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

Introduction

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

Introduction

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

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

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

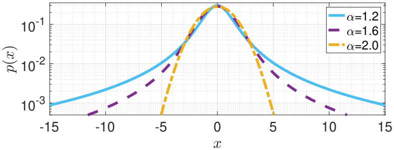
Introduction

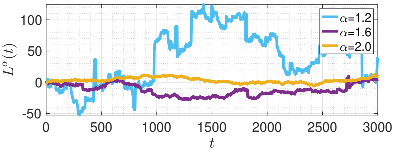
- **Non-convex optimization problem**: $\min f(x)$
- Fractional Langevin Algorithm (FLA) (Simsekli, 2017):

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

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

- **α-stable Distribution** **α-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)

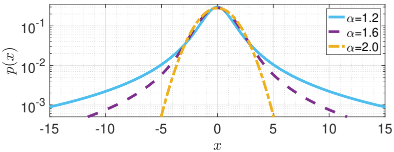
Introduction

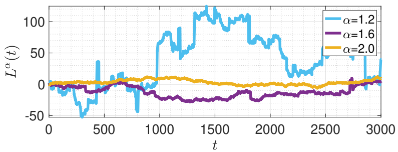
- **Non-convex optimization problem**: $\min f(x)$
- Fractional Langevin Algorithm (FLA) (Simsekli, 2017):

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

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

- **α-stable Distribution** **α-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 $\mathbb{E}[f(W^k) - f^\star]$, where $f^\star \triangleq \min f(x)$

Method of Analysis

- Define three stochastic processes:

$$\mathrm{d}X_1(t) = -c_\alpha \nabla f(X_1(t-))\mathrm{d}t + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t),$$

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

$$\mathrm{d}X_3(t) = -\mathcal{D}_{x_i}^{\alpha-2}\left(\phi(X_3(t-))\frac{\partial f(X_3(t-))}{\partial x_i}\right)\Big/\phi(X_3(t-))\mathrm{d}t \; + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t).$$

Method of Analysis

- Define three stochastic processes:

$$\mathrm{d}X_1(t) = -c_\alpha \nabla f(X_1(t-))\mathrm{d}t + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t),$$

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

$$\mathrm{d}X_3(t) = -\mathcal{D}_{x_i}^{\alpha-2}\Big(\phi(X_3(t-))\frac{\partial f(X_3(t-))}{\partial x_i}\Big)\Big)/\phi(X_3(t-))\mathrm{d}t \;+\beta^{-1/\alpha}\mathrm{d}L^\alpha(t).$$

– $\mathcal{D}$: 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 $\pi \propto \exp(-\beta f(x))\mathrm{d}x$ as its unique invariant distribution

Method of Analysis

- Define three stochastic processes:

$$\mathrm{d}X_1(t) = -c_\alpha \nabla f(X_1(t-))\mathrm{d}t + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t),$$

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

$$\mathrm{d}X_3(t) = -\mathcal{D}_{x_i}^{\alpha-2}\left(\phi(X_3(t-))\frac{\partial f(X_3(t-))}{\partial x_i}\right)\Big/\phi(X_3(t-))\mathrm{d}t \; + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t).$$

- $\mathcal{D}$: 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 $\pi \propto \exp(-\beta f(x))\mathrm{d}x$ as its unique invariant distribution

- Decompose the error $\mathbb{E}f(W^k) - f^*$ as:

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

- $\hat{W} \sim \pi \propto \exp(-\beta f(x))\mathrm{d}x$
- 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:*

$$\mathbb{E}[f(W^k)] - f^* \leq C\Bigg\{ k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}} + \frac{k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}} d}{\beta^{\frac{(q-1)\gamma}{\alpha q}}} + \frac{\beta b + d}{m}\exp(-\frac{\lambda_* k\eta}{\beta})\Bigg\} + \frac{Mc_\alpha^{-1}}{\beta^{\gamma+1}(1+\gamma)} + \frac{1}{\beta}\log\frac{(2e(b+\frac{d}{\beta}))^{\frac{d}{2}}\Gamma(\frac{d}{2}+1)\beta^d}{(dm)^{\frac{d}{2}}}.$$

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:

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

– Worse dependency on η and k than the case $\alpha = 2$

– Requires smaller η

Additional Results

- **Posterior Sampling:** sampling from $\pi \propto \exp(-\beta f(x))\mathrm{d}x$
- **Stochastic Gradients:**

$$f(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f^{(i)}(x)$$
$$\nabla f \approx \nabla f_k(x) \triangleq \Big(\sum_{i\in\Omega_k} \nabla f^{(i)}(x)\Big)/n_s$$

Additional Results

- **Posterior Sampling:** sampling from $\pi \propto \exp(-\beta f(x))\mathrm{d}x$
- **Stochastic Gradients:**

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

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

- For more information/questions, come to our poster #**198**!