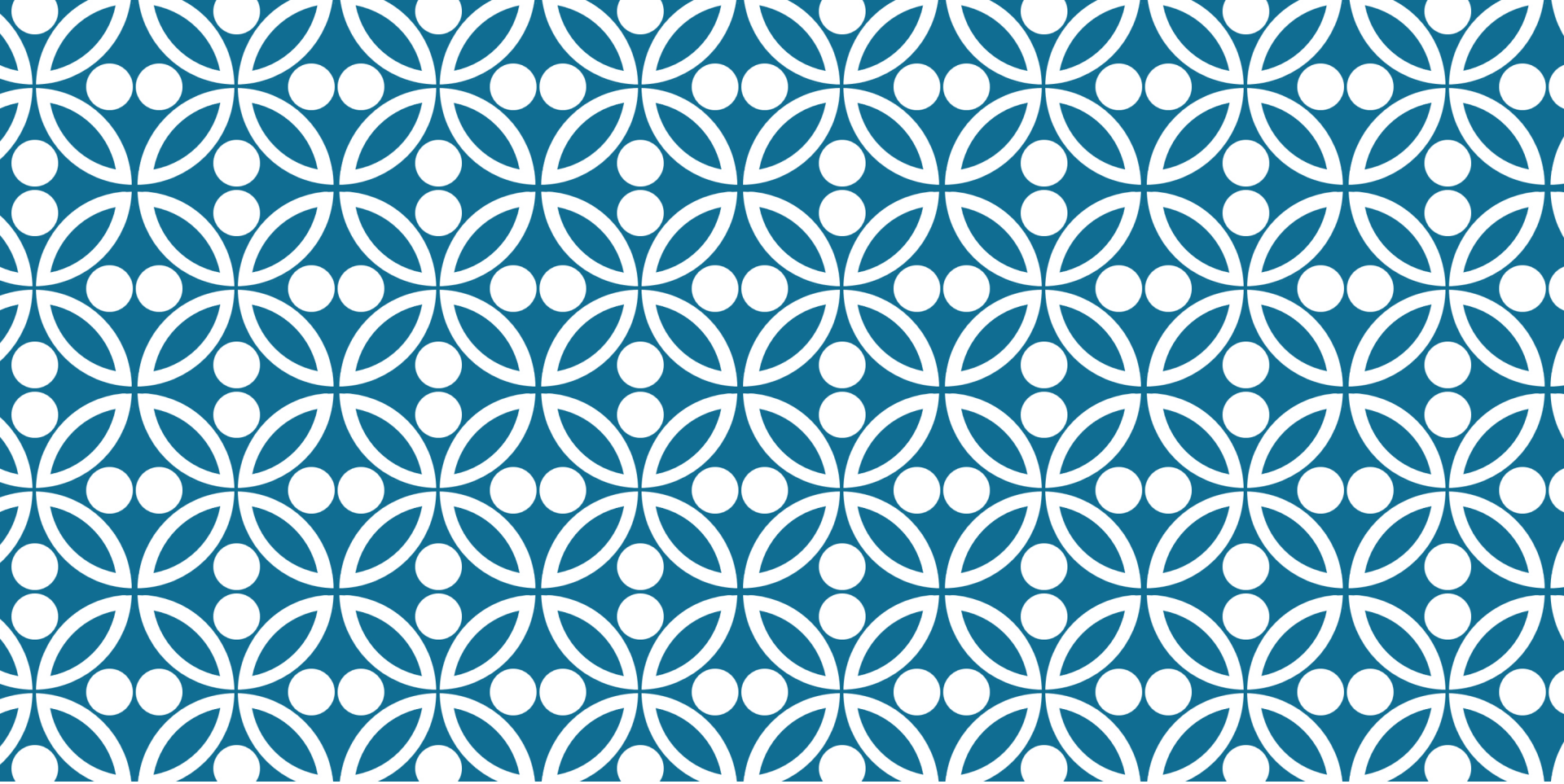


FRACTIONAL UNDERDAMPED LANGEVIN DYNAMICS: RETARGETING SGD WITH MOMENTUM UNDER HEAVY-TAILED GRADIENT NOISE

Umut Şimşekli*, Lingjiong Zhu*, Yee Whye Teh, Mert Gürbüzbalaban

(Florida State University) (University of Oxford) (Rutgers University)

