Global Non-convex Optimization with Discretized Diffusions

Murat A. Erdogdu
lmackey@microsoft.com

Lester Mackey
erdogdu@cs.toronto.edu

Ohad Shamir
ohad.shamir@weizmann.ac.il

From Optimization to Diffusions

Consider the unconstrained and possibly non-convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x).$$

- An example algorithm: **Langevin Gradient Descent**

$$X_{t+1} = X_t - \eta \nabla f(X_t) + \sqrt{2\eta} W_t,$$

where $\eta, \gamma > 0$ and $W_t \sim \mathcal{N}(0, I)$ independent of $X_t$ for $t \leq t^*$. This is the Euler discretization of the Langevin diffusion.

$$X_{t+1} - X_t \approx -\nabla f(X_t) + \sqrt{2\eta} B_t,$$

where $\eta \downarrow 0$.

- This algorithm is the Euler discretization of the Langevin diffusion.

$$\frac{dX}{dt} = -\nabla f(X_t) + \sqrt{2\eta} \frac{dB_t}{dt},$$

to obtain diffusion.

- This diffusion converges to Gibbs measure $X_t \sim p(x) \propto e^{-\frac{\gamma}{2}f(x)}$ concentrating around global minima. For small $\eta$, its discretization also concentrates around global minima, but current analysis requires $f$ to have **quadratic growth**.

Our focus is on general Itô diffusions $\frac{dX}{dt} = b(X_t) + \sigma(X_t) \, dB_t$ with $X_0 = x$, and their Euler discretization

$$X_{n+1} = X_n + \eta b(X_n) + \sqrt{\sigma(X_n)} W_n,$$

which can optimize a rich class of non-convex functions.

Conditions for Global Convergence

**Condition 1 (Coefficient growth)**: The drift and the diffusion coefficients satisfy the following growth condition for $\forall x \in \mathbb{R}^d$

$$\Vert b(x) \Vert_2 \leq \frac{1}{2} (1 + \|x\|), \quad \Vert \sigma(x) \Vert_F \leq \frac{1}{2} (1 + \|x\|), \quad \text{and} \quad \|\sigma \sigma^T(x)\|_{op} \leq \frac{1}{2} (1 + \|x\|).$$

**Condition 2 (Dissipativity)**: For $\alpha, \beta > 0$, the diffusion satisfies

$$\|A(x)\|_2^2 \leq -\alpha \|x\|_2^2 + \beta \quad \text{for} \quad A(x) = (b(x), \nabla \sigma(x)) + \frac{1}{2} \sigma(x) \sigma^T(x), \nabla \sigma^2(x).$$

$A$ is the generator of the diffusion, e.g., $\|A\|_2^2 = 2\|b\|_2 + \|\sigma\|_F^2$.

**Condition 3 (Finite Steinn factors)**: The function $u_2(s)$ solves the Stein equation

$$f - p(f) = A u_2 \quad \text{with} \quad p(f) = \mathbb{E}_{X \sim p}\left[f(X)\right],$$

has $s$-th order derivative with polynomial growth for $i = 1, 2, 3, 4$, i.e.,

$$\|\nabla^i u_j(x)\|_2 \leq \zeta_i (1 + \|x\|), \quad \text{for} \quad i \in \{1, 2, 3, 4\} \quad \text{and} \quad x \in \mathbb{R}^d.$$

with $\max_{i=1,2,3,4} \zeta_i < \infty$.

Explicit Bounds on Integration Error

**Theorem:** Integration error of discretized diffusions

Let Conditions 1, 2, 3 hold. For a step size small enough

$$\left| \frac{1}{M} \sum_{m=1}^{M} E[f(X_m)] - p(f) \right| \leq \left( \frac{1}{\eta M} + c_2 \sigma + c_3 \rho \right) \left( \zeta_0 + E[\|X_0\|_2^2] \right),$$

where

$$c_1 = 6c_2, \quad c_2 = \frac{1}{2} \zeta_0 \lambda_1^2 + 2 \zeta_0 \lambda_2^2 + 6 c_2 (\lambda_1^2 + 25 \lambda_2^2) (\lambda_1 + \lambda_2),$$

$$c_3 = \frac{1}{2} (2 \zeta_1 \lambda_1^2 + \zeta_0 \lambda_2^2 + 2 \zeta_1 \lambda_1^2), \quad \zeta_0 = 2 + \frac{d}{2} + \frac{d}{2} + \left( \frac{1}{\eta M} \right).$$

Remark 1: Convergence rate is $O(\frac{1}{\sqrt{M}})$ to the invariant measure.

