

# Non-Asymptotic Analysis of Fractional Langevin Monte Carlo for Non-Convex Optimization

**Thanh Huy Nguyen, Umut Şimşekli, Gaël Richard**

LTCI, Télécom Paris, Institut Polytechnique de Paris, France



- **Non-convex optimization problem:**  $\min f(x)$

- **Non-convex optimization problem:**  $\min f(x)$
- Fractional Langevin Algorithm (FLA) (Simsekli, 2017):

$$W^{k+1} = W^k - \eta c_\alpha \nabla f(W^k) + (\eta/\beta)^{1/\alpha} \Delta L_{k+1}^\alpha$$

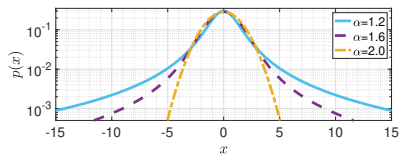
- $\{\Delta L_k^\alpha\}_{k \in \mathbb{N}_+}$ :  $\alpha$ -stable random variables
- $\alpha \in (1, 2]$ : the characteristic index,  $c_\alpha$ : a known constant

- **Non-convex optimization problem:**  $\min f(x)$
- Fractional Langevin Algorithm (FLA) (Simsekli, 2017):

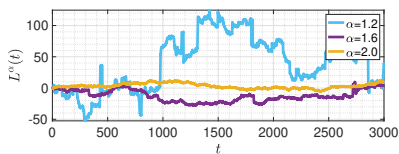
$$W^{k+1} = W^k - \eta c_\alpha \nabla f(W^k) + (\eta/\beta)^{1/\alpha} \Delta L_{k+1}^\alpha$$

- $\{\Delta L_k^\alpha\}_{k \in \mathbb{N}_+}$ :  $\alpha$ -stable random variables
- $\alpha \in (1, 2]$ : the characteristic index,  $c_\alpha$ : a known constant

- **$\alpha$ -stable Distribution**



- **$\alpha$ -stable Lévy Motion:**



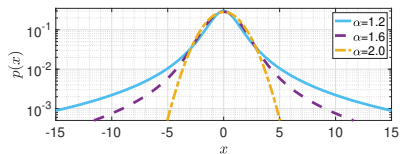
- Generalizes **Stochastic Gradient Langevin Dynamics** ( $\alpha = 2$ ) (Welling and Teh, 2011)
- Strong links with **SGD for Deep Neural Networks** (Simsekli et al. 2019)

- **Non-convex optimization problem:**  $\min f(x)$
- Fractional Langevin Algorithm (FLA) (Simsekli, 2017):

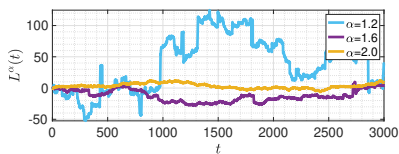
$$W^{k+1} = W^k - \eta c_\alpha \nabla f(W^k) + (\eta/\beta)^{1/\alpha} \Delta L_{k+1}^\alpha$$

- $\{\Delta L_k^\alpha\}_{k \in \mathbb{N}_+}$ :  $\alpha$ -stable random variables
- $\alpha \in (1, 2]$ : the characteristic index,  $c_\alpha$ : a known constant

- **$\alpha$ -stable Distribution**



- **$\alpha$ -stable Lévy Motion:**



- Generalizes **Stochastic Gradient Langevin Dynamics** ( $\alpha = 2$ ) (Welling and Teh, 2011)
- Strong links with **SGD for Deep Neural Networks** (Simsekli et al. 2019)
- **Our Goal:** Analyze  $\mathbb{E}[f(W^k) - f^*]$ , where  $f^* \triangleq \min f(x)$

- Define three stochastic processes:

$$dX_1(t) = -c_\alpha \nabla f(X_1(t-)) dt + \beta^{-1/\alpha} dL^\alpha(t),$$

$$dX_2(t) = -c_\alpha \sum_{k=0}^{\infty} \nabla f(X_2(j\eta)) \mathbb{I}_{[j\eta, (j+1)\eta]}(t) dt + \beta^{-1/\alpha} dL^\alpha(t),$$

$$dX_3(t) = -\mathcal{D}_{x_i}^{\alpha-2} \left( \phi(X_3(t-)) \frac{\partial f(X_3(t-))}{\partial x_i} \right) / \phi(X_3(t-)) dt + \beta^{-1/\alpha} dL^\alpha(t).$$