*equal contribution

International Conference on Machine Learning, 2020

Umut Şimşekli

LTCI, Télécom Paris,

Institut Polytechnique de Paris

DEEP LEARNING & SGD-MOMENTUM

- Deep learning (in general)

$$\mathbf{x}^\star = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min}\left\{ f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f^{(i)}(\mathbf{x}) \right\}$$

network weights; non-convex cost function; data points

- Optimization Algorithm – **S**tochastic **G**radiend **D**escent with **momentum**

velocity (momentum); step-size (learning rate); stochastic gradient

$$\tilde{\mathbf{v}}^{k+1} = \tilde{\gamma}\tilde{\mathbf{v}}^{k} - \tilde{\eta}\nabla\tilde{f}_{k+1}(\mathbf{x}^k)$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \tilde{\mathbf{v}}^{k+1}$$

Momentum decay

$$\nabla\tilde{f}_k(\mathbf{x}) \triangleq \frac{1}{b}\sum_{i\in\Omega_k} f^{(i)}(\mathbf{x})$$

minibatch size; minibatch

UNDERSTANDING SGD-M

- Theory: better-established for **convex** problems

  still in early phase for **non-convex** problems

- Useful approach for analysis → **S**tochastic **D**ifferential **E**quations (SDE)

  gradient noise: $U_k(\mathbf{x}) \triangleq \nabla \tilde{f}_k(\mathbf{x}) - \nabla f(\mathbf{x})$

- If we assume $\mathbb{E}\|U_k(\mathbf{x})\|^2 < \infty$ and invoke CLT: $U_k \sim \text{Gaussian}$

- SGD-m → Euler-Maruyama **Discretization of the SDE:**

Underdamped (a.k.a. Kinetic) Langevin Dynamics

$$\mathrm{d}\mathbf{v}_t = -(\gamma \mathbf{v}_t + \nabla f(\mathbf{x}_t))\mathrm{d}t + \sqrt{\frac{2\gamma}{\beta}}\mathrm{dB}_t$$

$$\mathrm{d}\mathbf{x}_t = \mathbf{v}_t \mathrm{d}t$$

friction

Inverse temperature

Brownian Motion

Dalalyan&Riou-Durand'18
Gao et al.'18

UNDERDAMPED LANGEVIN DYNAMICS

$$\mathrm{d}\mathbf{v}_t = -(\gamma \mathbf{v}_t + \nabla f(\mathbf{x}_t))\mathrm{d}t + \sqrt{\frac{2\gamma}{\beta}}\mathrm{dB}_t$$
$$\mathrm{d}\mathbf{x}_t = \mathbf{v}_t \mathrm{d}t$$

- Favorable properties:

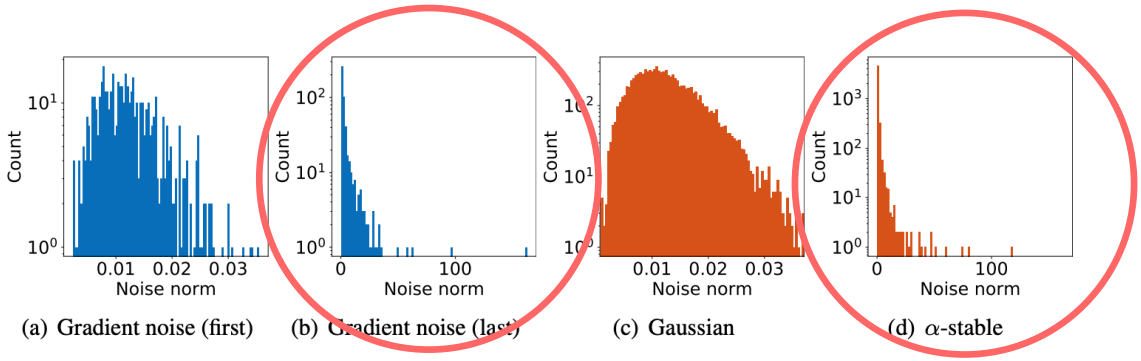
  Targets **Boltzmann-Gibbs** measure → marginal density $\propto \exp(-\beta f(x))$

