Global non-convex optimization with discretized diffusions

Murat Erdogdu, Lester Mackey, Ohad Shamir (NeurIPS 2018)

Geelon So, agso@eng.ucsd.edu
Sampling/Optimization Reading Group — April 14, 2021

Optimization goal

Consider the unconstrained (non-convex) optimization problem:

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

- **Sampling approach:** sample from a distribution maximized at the minima of f.
  - e.g. the **Gibbs measure** with inverse temperature $\gamma > 0$

$$p_\gamma(x) \propto \exp\big(-\gamma f(x)\big).$$

The Langevin algorithm

Gradient descent with noise

The **Langevin algorithm** is gradient descent on f with Gaussian noise:

$$X_{m+1} = X_m - \eta \nabla f(X_m) + \sqrt{\frac{2\sigma}{\eta}} W_{m+1},$$

where $W_{m+1} \sim \mathcal{N}(0, I_{d\times d})$.

- **Output:** at time M, return $\min_{m\in[M]} f(X_m)$.

Analysis: Langevin algorithm as discretized diffusion

The Langevin update can be viewed as a **discretization** of the (overdamped) Langevin diffusion for p_γ, which is the solution to this stochastic differential equation (SDE):

$$dZ_t = -\nabla f(Z_t)\, dt + \sqrt{\frac{2}{\gamma}}\, dB_t \quad \text{with} \quad Z_0 = X_0.$$

- This SDE has a limiting invariant distribution p_γ.

Analysis: optimization error

If p is a distribution, denote $p(f) = \mathbb{E}_{Z \sim p}[f(Z)]$. The optimization error after M steps is:

$$\min_{m \in [M]} \mathbb{E}[f(X_m)] - \min_x f(x) \leq \underbrace{\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[f(X_m) - p(f)]}_{\text{integration error}} + \underbrace{p(f) - \min_x f(x)}_{\text{expected suboptimality}} .$$

- The integration error bounds the short-term non-stationarity and long-term discretization bias.
- For the Gibbs measure $p = p_\gamma$, the suboptimality gap is controlled by the inverse temperature $\gamma > 0$.

This paper: beyond the overdamped Langevin diffusion

Question: what about other distributions p besides the Gibbs measure such that

$$p(f) \approx \min_x f(x)?$$

- This paper considers p that are the limiting invariant distributions of more general diffusion processes:
$$dZ_t = b(Z_t)\,dt + \sigma(Z_t)\,dB_t.$$
- Here, the covariance coefficient $\sigma(x)\sigma(x)^\top$ can be nonconstant over $\mathbb{R}^d$ (this is constant for the overdamped Langevin diffusion; c.f. preconditioning).

This paper: techniques

$$\min_{m\in[M]} \mathbb{E}[f(X_m)] - \min_x f(x) \leq \underbrace{\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[f(X_m) - p(f)]}_{\text{integration error}} + \underbrace{p(f) - \min_x f(x)}_{\text{expected suboptimality}} .$$

- Bounds the integration error via **Stein's method**.
- Bounds expected suboptimality for **generalized Gibbs measures** of the form:

$$p_{\gamma,\theta}(x) \propto \exp\left(-\gamma\big(f(x) - f(x^*)\big)^{\theta}\right),$$

  where x^* is a global maximizer with $\nabla f(x^*) = 0$.
  - **Note:** this assumes knowledge of $f(x^*)$, which is often 0 in many settings. If $f(x^*)$ is unknown, can carry out analysis just using an estimate.

Motivation: techniques to bound integration error

Goal: bounding the integration error

- **Invariant measure:** let p be the stationary distribution of a diffusion $(Z_t)_{t \geq 0}$
- **Discretization:** let $(X_m)_{m=0}^{\infty}$ be an appropriate discretization with step size η
- **Integration error:** $\frac{1}{M} \sum_{m \in [M]} \mathbb{E}[f(X_m) - p(f)]$

Theorem (Integration error, informal)

The integration error at time M is $O\left(\frac{1}{\eta M} + \eta\right)$.

Broad strokes

- Let $(Z_t)_{t\geq 0}$ be the continuous-time diffusion.
  - The distribution of Z_t converges to the stationary distribution p, so that:

$$\lim_{t\to\infty} \mathbb{E}[f(Z_t) - p(f)] = 0.$$

  - In fact, a quantitative **mean ergodic theorem** states:

$$\mathbb{E}\left[\left(\frac{1}{T}\int_0^T f(Z_t)\,dt - p(f)\right)^2\right] = O\left(\frac{1}{T}\right),$$

    which can be read as: *'the time average converges to the space average.'*

- The discretization using step size η leads to some additional bias:

$$\left|\frac{1}{M}\sum_{m\in[M]} \mathbb{E}[f(X_m) - p(f)]\right| \leq O\left(\frac{1}{\eta M} + \eta\right).$$

Roadmap

1. Interlude I: Markov semigroups
2. Interlude II: the Poisson equation and mean ergodic theorem
3. Interlude III: Connection to Stein's method
4. Optimization using discretized diffusion

Interlude I: Markov semigroups and generators[1]

[1]This section closely follows (Bakry et al., 2013).

Markov process

Definition (Markov process)

Let $(X_t)_{t\geq 0}$ be a stochastic process. Then, it is a (time-homogeneous) ***Markov process*** *if for all $t > s$, the law of X_t given $(X_u)_{0\leq u\leq s}$ is equal to:*

- *the law of X_t given X_s*
- *the law of X_{t-s} given X_0.*

Notation: we often write $(X_t^x)_{t\geq 0}$ to denote that the process is initially $X_0 = x$.

