First Exit Time Analysis of Stochastic Gradient Descent Under Heavy-Tailed Gradient Noise

Thanh Huy Nguyen[1] Umut Şimşekli[1,2] Mert Gürbüzbalaban[3] Gaël Richard[1]

1: Télécom Paris, Institut Polytechnique de Paris 2: University of Oxford 3: Rutgers Business School

Supported by the French National Research Agency (**ANR**) as a part of the **FBIMATRIX** project (ANR-16-CE23-0014)

Introduction & Context

- **Non-convex optimization problem:**

$$\min_{w\in\mathbb{R}^d} f(w) = (1/n)\sum_{i=1}^{n} f^{(i)}(w),$$

$w\in\mathbb{R}^d$, $f^{(i)}:\mathbb{R}^d\mapsto\mathbb{R}$: corresponds to the i-th data point.

- **SGD iterations:** $W^{k+1} = W^k - \eta\nabla\tilde{f}_k(W^k)$

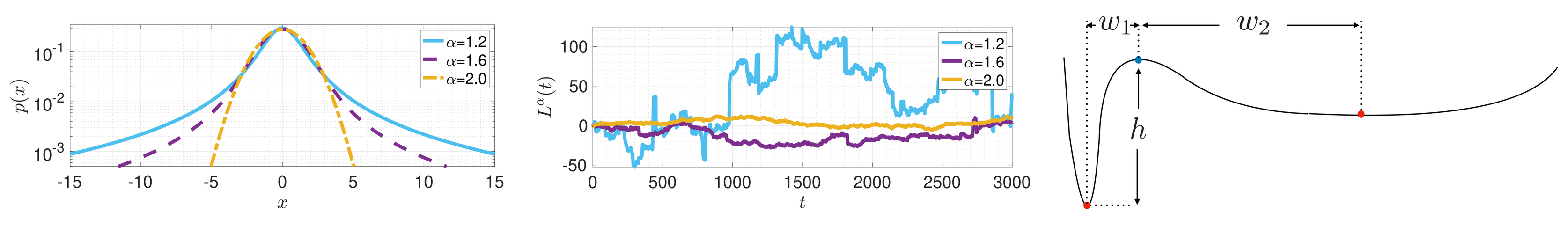
$$\nabla\tilde{f}_k(W^k) \triangleq \nabla\tilde{f}_{\Omega_k}(W^k) \triangleq (1/b)\sum_{i\in\Omega_k}\nabla f^{(i)}(W^k)$$

Stochastic Gradient Noise: $\nabla\tilde{f}_k(W^k) - \nabla f(W^k)$

- **Deep Neural Networks:** noise can have heavy tails [1] .
- **SGD as a discretization of an SDE with α-stable noise:**

$$\mathrm{d}W(t) = -\nabla f(W(t))\mathrm{d}t + \eta^{\frac{\alpha-1}{\alpha}}\sigma\mathrm{d}L^\alpha(t),$$

$L^\alpha(t)$: d-dim. α-stable motion with indep. components.



- **The first exit time → Wide minima** [2]
  α-stable systems: polynomial in the **width** of the basin
  for Brownian systems: exponential in the **height** of the basin.
- **"Does the discrete-time system have the same behaviour?"**

Theoretical Framework

- In this work, consider

$$\mathrm{d}W(t) = -\nabla f(W(t-))\mathrm{d}t + \varepsilon\sigma\mathrm{d}B(t) + \varepsilon\mathrm{d}L^\alpha(t),$$
$$W^{k+1} = W^k - \eta\nabla f(W^k) + \varepsilon\sigma\eta^{1/2}\xi_k + \varepsilon\eta^{1/\alpha}\zeta_k,$$

$\xi_k\sim\mathcal{N}(0,I)$, the components of ζ_k are i.i.d with $\mathcal{S}\alpha\mathcal{S}(1)$.

- The *first exit times* for $W(t)$ and W^k:

$$\tau_{\psi,a}(\varepsilon) \triangleq \inf\{t\geq 0 : \|W(t)-\bar{w}\|\notin[0,a+\psi]\},$$
$$\bar{\tau}_{\psi,a}(\varepsilon) \triangleq \inf\{k\in\mathbb{N} : \|W^k-\bar{w}\|\notin[0,a+\psi]\},$$

$\bar{w}$: local minimum of f, $a>0$, initial $W(0)$: $\|W(0)-\bar{w}\|\leq a$.

- **Goal:** Derive explicit conditions for the step-size s.t
  First exit time of discrete syst. ≈ First exit time of cont. syst.