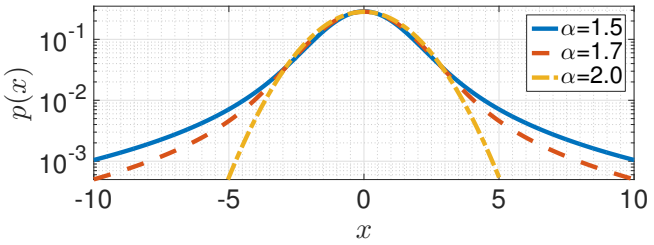
  Minima of f $\Leftrightarrow$ Maxima of the target measure

  As $\beta \to \infty$, target concentrates on the global minimum of f

- **Problem:** gradient noise → **non-Gaussian heavy-tailed** in deep nets

(Simsekli et al.'19)





Gaussian **when α=2**
Infinite variance **when α≠2**

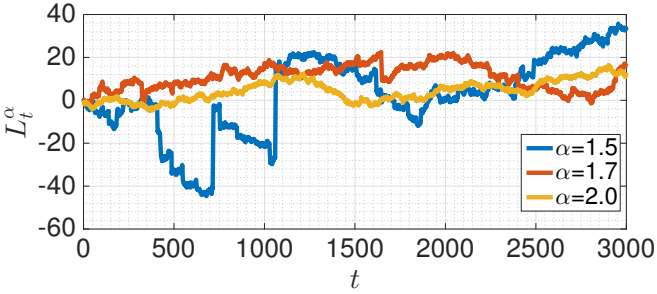
DYNAMICS WITH HEAVY-TAILED NOISE

- Gaussian increments → Brownian motion

  α-stable increments → α-stable Lévy motion

$$\mathrm{d}\mathbf{v}_t = -(\gamma \mathbf{v}_{t-} + \nabla f(\mathbf{x}_t))\mathrm{d}t + \sqrt{\frac{2\gamma}{\beta}}\mathrm{dL}_t^{\alpha}$$

$$\mathrm{d}\mathbf{x}_t = \mathbf{v}_t \mathrm{d}t$$

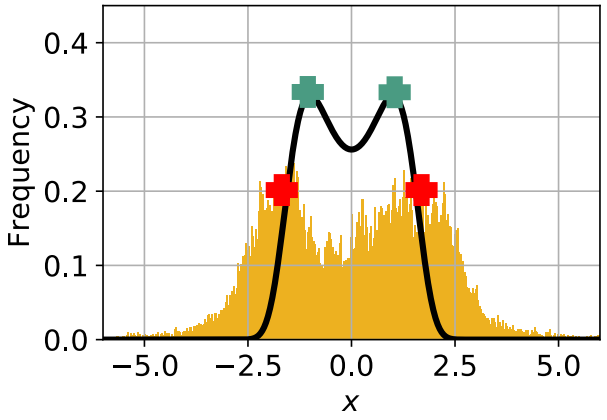


- **Problem:** doesn't target the Gibbs measure!

  The modes are **shifted** → **bias**

  Analytical results for bias when $\gamma \to \infty$

  (Sliusarenko et al.'13)
  (Capała & Dybiec'19)



MAIN QUESTION & GOALS

Can we design an SDE that targets the Gibbs measure under α–stable gradient noise?

