Global Non-convex Optimization with Discretized Diffusions

Murat A. Erdogdu\textsuperscript{1,2}, Lester Mackey\textsuperscript{3} and Ohad Shamir\textsuperscript{4}
\textsuperscript{1} University of Toronto \textsuperscript{2} Vector Institute \textsuperscript{3} Microsoft Research \textsuperscript{4} Weizmann Institute of Science

Presented by Chenyang Tao

Mar 1, 2018
Problem Setup

- Let $f(x) : \mathbb{R}^d \to \mathbb{R}$ be the target function.
- The goal is to carry out **unconstrained minimization** of $f(x)$
  with a diffusion process defined by
  \[ dX_t = b(X_t)\, dt + \sigma(X_t)\, dB_t \quad (1) \]
- $b(x) : \mathbb{R}^d \to \mathbb{R}^d$ is the drift term
- $\sigma(x) : \mathbb{R}^d \to \mathbb{R}^{d \times l}$ is the diffusion term
- $B_t$ is an $l$-dimensional Wiener process (e.g., Brownian motion)
\[ dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t \] (1)

For the above diffusion process defined by Eqn (1) to have target distribution \( p(x) \) as its invariant measure, set the diffusion term \( b(x) \) to

\[
b(x) = \frac{1}{2p(x)} \langle \nabla, p(x) \left( \sigma(x)\sigma(x)^T + C(x) \right) \rangle 
\] (2)

- \( C(x) \in \mathbb{R}^{d \times d} \) is skew-symmetric (\( C(x) = -C(x)^T \))
- \( \left[ \langle \nabla, M(x) \rangle \right]_j \triangleq \sum_k \partial_{x_k} M_{jk}(x) \) is the divergence operator
- \( M(x) \in \mathbb{R}^{d \times d} \)

\( A(x) \in \mathbb{R}^{d \times d} \)
Diffusion Process & Its Invariant Measure

\[ dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t \quad (1) \]

\[ b(x) = \frac{1}{2p(x)} \langle \nabla, p(x) (A(x) + C(x)) \rangle \quad (2) \]

A concrete example

- Inserting \( \sigma(x) = \sqrt{2}I \) and \( C(x) = 0 \), (1) recovers the standard Langevin diffusion

\[ X_t = \nabla \log p(X_t) \, dt + \sqrt{2} \, dB_t \]

with invariant measure \( p(x) \).
Let $\gamma > 0$ be the inverse temperature

Define the Gibbs measure: $p_\gamma(x) \propto \exp(-\gamma f(x))$

For $\gamma \to \infty$, we have $f(X_\gamma) \approx f(x^*)$, where $X_\gamma \sim p_\gamma(x)$

Use diffusion process (1) to sample $p_\gamma$

Euler discretization is used to simulate the diffusion

$$X_{m+1} = X_m + \eta b(X_{m-1}) + \sqrt{\eta} \sigma(X_m) \xi_m$$  \hspace{1cm} (3)

- $\eta > 0$ is the step size
- $\{\xi_m \in \mathbb{R}^d\}$ are i.i.d. standard Gaussians
Optimization Errors with Discretized Diffusion

- Optimization error decomposition for the discretized diffusion

\[
\min_{m=1, \ldots, M} \{ \mathbb{E}[f(X_m)] \} - \min_x f(x) 
\leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[f(X_m) - p(f)] + p(f) - \min_x f(x) \tag{4}
\]

Integration Error

Expected Suboptimality

- Causes of integration error
  - Short-term non-stationarity (initialization)
  - Long-term bias (accumulation of discretization errors)

- **Question:** What is the number of steps needed to assure integration error is less than \( \epsilon \)?
Error \leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[f(X_m) - p(f)] + p(f) - \min_x f(x)

(4)

