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):

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

- $\{\Delta L^{\alpha}_k\}_{k \in \mathbb{N}^+}$: $\alpha$-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):**

\[
W_{k+1} = W_k - \eta c_\alpha \nabla f(W_k) + \left(\frac{\eta}{\beta}\right)^{1/\alpha} \Delta L_{k+1}^\alpha
\]

- \( \{\Delta L_k^\alpha\}_{k \in \mathbb{N}_+} \): \( \alpha \)-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):
\[
W_{k+1} = W_k - \eta c_\alpha \nabla f(W_k) + \left( \frac{\eta}{\beta} \right)^{1/\alpha} \Delta L^\alpha_{k+1}
\]

- \( \{\Delta L^\alpha_k\}_{k \in \mathbb{N}_+} \): \( \alpha \)-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:

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

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

\[ dX_3(t) = -D_{x_i}^{\alpha-2} \left( \phi(X_3(t-)) \frac{\partial f(X_3(t-))}{\partial x_i} \right) / \phi(X_3(t-))dt + \beta^{-1/\alpha}dL^\alpha(t). \]
Define three stochastic processes:

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

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

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

- \( D \): 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:

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

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

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

- \( D \): 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:

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

- \( \hat{W} \sim \pi \propto \exp(-\beta f(x))dx \)
- 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 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:

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

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

- 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**:  
  \[ f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(x) \]  
  \[ \nabla f \approx \nabla f_k(x) \triangleq \left( \sum_{i \in \Omega_k} \nabla f^{(i)}(x) \right) / n_s \]
Stochastic Gradients:

\[ f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x) \]

\[ \nabla f \approx \nabla f_k(x) = \sum_i k \eta \nabla f_i(x) / n \]

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