Main Assumptions

Assumption: (Hölder continuity) For $\frac{1}{2} < \gamma < \min\{\frac{1}{\sqrt{2}}, \frac{\alpha}{2}\}$,

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x-y\|^\gamma, \qquad \forall x,y\in\mathbb{R}^d.$$

Assumption: (Dissipativity) For $m>0$ and $b\geq 0$:

$$\langle x, \nabla f(x)\rangle \geq m\|x\|^{1+\gamma} - b, \forall x\in\mathbb{R}^d.$$

Assumption: For $\delta>0$, $t=K\eta$, and for some $C>0$: $0<\eta\leq$

$$\min\left\{1, \frac{m}{M^2}, \left(uK_1\right)^{\frac{-1}{\gamma^2+2\gamma-1}}, \left(uK_2\right)^{\frac{-1}{2\gamma}}, \left(uK_3\right)^{\frac{-\alpha}{2\gamma}}, \left(uK_4\right)^{\frac{-1}{\gamma}}\right\},$$

where $u = 2t^2/\delta^2$, $K_1 = \mathcal{O}(d\varepsilon^{2\gamma^2-2})$, $K_2 = \mathcal{O}(\varepsilon^{-2})$, $K_3 = \mathcal{O}(d^{2\gamma}\varepsilon^{2\gamma-2})$, $K_4 = \mathcal{O}(d^{2\gamma}\varepsilon^{2\gamma-2})$.

Method of Analysis

- Define a *linearly interpolated* version of $\{W^k\}_{k\in\mathbb{N}_+}$:

$$\mathrm{d}\hat{W}(t) = b(\hat{W})\mathrm{d}t + \varepsilon\sigma dB(t) + \varepsilon\mathrm{d}L^\alpha(t),$$

$\hat{W}\equiv\{\hat{W}(t)\}_{t\geq 0}$ denotes the whole process and

$$b(\hat{W}) \triangleq -\sum_{k=0}^{\infty}\nabla f(\hat{W}(k\eta))\mathbb{I}_{[k\eta,(k+1)\eta)}(t).$$

$\mathbb{I}$: the indicator function. We have $\hat{W}(k\eta) = W^k$ $\forall k\in\mathbb{N}_+$.

- Using Girsanov-like change of measures to upper-bound KL divergence between $\{W(s)\}_{s\in[0,t]}\sim\mu_t$ and $\{\hat{W}(s)\}_{s\in[0,t]}\sim\hat{\mu}_t$:

Theorem 1 *The following inequality holds:*

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq 2\delta^2.$$

- Upper-bound the total variation:

$$\|\mu_{K\eta} - \hat{\mu}_{K\eta}\|_{TV} \leq \left(\frac{1}{2}\mathrm{KL}(\hat{\mu}_{K\eta}, \mu_{K\eta})\right)^{\frac{1}{2}}.$$

$\|\mu-\nu\|_{TV} \triangleq 2\sup_{A\in\mathcal{B}(\Omega)}|\mu(A)-\nu(A)|$, $\mathcal{B}(\Omega)$: Borel set of Ω.

- By an optimal coupling argument: $\|\mu_{K\eta} - \hat{\mu}_{K\eta}\|_{TV} \geq$

$$\mathbb{P}_{\mathbf{M}}[(W(\eta),\dots,W(K\eta)) \neq (\hat{W}(\eta),\dots,\hat{W}(K\eta))]$$

$\mathbf{M}$: optimal coupling of $\{W(s)\}_{s\in[0,K\eta]}$ and $\{\hat{W}(s)\}_{s\in[0,K\eta]}$

- Relate the first exit time of discrete system to cont. system.

Main Result

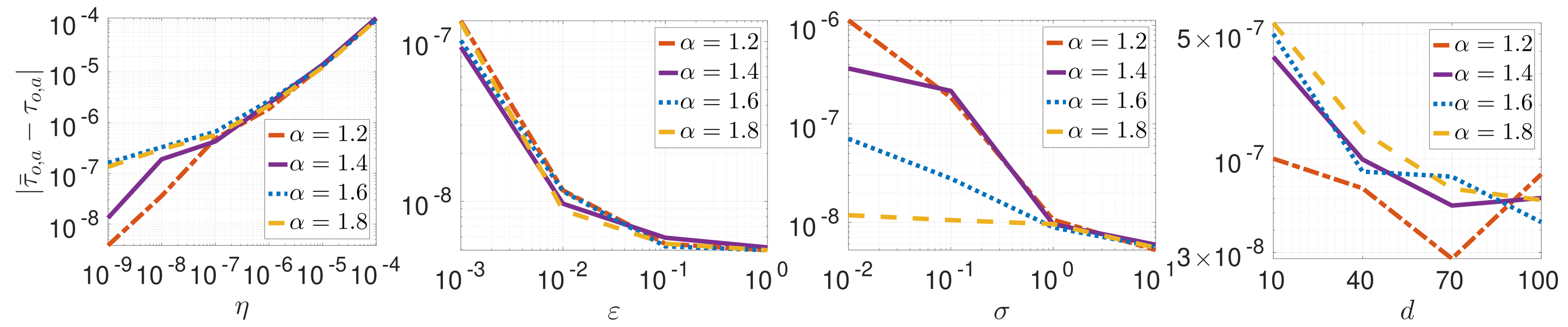
Theorem 2 *The following inequalities hold:*