- Define three stochastic processes:

$$dX_1(t) = -c_\alpha \nabla f(X_1(t-)) dt + \beta^{-1/\alpha} dL^\alpha(t),$$

$$dX_2(t) = -c_\alpha \sum_{k=0}^{\infty} \nabla f(X_2(j\eta)) \mathbb{I}_{[j\eta, (j+1)\eta]}(t) dt + \beta^{-1/\alpha} dL^\alpha(t),$$

$$dX_3(t) = -\mathcal{D}_{x_i}^{\alpha-2} \left( \phi(X_3(t-)) \frac{\partial f(X_3(t-))}{\partial x_i} \right) / \phi(X_3(t-)) dt + \beta^{-1/\alpha} dL^\alpha(t).$$

- $\mathcal{D}$ : Riesz fractional (directional) derivative
- $X_1$  is the continuous-time limit of the FLA algorithm
- $X_2$  is a linearly interpolated version of  $W^k$ :  $X_2(k\eta) = W^k, \forall k \in \mathbb{N}_+$
- $X_3$  admits  $\pi \propto \exp(-\beta f(x)) dx$  as its unique invariant distribution

# Method of Analysis

- Define three stochastic processes:

$$dX_1(t) = -c_\alpha \nabla f(X_1(t-)) dt + \beta^{-1/\alpha} dL^\alpha(t),$$

$$dX_2(t) = -c_\alpha \sum_{k=0}^{\infty} \nabla f(X_2(j\eta)) \mathbb{I}_{[j\eta, (j+1)\eta]}(t) dt + \beta^{-1/\alpha} dL^\alpha(t),$$

$$dX_3(t) = -\mathcal{D}_{x_i}^{\alpha-2} \left( \phi(X_3(t-)) \frac{\partial f(X_3(t-))}{\partial x_i} \right) / \phi(X_3(t-)) dt + \beta^{-1/\alpha} dL^\alpha(t).$$

- $\mathcal{D}$ : Riesz fractional (directional) derivative
- $X_1$  is the continuous-time limit of the FLA algorithm
- $X_2$  is a linearly interpolated version of  $W^k$ :  $X_2(k\eta) = W^k, \forall k \in \mathbb{N}_+$
- $X_3$  admits  $\pi \propto \exp(-\beta f(x)) dx$  as its unique invariant distribution
- Decompose the error  $\mathbb{E}f(W^k) - f^*$  as:

$$\begin{aligned} & [\mathbb{E}f(X_2(k\eta)) - \mathbb{E}f(X_1(k\eta))] + [\mathbb{E}f(X_1(k\eta)) - \mathbb{E}f(X_3(k\eta))] \\ & + [\mathbb{E}f(X_3(k\eta)) - \mathbb{E}f(\hat{W})] + [\mathbb{E}f(\hat{W}) - f^*] \end{aligned}$$

- $\hat{W} \sim \pi \propto \exp(-\beta f(x)) dx$
- Relate these terms to Wasserstein distance between processes



# Main Result

Main assumptions:

1) Hölder continuous gradients:  $c_\alpha \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|^\gamma$

2) Dissipativity:  $c_\alpha \langle x, \nabla f(x) \rangle \geq m \|x\|^{1+\gamma} - b$

# Main Result

Main assumptions:

- 1) Hölder continuous gradients:  $c_\alpha \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|^\gamma$
- 2) Dissipativity:  $c_\alpha \langle x, \nabla f(x) \rangle \geq m \|x\|^{1+\gamma} - b$

## Theorem

For  $0 < \eta < m/M^2$ , there exists  $C > 0$  such that:

$$\begin{aligned} \mathbb{E}[f(W^k)] - f^* \leq & C \left\{ k^{1+\max\{\frac{1}{q}, \gamma + \frac{\gamma}{q}\}} \eta^{\frac{1}{q}} + \frac{k^{1+\max\{\frac{1}{q}, \gamma + \frac{\gamma}{q}\}} \eta^{\frac{1}{q} + \frac{\gamma}{\alpha q}} d}{\beta^{\frac{(q-1)\gamma}{\alpha q}}} \right. \\ & \left. + \frac{\beta b + d}{m} \exp\left(-\frac{\lambda_* k \eta}{\beta}\right) \right\} + \frac{M c_\alpha^{-1}}{\beta^{\gamma+1} (1+\gamma)} \\ & + \frac{1}{\beta} \log \frac{(2e(b + \frac{d}{\beta}))^{\frac{d}{2}} \Gamma(\frac{d}{2} + 1) \beta^d}{(dm)^{\frac{d}{2}}}. \end{aligned}$$