Integration Error Expected Suboptimality

**Question:** What is the number of steps needed to assure integration error is less than $\epsilon$?

**Existing results:**
- **log-concavity:** $\mathcal{O}(\frac{1}{\epsilon^2}\text{poly}(\log(\frac{1}{\epsilon})))$, $\mathcal{O}(\frac{1}{\epsilon^4}\text{poly}(\log(\frac{1}{\epsilon})))$
  - strong
  - general
- Quadratic-growth of dissipativeness and Lipschitz gradient:
  $\mathcal{O}(\frac{1}{\epsilon^4}\text{poly}(\log(\frac{1}{\epsilon}))) \frac{1}{\lambda^*}$, $\lambda^*$ spectral gap
- Distant strong convexity: $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon}))$
Contributions of The Paper

- Providing explicit $O\left(\frac{1}{\epsilon^2}\right)$ bounds on the integration error
- Deriving explicit Stein factor bounds for pseudo-Lipschitz $f$
  - Applicable tools for computing the bounds
- Introducing new explicit bounds on the expected suboptimality
  - Combining with integration error bounds yields computable and convergent bounds on global optimization error
- Showing that different diffusions are appropriate for different objectives $f$
Pseudo-Lipschitz of order $n$:
\[
|g(x) - g(y)| \leq \tilde{\mu}_{1,n}(g) \left( 1 + \|x\|_2^n + \|y\|_2^n \right) \|x - y\|_2 \quad \forall x, y \in \mathbb{R}^d
\]

- $\phi_1(g) = \sup_{x \neq y} \left\{ \frac{\|g(x) - g(y)\|_F}{\|x - y\|_2} \right\}$,
- $\mu_0(g) = \sup_x \{ \|g(x)\|_{op} \}$,
- $\mu_i(g) = \sup_x \left\{ \frac{\|\nabla^i g(x)\|_{op}}{1 + \|x\|_2^n} \right\}$,
- $\tilde{\pi}_{i,n}(g) = \sup_x \left\{ \frac{\|\nabla^i g(x)\|_{op}}{1 + \|x\|_2^n} \right\}$
Condition 1. For some $r \in \{1, 2\}$ and $\forall x \in \mathbb{R}^d$, the drift and diffusion coefficients satisfy polynomial growth conditions:

$$
\|b(x)\|_2 \leq \frac{\lambda_b}{4} (1 + \|x\|_2), \quad \|\sigma(x)\|_F \leq \frac{\lambda_\sigma}{4} (1 + \|x\|_2), \quad \text{and}
$$

$$
\|a(x)\|_{op} \leq \frac{\lambda_a}{4} (1 + \|x\|_2^r)
$$

- Guarantees the existence of solution

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

$$
\mathcal{A}\|x\|_2^2 \leq -\alpha\|x\|_2^2 + \beta \quad \text{for} \quad \mathcal{A}g(x) \triangleq \langle b(x), \nabla g(x) \rangle + \frac{1}{2} \langle a(x), \nabla^2 g(x) \rangle
$$

- Diffusion travels inwards at distant location (does not diverge)
**Condition 3.** The function $u_f$ solves the Poisson equation $Au_f = f - p(f)$ is pseudo-Lipschitz of order $n$ with constant $\zeta_1$, and has $i$-th order derivative with degree-$n$ polynomial growth for $i = 2, 3, 4$, i.e.,

$$\| \nabla^i u_f(x) \|_{op} \leq \zeta_i (1 + \|x\|^n_2) \text{ for } i = 2, 3, 4 \forall x \in \mathbb{R}^d.$$ 

In other words, $\tilde{\mu}_{1,n}(u_f) = \zeta_1$, and $\tilde{\pi}_{i,n}(u_f) = \zeta_i$ for $i = 2, 3, 4$ with $\{\xi_i\}$ finite.