- In other words: "**retarget** SGD-m towards the Gibbs measure"
- Make sure → the modes match the minima of f

  Make sure → the dynamics is computationally tractable

  Overdamped case → retargeting possible but **not** tractable (Simsekli'17)

- Relation to **gradient clipping** and **differential geometric** approaches

- Asymptotic analysis of the discretization scheme

PROPOSED DYNAMICS

Theorem 1 (Fractional Underdamped Langevin Dynamics – FULD)

Consider the dynamics

Drift

$$\mathrm{d}\mathbf{v}_t = -(\gamma c(\mathbf{v}_{t-}, \alpha) + \nabla f(\mathbf{x}_t))\mathrm{d}t + \left(\frac{\gamma}{\beta}\right)^{1/\alpha} \mathrm{dL}_t^\alpha$$

$$\mathrm{d}\mathbf{x}_t = \nabla g(\mathbf{v}_t)\mathrm{d}t$$

with $(c(\mathbf{v}, \alpha))_i := \dfrac{\mathcal{D}_{v_i}^{\alpha-2}(\psi(\mathbf{v})\partial_{v_i} g(\mathbf{v}))}{\psi(\mathbf{v})}$, $\psi(\mathbf{v}) := e^{-g(\mathbf{v})}$ Kinetic Energy

Then, the Boltzmann-Gibbs measure is **an** invariant measure of this SDE.

- $\mathcal{D}_v^\gamma$: fractional **Riesz** derivative → non-local → often no analytical expression hard to approximate (Simsekli'17)
- General recipe: specify the Kinetic energy **g** → imposes a drift **c**
- Two choices of g: 1) Gaussian 2) α-stable → **analytical** Riesz derivatives!

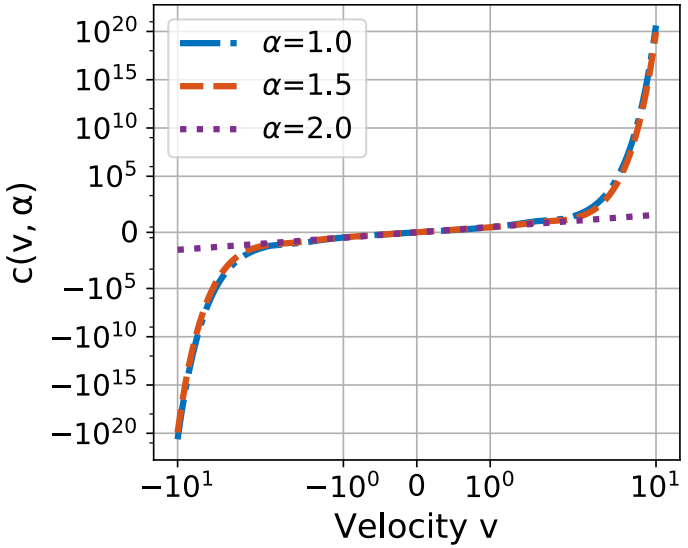
GAUSSIAN KINETIC ENERGY

Theorem 2 (FULD with Gaussian Kinetic energy)

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

Gaussian — Gamma fn. — Kummer confluent hypergeometric fn.

- α=2, recovers standard ULD
- Non-Lipschitz, explosive
- Uniqueness not guaranteed
- Doesn't have much practical value



ALPHA-STABLE KINETIC ENERGY

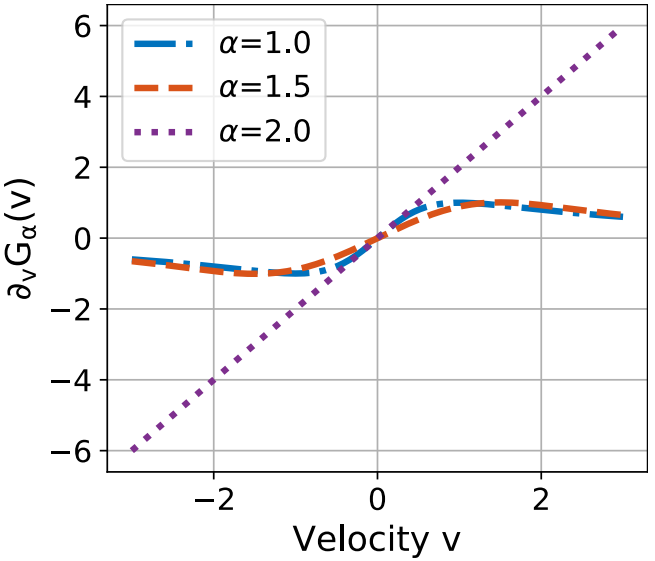
Theorem 3 (FULD with α-Stable Kinetic energy)

Let $e^{-g_\alpha(\mathbf{v})}$ be the pdf of the α-stable distribution.
Define the kinetic energy: $\psi(\mathbf{v}) = e^{-G_\alpha(\mathbf{v})}$ with $G_\alpha(\mathbf{v}) = \sum_{i=1}^{d} g_\alpha(v_i)$.