# Main Result

Main assumptions:

- 1) Hölder continuous gradients:  $c_\alpha \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|^\gamma$
- 2) Dissipativity:  $c_\alpha \langle x, \nabla f(x) \rangle \geq m \|x\|^{1+\gamma} - b$

## Theorem

For  $0 < \eta < m/M^2$ , there exists  $C > 0$  such that:

$$\begin{aligned} \mathbb{E}[f(W^k)] - f^* \leq & C \left\{ k^{1+\max\{\frac{1}{q}, \gamma + \frac{\gamma}{q}\}} \eta^{\frac{1}{q}} + \frac{k^{1+\max\{\frac{1}{q}, \gamma + \frac{\gamma}{q}\}} \eta^{\frac{1}{q} + \frac{\gamma}{\alpha q}} d}{\beta^{\frac{(q-1)\gamma}{\alpha q}}} \right. \\ & \left. + \frac{\beta b + d}{m} \exp\left(-\frac{\lambda_* k \eta}{\beta}\right) \right\} + \frac{M c_\alpha^{-1}}{\beta^{\gamma+1}(1+\gamma)} \\ & + \frac{1}{\beta} \log \frac{(2e(b + \frac{d}{\beta}))^{\frac{d}{2}} \Gamma(\frac{d}{2} + 1) \beta^d}{(dm)^{\frac{d}{2}}}. \end{aligned}$$

- Worse dependency on  $\eta$  and  $k$  than the case  $\alpha = 2$
- Requires smaller  $\eta$

# Additional Results

- **Posterior Sampling:** sampling from  $\pi \propto \exp(-\beta f(x))dx$
- **Stochastic Gradients:**

$$f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f^{(i)}(x)$$


$$\nabla f \approx \nabla f_k(x) \triangleq \left( \sum_{i \in \Omega_k} \nabla f^{(i)}(x) \right) / n_s$$

# Additional Results


- **Posterior Sampling:** sampling from  $\pi \propto \exp(-\beta f(x))dx$
- **Stochastic Gradients:**

$$f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f^{(i)}(x)$$

$$\nabla f \approx \nabla f_k(x) \triangleq \left( \sum_{i \in \Omega_k} \nabla f^{(i)}(x) \right) / n_s$$
- For more information/questions, come to our poster #198!

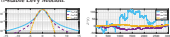


**NON-ASYMPTOTIC ANALYSIS OF FRACTIONAL LANGEVIN MONTE CARLO FOR NON-CONVEX OPTIMIZATION**  
 Thanh Huy Nguyen<sup>1</sup>, Umut Şimşekli<sup>2</sup>, Gaël Richard<sup>3</sup>  
1. LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France.  
 2. supported by the French National Research Agency (ANR) as a part of the FEDIMATREX project (ANR-18-CET2-0014)



### INTRODUCTION

- Non-convex optimization problem:  $\min_x f(x)$
- Fractional Langevin Algorithm (FLA) [1]
 
$$W^{t+\Delta t} = W^t - \eta \nabla f(W^t) + (\eta/\beta)^{1-\alpha} \Delta L_{\alpha, \beta}^t$$

$-\Delta L_{\alpha, \beta}^t$ :  $\alpha$ -stable random variables  
 $-\alpha \in (1, 2]$ : the characteristic index,  $c_{\alpha, \beta}$  a known constant
- $\alpha$ -stable Lévy Motion:
 
- Generalizes Stoch. Grad. Langevin Dynamics [2] ( $\alpha = 2$ )
- Strong links with SGD for Deep Neural Networks [3]
- Has better empirical generalization properties
- Our Goal: Analyze the expected error:  $\mathbb{E}[f(W^t)] - f^*$ , where  $f^* \triangleq \min_x f(x)$

