Global Non-convex Optimization with Discretized Diffusions

Murat A. Erdogdu — erdogdu@cs.toronto.edu — University of Toronto, Vector Institute

Lester Mackey — lmackey@microsoft.com — Microsoft Research

Ohad Shamir — ohad.shamir@weizmann.ac.il — Weizmann Institute of Science

From Optimization to Diffusions

Consider the unconstrained and possibly non-convex optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x).$$

- An example algorithm: Langevin Gradient Descent

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

where $\eta, \gamma > 0$ and $W_t \sim \mathcal{N}(0, I)$ independent of X_τ for $\tau \le t$.

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

$$\frac{X_{t+\eta} - X_t}{\eta} = -\nabla f(X_t) + \sqrt{\tfrac{2}{\gamma}}\,\frac{B_{t+\eta} - B_t}{\eta}\ (\text{let } \eta \downarrow 0),$$
$$\frac{dX_t}{dt} = -\nabla f(X_t) + \sqrt{\tfrac{2}{\gamma}}\,\frac{dB_t}{dt},\ (\text{to obtain diffusion}).$$

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

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

$$X_{m+1} = X_m + \eta\, b(X_m) + \sqrt{\eta}\sigma(X_m)\, W_m,$$

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* $\forall x \in \mathbb{R}^d$

$$\|b(x)\|_2 \le \tfrac{\lambda_b}{4}(1 + \|x\|_2),\ \|\sigma(x)\|_{\mathrm{F}} \le \tfrac{\lambda_\sigma}{4}(1 + \|x\|_2),\ \textit{and}\ \|\sigma\sigma^\top(x)\|_{\mathrm{op}} \le \tfrac{\lambda_a}{4}(1 + \|x\|_2^2)$$

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

$$\mathcal{A}\|x\|_2^2 \le -\alpha\|x\|_2^2 + \beta \quad \textit{for} \quad \mathcal{A}g(x) \triangleq \langle b(x), \nabla g(x)\rangle + \tfrac{1}{2}\langle \sigma(x)\sigma(x)^\top, \nabla^2 g(x)\rangle.$$

$\mathcal{A}$ is the generator of the diffusion, e.g., $\mathcal{A}\|x\|_2^2 = 2\langle b(x), x\rangle + \|\sigma(x)\|_{\mathrm{F}}^2$.

Condition 3 (Finite Stein factors). *The function u_f solves the Stein equation*

$$f - p(f) = \mathcal{A}u_f, \quad \textit{with}\ \ p(f) = \mathbb{E}_{X\sim p}[f(X)]$$

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

$$\|\nabla^i u_f(x)\|_{\mathrm{op}} \le \zeta_i(1 + \|x\|_2) \ \textit{for}\ i \in \{1, 2, 3, 4\}\ \textit{and all}\ x \in \mathbb{R}^d.$$

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

Explicit Bounds on Integration Error

Theorem: Integration error of discretized diffusions

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

$$\left|\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[f(X_m)] - p(f)\right| \le \left(c_1\frac{1}{\eta M} + c_2\eta + c_3\eta^{1.5}\right)\left(\kappa + \mathbb{E}\left[\|X_0\|_2^6\right]\right)$$

where $c_1 = 6\zeta_1$, $\quad c_3 = \frac{1}{48}\left[\zeta_3\lambda_b^3 + 2\zeta_4\lambda_b^4 + 6\zeta_4(\lambda_b^4 + 25\lambda_\sigma^4)(\lambda_b + \lambda_\sigma)\right]$,

$c_2 = \frac{1}{16}\left[2\zeta_2\lambda_b^2 + \zeta_3\lambda_b\lambda_\sigma^2 + 2\zeta_4\lambda_\sigma^4\right]$, $\quad \kappa = 2 + \frac{2\beta}{\alpha} + \frac{3\lambda_a}{2\alpha} + \left(\frac{3\lambda_a + 3\beta}{\alpha}\right)^6$.

Remark 1: Convergence rate is $\mathcal{O}(\frac{1}{\epsilon^2})$ to the invariant measure.

Remark 2: Stein factors ζ_i depend on f and the chosen diffusion.

Condition 4 (Wasserstein decay). *The diffusion has L_1-Wasserstein decay ϱ*

$$\inf_{\text{couplings } (X_t^x, X_t^y)} \mathbb{E}[\|X_t^x - X_t^y\|_2] \le \varrho(t)\|x - y\|_2 \quad \textit{for all}\ x, y \in \mathbb{R}^d\ \textit{and}\ t \ge 0.$$

Theorem: Explicit bounds on the Stein factors

For an objective function f satisfying

$$|f(x) - f(y)| \le \pi_1(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2, \quad \text{for all } x, y \in \mathbb{R}^d,$$
$$\|\nabla^i f(x)\|_{\mathrm{op}} \le \pi_i(1 + \|x\|_2) \text{ for } i = 2, 3, 4 \text{ and for all } x, y \in \mathbb{R}^d,$$

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

$\zeta_i = \tau_i + \xi_i \int_0^\infty \varrho(t)dt$ where τ_i, and ξ_i have **explicit** forms.