Then, $(c(\mathbf{v}, \alpha))_i = v_i$.

- Resulting SDE:

$$\mathrm{d}\mathbf{v}_t = -\gamma \mathbf{v}_{t-}\mathrm{d}t - \nabla f(\mathbf{x}_t)\mathrm{d}t + \gamma^{1/\alpha}\mathrm{dL}_t^\alpha,$$
$$\mathrm{d}\mathbf{x}_t = \nabla G_\alpha(\mathbf{v}_t)\mathrm{d}t$$



- **Proposition:** ∇G_α is Lipschitz for all α.
- **Uniqueness:** with standard conditions on f

DISCRETIZATION

- Analysis: weak convergence of the Euler-Maruyama discretization

$\tilde{\gamma}_k = 1 - \gamma\eta_k$

$$\mathbf{v}^{k+1} = \tilde{\gamma}_k \mathbf{v}^k - \eta_k \nabla f(\mathbf{x}^k) + \left(\frac{\eta_k \gamma}{\beta}\right)^{1/\alpha} \mathbf{s}^{k+1}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \eta_k \nabla G_\alpha(\mathbf{v}^{k+1})$$

$\mathbf{s}^{k+1}$ Standard α-stable r.v.

- Our interest: $\pi(h) := \mathbb{E}_{X\sim\pi}[h(X)]$ via $\bar{\pi}_K(h) := \dfrac{\sum_{k=1}^K \eta_k h(\mathbf{x}^k)}{\sum_{k=1}^K \eta_k}$

Corollary 1 (Weak Convergence of Sample Averages)

Assume: $\lim_{k\to\infty} \eta_k = 0$, $\lim_{k\to\infty} \sum_{i=1}^k \eta_i = \infty$

Assume: ∇f is Lipschitz + linear growth + standard Lyapunov condition

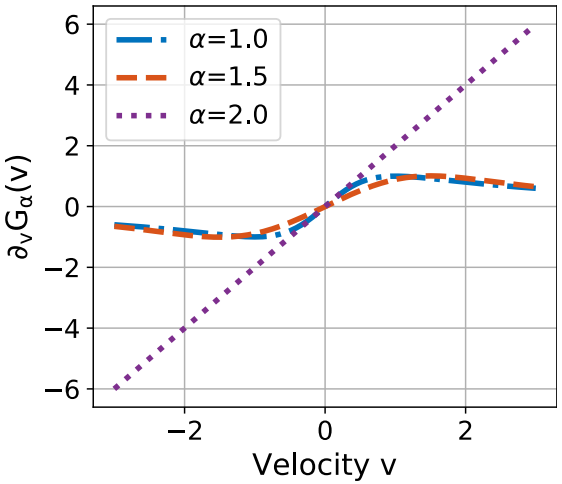
Then, $\bar{\pi}_K(h) \to \pi(h)$ almost surely, as $K \to \infty$.

CONNECTIONS

- **Gradient Clipping**: used when "exploding gradients" (Pascanu et al.'13)

Heavy-tails → exploding grads → clipping improves convergence (Zhang et al.'19)

- Naturally arises in our scheme:



- **Natural gradient** (Amari'98)
- Precondition by Fisher Information

$$\mathbb{E}[\nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top]$$

- The "Composite" gradient: (α=1)

$$\nabla G_1(\nabla \tilde{f}_k(\mathbf{x})) = \mathbf{M}_k(\mathbf{x})^{-1} \nabla \tilde{f}_k(\mathbf{x})$$