- $\{\xi_i\}$ are termed as Stein factors.
Theorem 3.1 (Integration error) Under Conditions 1 to 3 for some \( r \in \{1, 2\} \), for any even integer \( n_e \geq n + 4 \) and step size \( \eta < 1 \land \frac{\alpha}{2(n_e-1)!!(1+\lambda_b/2+\lambda_\sigma/2)^{n_e}} \),

\[
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[f(X_m) - p(f)] \leq C(\kappa_r(n_e) + \mathbb{E}[\|X_0\|_2^{n_e}]),
\]

where \( C = (c_1 \frac{1}{\eta M} + c_2 \eta + c_3 \eta^{1+|1\land n/2|}) \), and \( c_1, c_2, c_3 \) depend on \( \lambda_a, \lambda_b, \lambda_\sigma, \alpha, \beta, n_e \) and Stein factors \( \{\xi_i\} \).

- The above bound is dominated by \( \mathcal{O}(\frac{1}{\eta M} + \eta) \), which implies \( \eta = \mathcal{O}(\epsilon) \Rightarrow M = \mathcal{O}(\epsilon^{-2}) \) to get the desired tolerance.
While verifying Condition 1 and 2 are often straightforward, Condition 3 is usually unverifiable in practice.

**Condition 4.** (Wasserstein rate) The diffusion $Z^x_t$ has $L_p$-Wasserstein rate $\varrho_p(t) : \mathbb{R}_+ \to \mathbb{R}_+$ if

$$\mathbb{W}_p(X^x_t, X^y_t) \leq \varrho_p(t) \|x - y\|_2 \forall x, y \in \mathbb{R}^d \text{ and } t \geq 0,$$

where $\mathbb{W}_p(X, Y)$ denotes the $p$-Wasserstein distance between random variable $X$ and $Y$, and $X^x_t$ denotes the r.v. defined by diffusion (1) with $x$ as starting point.
Theorem 3.2 (Finite Stein factors from Wasserstein decay)
Assume that Conditions 1, 2 and 4 hold and that \( f \) is pseudo-Lipschitz continuous of order \( n \) with, for \( i = 2, 3, 4 \), at most degree-\( n \) polynomial growth of its \( i \)-th order derivatives. Then, the Stein factors in Condition 3 are can be expressed in terms of Wasserstein rates.
Two sufficient conditions for Wasserstein decay.
- Uniform and distant dissipativity

**Proposition 3.3** (Wasserstein decay from uniform dissipativity [F. Wang (2016), Thm. 2.5]). A diffusion with drift and diffusion coefficients $b(x)$ and $\sigma(x)$ has Wasserstein rate $\varrho_p(t) = \exp\left(-\frac{kt}{2}\right)$ if, for all $x, y \in \mathbb{R}^d$,

$$
2\langle b(x) - b(y), x - y \rangle + \left\| \sigma(x) - \sigma(y) \right\|^2_F \\
+ (p - 2) \left\| \sigma(x) - \sigma(y) \right\|_{op}^2 \leq -k \left\| x - y \right\|^2_2
$$

(6)
Proposition 3.4 (Wasserstein decay from distant dissipativity [J. Gorham, et al. (2016), Col. 4.2]). A diffusion with drift and diffusion coefficients $b(x)$ and $\sigma(x)$ has Wasserstein rate $\varrho_p(t) = 2 \exp(LR^2/8) \exp(-kt/2)$ for $R, L \geq 0$ if, for all $x, y \in \mathbb{R}^d$,

\[
\begin{align*}
\text{a very involved equation} \leq \begin{cases} 
-K & \|x - y\|_2 > R, \\
L & \|x - y\|_2 \leq R,
\end{cases}
\end{align*}
\]

with $K > 0$ and additional conditions on $s \in (0, 1/\mu_0(\sigma^{-1}))$.