(More generally, can support any polynomial growth in f and its derivatives.)

Explicit Bounds on Optimization Error

Proposition: Sampling yields near-optima

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

$$-p(\log p) + \log p(x^*) \ \le\ \tfrac{d}{2\theta}\log(\tfrac{2C}{d}) + \tfrac{d}{2}\log(\tfrac{e\beta}{\alpha}).$$

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

$$p_{\gamma,\theta}(f(x)) - f(x^*) \ \le\ \sqrt[\theta]{\tfrac{d}{2\gamma}\{\tfrac{1}{\theta}\log(\tfrac{2\gamma}{d}) + \log(\tfrac{e\beta\mu_2(f)}{2\alpha})\}}.$$

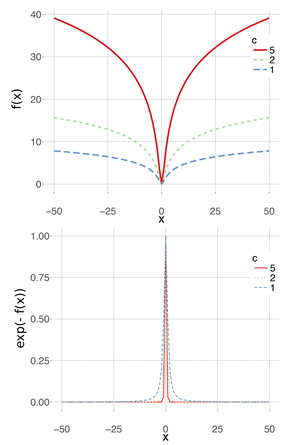
Corollary: Optimization error of discretized diffusions

If the diffusion has the generalized Gibbs stationary density $p_{\gamma,\theta}(x)$, then

$$\min_{m=1,..,M}\mathbb{E}[f(X_m)] - f(x^*) \le \left(c_1\tfrac{1}{\eta M} + (c_2 + c_3)\eta\right)\left(\kappa + \mathbb{E}\left[\|X_0\|_2^6\right]\right) + \sqrt[\theta]{\tfrac{d}{2\gamma}\{\tfrac{1}{\theta}\log(\tfrac{2\gamma}{d}) + \log(\tfrac{e\beta\pi_2}{2\alpha})\}}.$$

An Example with Sublinear Growth

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

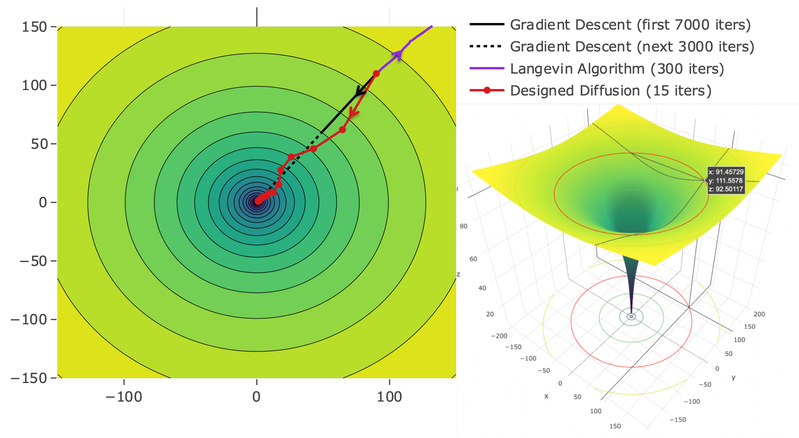


- $f(x)$ is non-convex with **sublinear growth**, so Langevin algorithm is not guaranteed to work!
- Choose $\sigma(x) = \frac{1}{\sqrt{\gamma}}\sqrt{1 + \frac{1}{2}\|x\|_2^2}\,I$, and $b(x) = -\frac{1}{2}(c - \frac{1}{\gamma})x$.
- Diffusion has target invariant measure $p(x) \propto e^{-\gamma f(x)}$.
- The diffusion is uniformly dissipative

$$2\langle b(x) - b(y), x - y\rangle + \|\sigma(x) - \sigma(y)\|_{\mathrm{F}}^2, \le -\alpha\|x - y\|_2^2, \text{ for } \alpha = c - \tfrac{d+3}{2\gamma};$$

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 $\gamma = 1$, $X_0 = (91, 111)$.



- In fact, the **optimization error** can be made of order ϵ by choosing the inverse temperature $\gamma = \mathcal{O}(\epsilon^{-1})$, the step size $\eta = \mathcal{O}(\epsilon^{1.5})$, and the number of iterations $M = \mathcal{O}(\epsilon^{-2.5})$.
- See the paper for additional examples like learning with non-convex losses, e.g., $f(x) = \frac{1}{n}\sum_{i=1}^n \psi_i(\langle x, v_i\rangle) + \rho(\frac{1}{2}\|x\|_2^2)$.