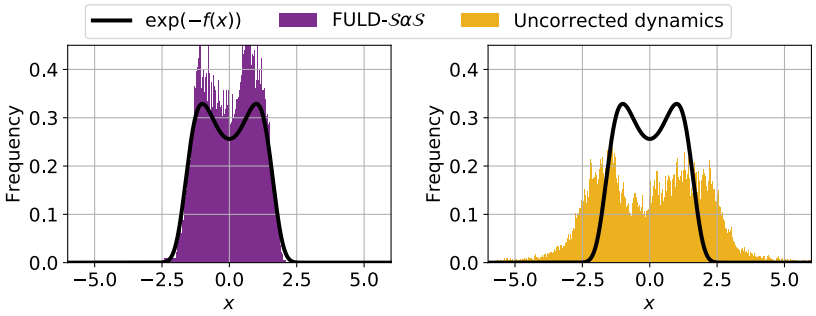
with

$$m_{ii} = ((\nabla \tilde{f}_k(\mathbf{x}))_i^2 + 1)/2$$

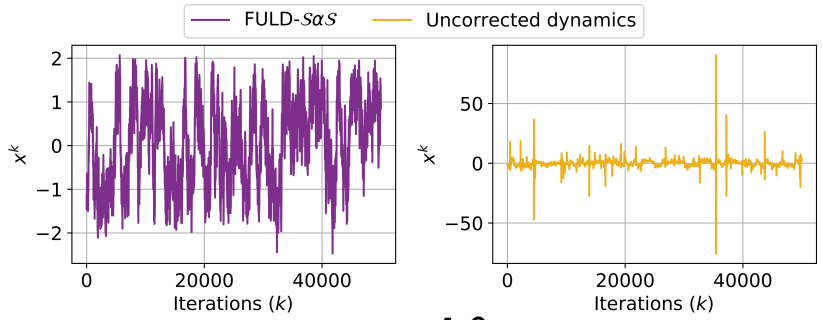
- Can be seen as a (biased) estimator of the diagonal of FIM

EXPERIMENTS – SYNTHETIC

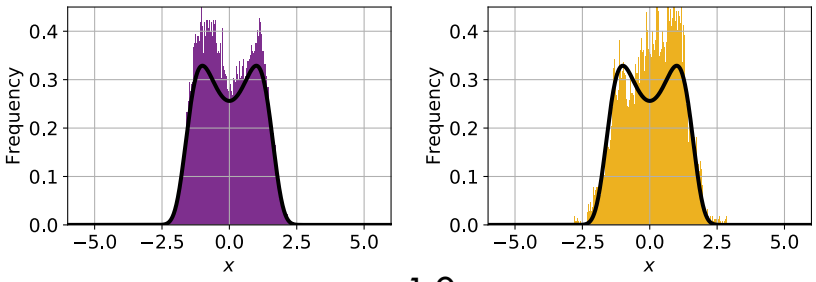
- f: "Double-well" (4-th order polynomial)



