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: LTCI, Télécom Paris, Institut Polytechnique de Paris, France

2: Department of Statistics, University of Oxford, UK

3: Dept. of Management Science and Information Systems, Rutgers Business School, NJ, USA

INTRODUCTION

- 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:

$$W^{k+1} = W^k - \eta\nabla\tilde{f}_k(W^k), \quad k \geq 0,$$

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

- Under the Gaussian noise assumption, consider:

$$\mathrm{d}W(t) = -\nabla f(W(t))\mathrm{d}t + \sqrt{\eta}\sigma\mathrm{d}B(t)$$

$B(t)$: standard Brownian motion, σ: noise variance.

- Under α-stable noise model (Şimşekli et al., 2019):

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

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

THIS WORK

- **In this work**, consider

$$\mathrm{d}W(t) = -\nabla f(W(t-))\mathrm{d}t + \varepsilon\sigma\mathrm{d}B(t) + \varepsilon\mathrm{d}L^\alpha(t) \quad (1)$$

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

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

- Define the *first exit times*, respectively for $W(t)$ and W^k as follows:

$$\tau_{\psi,a}(\varepsilon) \triangleq \inf\{t \geq 0 : \|W(t) - \bar{w}\| \notin [0, a+\psi]\},$$

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

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

- **Goal:** Derive explicit conditions for the step-size such that the probability to exit a given neighborhood of the local optimum at a fixed time t of the discretization process approximates that of the continuous process.

ASSUMPTIONS

Assumption: The SDE (1) admits a unique strong solution.

Assumption: The process $\phi_t \triangleq -\frac{b(W)+\nabla f(W(t))}{\varepsilon\sigma}$ satisfies $\mathbb{E}\exp\left(\frac{1}{2}\int_0^T \phi_t^2\mathrm{d}t\right) < \infty$.

Assumption: The gradient of f is γ-Hölder continuous with $\frac{1}{2} < \gamma < \min\{\frac{1}{\sqrt{2}}, \frac{\alpha}{2}\}$:

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

Assumption: The gradient of f satisfies the following assumption: $\|\nabla f(0)\| \leq B$.

Assumption: For some $m > 0$ and $b \geq 0$, f is (m, b, γ)-dissipative: $\langle x, \nabla f(x)\rangle \geq m\|x\|^{1+\gamma} - b, \forall x \in \mathbb{R}^d$.

Assumption: For a given $\delta > 0$, $t = K\eta$, and for some $C > 0$, the step-size satisfies the following condition: $0 < \eta \leq \min\Big\{$

$$1, \frac{m}{M^2}, \left(\frac{\delta^2}{2K_1t^2}\right)^{\frac{1}{\gamma^2+2\gamma-1}}, \left(\frac{\delta^2}{2K_2t^2}\right)^{\frac{1}{2\gamma}}, \left(\frac{\delta^2}{2K_3t^2}\right)^{\frac{\alpha}{2\gamma}}, \left(\frac{\delta^2}{2K_4t^2}\right)^{\frac{1}{\gamma}}\Big\},$$

where ε is as in (2), and $K_1 = \mathcal{O}(d\varepsilon^{2\gamma^2-2})$, $K_2 = \mathcal{O}(\varepsilon^{-2})$, $K_3 = \mathcal{O}(d^{2\gamma}\varepsilon^{2\gamma-2})$, $K_4 = \mathcal{O}(d^{2\gamma}\varepsilon^{2\gamma-2})$.

METHOD OF ANALYSIS

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

$$\mathrm{d}\hat{W}(t) = b(\hat{W})\mathrm{d}t + \varepsilon\sigma\mathrm{d}B(t) + \varepsilon\mathrm{d}L^\alpha(t), \quad (3)$$

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

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

Here, $\mathbb{I}$ denotes the indicator function, i.e. $\mathbb{I}_S(x) = 1$ if $x \in S$ and $\mathbb{I}_S(x) = 0$ if $x \notin S$. We have $\hat{W}(k\eta) = W^k$ for all $k \in \mathbb{N}_+$.

- Develop a Girsanov-like change of measures to express the Kullback-Leibler (KL) divergence between μ_t and $\hat{\mu}_t$:

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) \triangleq \int \log\frac{\mathrm{d}\hat{\mu}_t}{\mathrm{d}\mu_t}\mathrm{d}\hat{\mu}_t,$$

where $\mu_t \sim \{W(s)\}_{s\in[0,t]}$, $\hat{\mu}_t \sim \{\hat{W}(s)\}_{s\in[0,t]}$, and $\mathrm{d}\mu_t/\mathrm{d}\hat{\mu}_t$ is the Radon-Nikodym derivative of μ_t with respect to $\hat{\mu}_t$.

THEORETICAL RESULTS

Theorem 1 *The following inequality holds:*

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

Theorem 2 *The following inequalities hold:*

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

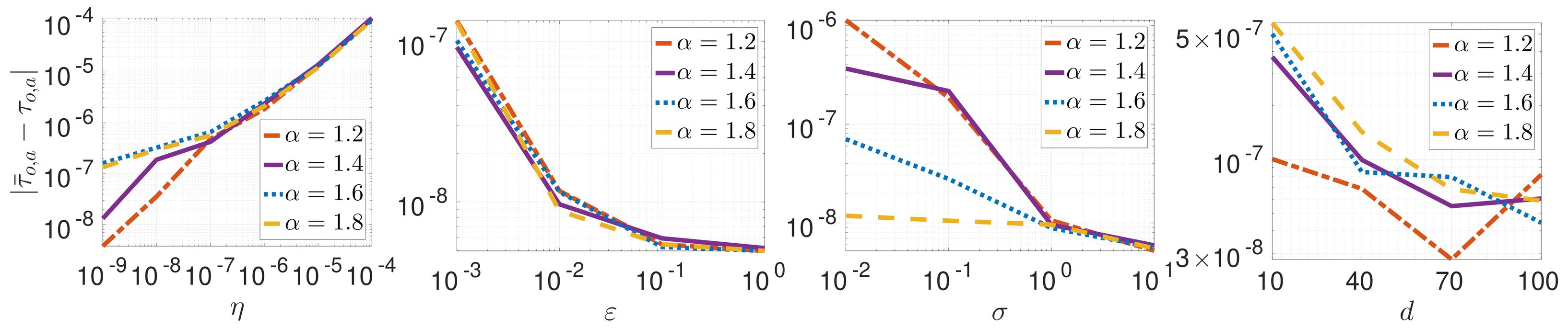
$$\mathbb{P}[\bar{\tau}_{0,a}(\varepsilon) > K] \leq \mathbb{P}[\tau_{\psi,a}(\varepsilon) > K\eta] + C_{K,\eta,\varepsilon,d,\psi} + \delta$$

where $C_{K,\eta,\varepsilon,d,\psi}$ is constant.

Remark. Theorem 2 enables the use of the metastability results for Lévy-driven SDEs for their discretized counterpart, which is our most important contribution.

NUMERICAL ILLUSTRATION

- Results of the synthetic experiments.



- Results of the neural network experiments.