\[
2\langle b(x) - b(y), x - y \rangle + \|\sigma(x) - \sigma(y)\|_F^2 \\
+ (p - 2)\|\sigma(x) - \sigma(y)\|_{op}^2 \leq -k\|x - y\|_2^2
\]
Proposition 3.5 (User-friendly Wasserstein decay). Fix any diffusion and skew-symmetric stream coefficients $\sigma(x)$ and $C(x)$ satisfying $L^* = \sqrt{F_1(\tilde{\sigma})^2 + \sup_x \lambda_{\max}(\nabla \langle \nabla, m(x) \rangle)} < \infty$ for $m(x) \doteq A(x) + C(x)$, $\tilde{\sigma}(x) \doteq \left(A(x) - s_0^2 I\right)^{1/2}$, and $s_0 \in (0, 1/\mu_0(\sigma^{-1})$. If

$$-\langle m(x) \nabla f(x) - m(y) \nabla f(y), x - y \rangle \leq \begin{cases} -K_m, & \text{if } \|x - y\|_2 > R_m \\ L_m, & \text{if } \|x - y\|_2 \leq R_m \end{cases}$$

holds for $R_m, L_m \geq 0, K_m > 0$, then, or any inverse temperature $\gamma > L^*/K_m$, the diffusion with drift and diffusion coefficient $b_\gamma(x) = \frac{1}{2} m(x) \nabla f + \frac{1}{2\gamma} \langle \nabla, m(x) \rangle$ and $\sigma_\gamma = \frac{1}{\sqrt{\gamma}} \sigma$ has stationary distribution $p_\gamma(x) \propto \exp(-\gamma f(x))$ and satisfies distant dissipativity.
Proposition 4.1 (Expected suboptimality: Sampling near-optima) Suppose $p(x)$ is the stationary density of an $(\alpha, \beta)$-dissipative diffusion (Condition 2) with global maximizer $x^*$. Fix $C > 0$ and $\theta \in (0, 1]$. If
\[
\log p(x^*) - \log p(x) \leq C \|x - x^*\|_2^{2\theta} \quad \text{for all } x,
\]
then
\[
-p(\log p) + \log p(x^*) \leq \frac{d}{2\theta} \log \left( \frac{2C}{d} \right) + \frac{d}{2} \log \left( \frac{e\beta}{\alpha} \right). \tag{10}
\]
If this $p$ takes the generalized Gibbs form
\[
p_{\gamma, \theta}(x) \propto \exp(-\gamma (f(x) - f(x^*))^\theta) \quad \text{for } \gamma > 0,
\]
we have
\[
p_{\gamma, \theta}(f(x)) - f(x^*) \leq \left( \frac{d}{2\gamma} \left( \frac{1}{\theta} \log \left( \frac{2\gamma}{d} \right) + \log \left( \frac{e\beta \mu_2(f)}{2\alpha} \right) \right) \right)^{1/\theta}. \tag{11}
\]
**Corollary 4.2** Under the assumptions of Theorem 3.1 and Proposition 4.1, if the diffusion has the generalized Gibbs stationary density $p_{\gamma, \theta} \propto \exp(-\gamma (f(x) - f(x^*))^\theta)$, then

$$\min_{m=1, \ldots, M} \mathbb{E} f(X_m) - f(x^*) \leq \left( c_1 \frac{1}{\eta M} + (c_2 + c_3) \eta \right) \left( \kappa_r(n_e) + \mathbb{E}[\|X_0\|_{n_e}^2] \right)$$

$$+ \left( \frac{d}{2\gamma} \left( \frac{1}{\theta} \log\left( \frac{2\gamma}{d} \right) + \log\left( \frac{e^{\beta \mu_2(f)}}{2\alpha} \right) \right) \right)^{1/\theta}$$

**Proposition 4.3** For quadratic function $f(x) = \langle x - b, A(x - b) \rangle$ for a positive definite $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$. Then for each positive integer $k$, we have

$$p_{\gamma, 1/k}(f) - f(x^*) \leq \left( \frac{k(1 + \frac{d}{2}) - 1}{\gamma} \right)^k.$$