### METHOD OF ANALYSIS

- Define these stochastic processes:
 
$$dX_t(t) = h_t(X_t(t-\cdot), \alpha) dt + \beta^{-1/\alpha} dL_{\alpha, \beta}^t(t),$$

$$dX_t^0(t) = h_0(X_t^0(t-\cdot), \alpha) dt + \beta^{-1/\alpha} dL_{\alpha, \beta}^t(t),$$

$$dX_t^1(t) = h_1(X_t^1(t-\cdot), \alpha) dt + \beta^{-1/\alpha} dL_{\alpha, \beta}^t(t),$$
- with
 
$$h_t(X_t, \alpha) \triangleq -c_{\alpha, \beta} \nabla f(X_t)$$

$$h_0(X_t, \alpha) \triangleq -c_{\alpha, \beta} \sum_{i=1}^n \nabla f_i(X_t) \otimes (\beta_{i, \alpha, \beta})_{i=1, \dots, n} \otimes \mathbb{1}_{\{i \neq t\}}$$

$$h_1(X_t, \alpha) \triangleq -c_{\alpha, \beta} \sum_{i=1}^n \nabla f_i(X_t) \otimes (\beta_{i, \alpha, \beta})_{i=1, \dots, n} \otimes \mathbb{1}_{\{i=t\}}$$
- $D_t$ : Riesz fractional (directional) derivative  
 $-X_t^0(t) = W^t$  for all  $k \in \mathbb{N}$  (i.e. linear interpolation)  
 $-X_t^k$  targets  $\pi \propto \exp(-\beta f)$  (i.e.  $f^*$ )  
 • Decompose the error  $\mathbb{E}[f(W^t)] - f^*$  as:
 
$$\mathbb{E}[f(X_t^k(t)) - f^*] = \mathbb{E}[f(X_t^k(t)) - \mathbb{E}[f(X_t^k(t))]] + \mathbb{E}[f(X_t^k(t)) - f^*]$$

$$= \mathbb{E}[f(X_t^k(t)) - \mathbb{E}[f(X_t^k(t))]] + \mathbb{E}[f(X_t^k(t)) - f^*]$$
- $W^t \sim \pi$ ,  $\pi \propto \exp(-\beta f)$  (i.e.  $f^*$ )  
 • Relate these terms to Wasserstein distance between processes

### ASSUMPTIONS & INTERMEDIATE RESULTS

**Assumption:** There exist constants  $M > 0, 0 \leq \gamma < 1$ :

$$c_{\alpha, \beta} \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|^\gamma, \quad x, y \in \mathbb{R}^d.$$

**Assumption:** For some  $m > 0$  and  $b \geq 0$

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

**Assumption:**  $\exists \eta, \rho_1, \rho_2, \rho_3 > 0$  such that:  $\eta < c_{\alpha, \beta} \gamma < 1, \eta \rho_1 < 1, (\eta - 1) \rho_2 < 1$  and  $1/\rho_1 + 1/\rho_2 = 1/\rho_3 = 1$ .

**Assumption:** 1) For some  $\gamma \in [0, 1], k_0 \geq 0, K_1, K_2 > 0$

$$\frac{\|h_t(x) - h_t(y)\|}{\|x - y\|} \leq \begin{cases} K_1 \|x - y\|^\gamma, & \|x - y\| \leq k_0, \\ -K_2 \|x - y\|, & \|x - y\| \geq k_0. \end{cases}$$

2) For any coupling  $P_t$  of  $X_t(t)$  and  $W^t \sim \pi, \mathbb{C}(t) > 0$

$$\int \|X_t(t) - W^t\| dP_t < C, \quad t > 0, \gamma \in (0, \alpha).$$

**Assumption:** There exists  $L > 0$  such that  $L < c_{\alpha, \beta}$  and

$$\sup_{x \in \mathbb{R}^d} \|\nabla f(x) + h_t(x, \alpha)\| \leq L.$$

**Lemma 1** Let  $V \sim \mu$  and  $W \sim \nu$  and let  $\rho \in C^0(\mathbb{R}^d, \mathbb{R})$ . Assume that for some  $c_1 > 0, c_2 \geq 0$  and  $0 \leq \gamma < 1$ ,

$$\|\nabla \rho(x)\| \leq c_1 \|x\|^\gamma + c_2, \quad \forall x \in \mathbb{R}^d,$$

and  $\max\left\{\mathbb{E}[\|W\|^{1+\gamma}], \mathbb{E}[\|V\|^{1+\gamma}]\right\} < \infty$ . Then we have:

$$\|\mathbb{E}[\rho(V)] - \mathbb{E}[\rho(W)]\| \leq C \|W\|^{c_1, \alpha}, \quad \text{for some } C > 0.$$