$$\mathbb{P}[\tau_{-\psi,a}(\varepsilon) > K\eta] - C_{K,\eta,\varepsilon,d,\psi} - \delta$$
$$\leq \mathbb{P}[\bar{\tau}_{0,a}(\varepsilon) > K] \leq$$
$$\mathbb{P}[\tau_{\psi,a}(\varepsilon) > K\eta] + C_{K,\eta,\varepsilon,d,\psi} + \delta$$

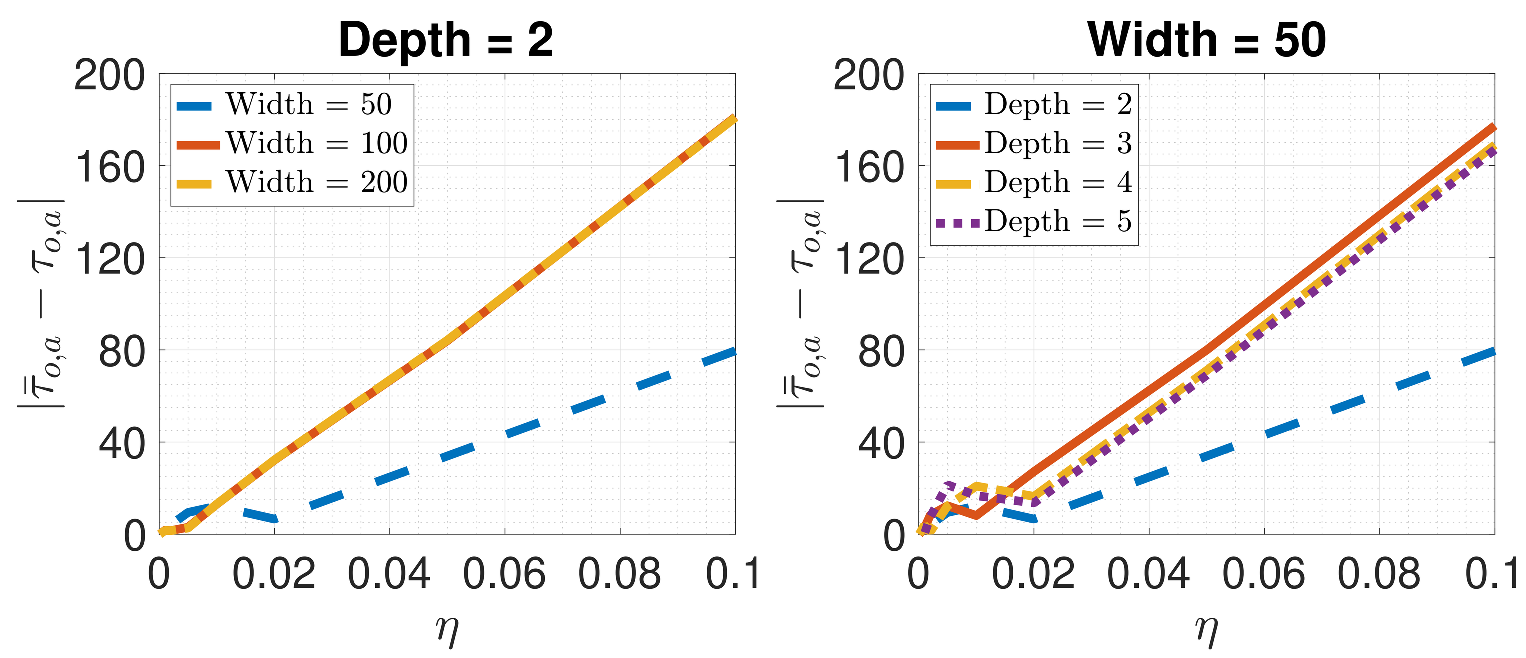
- First exit times for discrete process and cont. process become more similar with the decrease of η, d or ε.
- Theorem 2 enables the use of the metastability results for Lévy-driven SDEs for their discretized counterpart.

Numerical Illustration

- Results of the synthetic experiments.



- Results of the neural network experiments.



- Our experimental results are in accordance with the theoretical result shown in Theorem 2.

References

[1] U. Şimşekli, L. Sagun, and M. Gürbüzbalaban. "A tail-index analysis of stochastic gradient noise in deep neural networks." ICML 2019.

[2] P. Imkeller and I. Pavlyukevich. "First exit times of sdes driven by stable Lévy processes." Stochastic Processes and their Applications 2006.