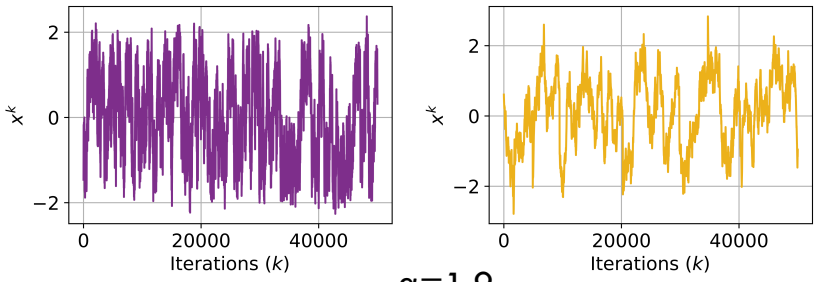
α=1.0



α=1.0



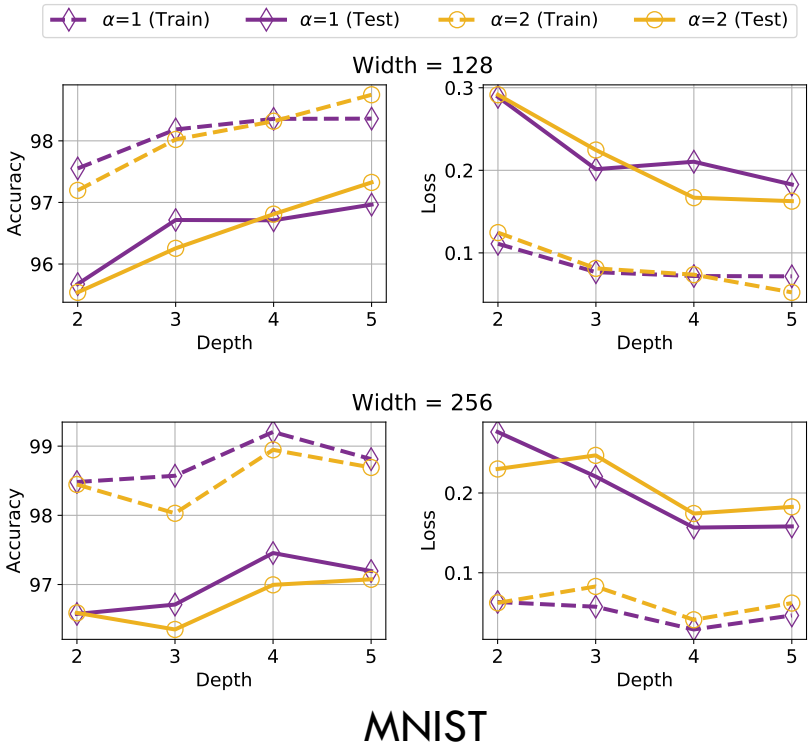
α=1.9



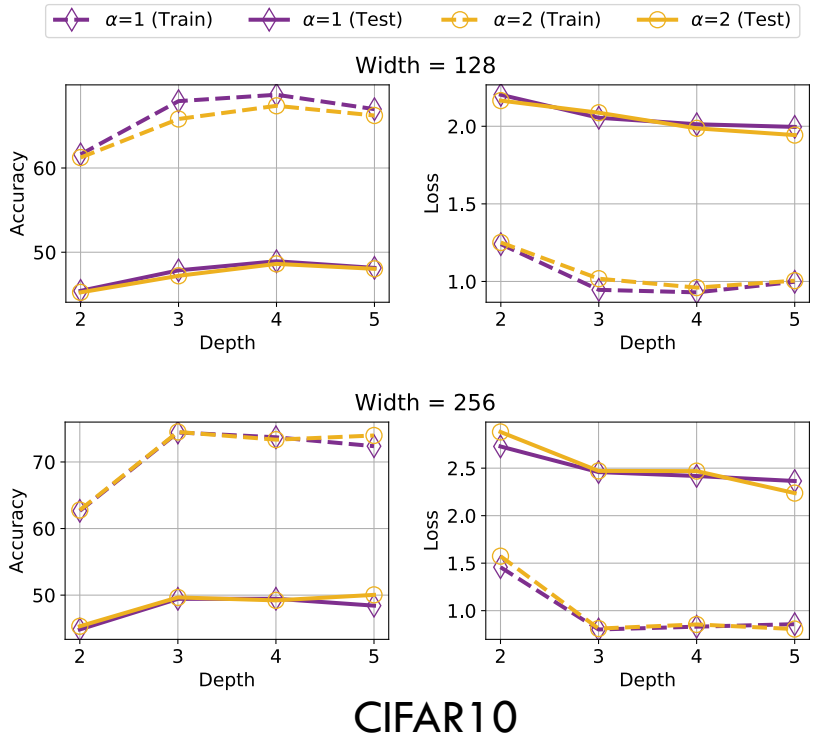
α=1.9

EXPERIMENTS – NEURAL NETWORKS

- Fully connected networks – ReLU activations – Cross Entropy loss
- No additional noise – true gradient noise



MNIST



CIFAR10

CONCLUSION

- Gradient noise in deep networks can be heavy-tailed
- Heavy-tails → might shift the modes
- Proposed dynamics → targets the Gibbs measure → no shift
- Brings theoretical understanding for gradient clipping

THANK YOU FOR YOU ATTENTION!