Remark 2: Stein factors $\zeta_i$ depend on $f$ and the chosen diffusion.

**Condition 4:** The diffusion has $L_2$-Wasserstein decay $g$

$$\inf_{(X, X')} E[\|X - X\|_2^2] \leq g(t) \forall x, y \in \mathbb{R}^d$$

for all $x, y \in \mathbb{R}^d$.

**Theorem:** Explicit bounds on the Stein factors

For an objective function $f$ satisfying

$$\|f(x) - f(y)\| \leq \alpha (1 + \|x\|_2 + \|y\|_2 + \|x - y\|_2), \quad \text{for all} \quad x, y \in \mathbb{R}^d,$$

and a diffusion satisfying Conditions 1, 2, 4, the Stein factors are given as

$$\zeta_1 = \tau_{\gamma} \zeta_2 + \zeta_3 \int_0^\tau_{\gamma} \sigma(d)$$

where $\tau_{\gamma}$ and $\zeta_i$ have explicit forms.

An Example with Sublinear Growth

minimize $f(x) = c \log(1 + \frac{1}{2} \|x\|_2^2)$ by sampling from $p(x) \propto e^{-\gamma f(x)}$.

- $f(x)$ is non-convex with sublinear growth, so Langevin algorithm is not guaranteed to work!

- Choose $\sigma(x) = \frac{1}{\sqrt{\gamma}} \sqrt{1 + \frac{1}{2} \|x\|_2^2}$.

- Diffusion has target invariant measure $p(x) \propto e^{-\gamma f(x)}$.

The diffusion is uniformly dissipative

$$2(b(x) - b(y), x - y) + \|\sigma(x) - \sigma(y)\|_2^2 \leq \alpha \|x - y\|_2^2,$$

for $\alpha = c - \frac{d}{2}$, hence it satisfies Conditions 1, 2, 4, and our theorems apply!

- In $d = 2$ dimension, for $c = 5$, step size $\eta = 0.1$, inverse temperature $\gamma = 1$, $X_0 = (91, 111)$.

Explicit Bounds on Optimization Error

**Proposition:** Sampling yields near-optima

Fix $C > 0$, $\theta \in (0, 1]$, and $x^* \in \arg \min f(x)$. For a diffusion with invariant measure $p$ and satisfying Condition 2, if $\log(p(x^*)) - \log(p(x)) \leq C \|x - x^*\|_2^2 \forall x$, then

$$-p(x^*) \log(p(x)) \leq \frac{\lambda_1}{2} \log \left( \frac{\lambda_1}{2 \lambda_2} \right) + \frac{\lambda_1}{2} \log \left( \frac{\lambda_1}{2 \lambda_2} \right).$$

If $p$ takes the generalized Gibbs form $p_n(x^*) \propto \exp(-\gamma (f(x) - f(x^*)))$, then

$$p_n(f(x^*)) \leq \sqrt{\frac{1}{1 + \log \left( \frac{\lambda_1}{2 \lambda_2} \right)}}.$$

**Corollary:** Optimization error of discretized diffusions

If the diffusion has the generalized Gibbs stationary density $p_n(x)$, then

$$\min_{m=1}^M E[f(X_m)] - f(x^*) \leq \left( c_1 \sigma + c_2 \rho \right) \left( \zeta_0 + E[\|X_0\|_2^2] \right)$$

+ $\frac{1}{\eta M} \left( \frac{1}{\eta M} \right)$.