**Lemma 2** We have the following identity:  $W_t(\mu, \nu, \beta) =$

$$\int \left( \mathbb{E} \left[ \int_0^1 \lambda \|\Delta X_{\lambda, \beta}(s)\|^{1-\alpha} \langle \Delta X_{\lambda, \beta}(s), \Delta h_{\lambda, \beta}(s-\cdot) \rangle ds \right] \right)^{1/\alpha},$$

where the infimum is taken over the couplings and

$$\Delta X_{\lambda, \beta}(s) \triangleq X_{\lambda, \beta}(s) - X_{\lambda, \beta}(s-),$$

$$\Delta h_{\lambda, \beta}(s-\cdot) \triangleq h_{\lambda, \beta}(X_{\lambda, \beta}(s-\cdot), \alpha) - h_{\lambda, \beta}(X_{\lambda, \beta}(s-\cdot), \alpha).$$

### MAIN RESULT

**Theorem 1** For  $0 < \alpha \leq \alpha_0/M^2$ , there exists  $C > 0$  such that:

$$\mathbb{E}[f(W^t)] - f^* \leq C \left\{ k^{1+\alpha} + (\beta^{-1/\alpha})^{\alpha+1} \eta + \beta^{-1/\alpha} \left( \frac{k^{1+\alpha} + (\beta^{-1/\alpha})^{\alpha+1} \eta}{\beta^{1-\alpha}} + \frac{M c_{\alpha, \beta}^{-1}}{\beta^{1-\alpha} (1+\gamma)} \right) \right.$$

$$\left. + \frac{\beta b + d}{m} \exp(-\frac{\lambda k}{\beta}) \right\} + \frac{M c_{\alpha, \beta}^{-1}}{\beta^{1-\alpha} (1+\gamma)}$$

$$+ \frac{1}{\beta} \log \frac{(2\beta b + \frac{d}{\beta}) \beta^{1-\alpha} (1+\gamma)}{(dm)^{1-\alpha}}$$

– Worse dependency on  $\alpha$  and  $k$  than the case  $\alpha = 2$   
 – Requires smaller  $\eta$

### ADDITIONAL RESULTS

- **Posterior Sampling:** If our aim is only to draw samples from the distribution  $\pi$ , we have the result:
 

**Corollary 1** For  $0 < \alpha \leq \alpha_0/M^2$ , the following bound holds:

$$W_t(\mu, \pi, \beta) \leq C \left( k^{1+\alpha} + (\beta^{-1/\alpha})^{\alpha+1} \eta + \beta^{-1/\alpha} \left( \frac{k^{1+\alpha} + (\beta^{-1/\alpha})^{\alpha+1} \eta}{\beta^{1-\alpha}} + \frac{M c_{\alpha, \beta}^{-1}}{\beta^{1-\alpha} (1+\gamma)} \right) + \beta c_{\alpha, \beta}^{-1} \right)^{1/\alpha}$$
- **Stochastic Gradients:** Assume:  $f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f^{(i)}(x)$   
 – Approximate  $\nabla f$  by:  $\nabla f_k(x) \triangleq \left( \sum_{i \in \Omega_k} \nabla f^{(i)}(x) \right) / n_k$   
 $\Omega_k$  is a random subset of  $\{1, \dots, n\}$  with  $|\Omega_k| = n_k \ll n$ .

**Theorem 2**  $\exists$  there exists  $\delta \in [0, 1]$  such that, for any  $k$ ,

$$\mathbb{E}_{\Omega_k} \|\nabla f_k(W^t) - \nabla f_k(x)\| \leq C \beta^{\delta} M^{\alpha} \eta^{\alpha} \|x\|^{1-\delta}, \quad x \in \mathbb{R}^d,$$

(here we have the following bound:

$$W_t^{\alpha}(\mu, \nu, \beta) \leq C(1 + \delta) k^{\alpha} \eta^{\alpha} + k^{\alpha} \beta^{1-\alpha} \gamma^{-\alpha} (\beta^{-1/\alpha})^{\alpha}.$$

### REFERENCES

[1] Spigler, D. Theoretical Langevin Monte Carlo: Righting Lefty Devotees. *International Conference on Machine Learning*, 2020.

[2] Hopkins, M., Radhakrishnan, A., Sridharan, M. Non-convex Sampling via Fractional Langevin Dynamics: a non-asymptotic analysis. *ICML 2021*.

[3] Spigler, D., Sridharan, M., Radhakrishnan, M. Theoretical Analysis of Fractional Gradient Descent on Deep Neural Networks. *ICML 2021*.