- Example: the solution $(X_t^x)_{t\geq 0}$ to the Itô SDE is Markov:

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t \quad \text{and} \quad X_0 = x.$$

Example: Brownian motion

Let $(B_t^x)_{t\geq 0}$ be d-dimensional **Brownian motion** with $B_0 = x$.

- Define the probability kernels for $t > 0$,

$$p_t(x,y) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x-y\|^2}{2t}\right),$$

  and let $p_0(x,\cdot)$ be the Dirac distribution at x, so that B_t^x has density $p_t(x,\cdot)$.

- We can characterize the evolution of $p_t(x,\cdot)$ by defining the operator P_t for $t \geq 0$,

$$P_t f(x) = \int_{\mathbb{R}^d} f(y)\, p_t(x, dy),$$

  for f bounded measurable map.

The associated semigroup

Let $(X_t)_{t\geq 0}$ be Markov. Define the operator P_t on bounded measurable functions f by:

$$P_t f(x) = E[f(X_t)|X_0 = x].$$

Read: P_t *maps* f *to the function describing the expected value of* f *after time* t.

- The Markov property implies that $\mathbf{P} = (P_t)_{t\geq 0}$ is a **semigroup**:

$$P_{t+s}f(x) = P_t\big(P_s f\big)(x).$$

  That is, $P_{t+s} = P_t \circ P_s$ and $P_0 = \mathrm{Id}$.

Properties of the semigroup

Let $\mathbf{P} = (P_t)_{t\geq 0}$ be the semigroup for a Markov process $(X_t)_{t\geq 0}$ on $\mathbb{R}^d$. Let $\mathcal{M}$ be the set of bounded measurable $f : \mathbb{R}^d \to \mathbb{R}$.

(i) **Linearity:** P_t is a linear operator on $\mathcal{M}$ for all $t \geq 0$.

(ii) **Initial condition:** $P_0 = \mathrm{Id}$.

(iii) **Mass conservation:** $P_t(\mathbf{1}) = \mathbf{1}$.

(iv) **Positivity:** if $f \geq 0$, then $P_t f \geq 0$.

(v) **Semigroup (Markov) property:** $P_{t+s} = P_t \circ P_s$ for all $t, s \geq 0$.

Question: can we define the derivative of the map $t \mapsto P_t$?

- Knowing evolution of P_t implies a lot about the Markov process.
  - However, Markov processes are too general to define derivative—let's impose a continuity condition, which we can get at through $\mathbf{P}$.

Invariant measure

Definition (Invariant measure, Bakry et al. (2013))

Let $\mathbf{P} = (P_t)_{t \geq 0}$ *be the semigroup for a Markov process on* $\mathbb{R}^d$*. The Markov process has* ***invariant measure*** μ *on* $\mathbb{R}^d$ *if for all* $t \geq 0$,

$$\int_{\mathbb{R}^d} P_t f \, d\mu = \int_{\mathbb{R}^d} f \, d\mu.$$

Markov semigroup

Definition (Markov semigroup, Bakry et al. (2013))

Let $\mathbf{P} = (P_t)_{t\geq 0}$ be a family of operators satisfying conditions (i–v) on Slide 17. Suppose it has an invariant measure μ. Then $\mathbf{P}$ is a ***Markov semigroup*** *if it further satisfies:*

(vi) ***Continuity:*** *for every $f \in L^2(\mu)$, we have convergence in $L^2(\mu)$,*

$$\lim_{t\downarrow 0} P_t f = f.$$

- If $(X_t)_{t\geq 0}$ has a Markov semigroup, then X is called a **Feller process**.
  - X has additional regularity properties (e.g. it has a *cádlág* modification).
  - If $t \mapsto X_t$ is continuous and Feller, then $(X_t)_{t\geq 0}$ is a **diffusion process**.

Defining the derivative

Let C_0 be the set of continuous $f : \mathbb{R}^d \to \mathbb{R}$ vanishing at infinity.

Definition (Infinitesimal generator, Revuz and Yor (2013))

Let $(X_t)_{t\geq 0}$ be a Feller process. Let $\mathbf{P} = (P_t)_{t\geq 0}$ be its Markov semigroup. A function $f \in C_0$ belongs to the domain $\mathcal{D}_{\mathcal{A}}$ of the ***infinitesimal generator*** *$\mathcal{A}$ if the limit exists in C_0,*

$$\mathcal{A}f = \lim_{t\downarrow 0} \frac{P_t f - P_0 f}{t}.$$

- It turns out that $\mathcal{A}$ and $\mathcal{D}_{\mathcal{A}}$ completely characterizes $\mathbf{P}$.

Interpretation of generator

Let $(X_t)_{t\geq 0}$ be a Markov process with generator $\mathcal{A}$ and $f \in \mathcal{D}_{\mathcal{A}}$. Taylor's formula shows:

$$E\big[\,f(X_{t+h}) - f(X_t) \,\big|\, X_t\big] = P_h f(X_t) - P_0 f(X_t) \approx h\mathcal{A}f(X_t).$$

- Therefore, $\mathcal{A}f(X_t)$ describes the *infinitesimal expected evolution* from $f(X_t)$.
- If f is a test function, the evolution of the statistic $E[f(X_t)]$ is described by $\mathcal{A}$.

Properties of the generator

Proposition

$$\partial_t P_t = \mathcal{A} P_t = P_t \mathcal{A}.$$

Proof.

Fix t and consider the map $s \mapsto P_s$. The semigroup property implies:

$$\frac{P_{t+h} - P_t}{h} = P_t \left(\frac{P_h - P_0}{h} \right) = \left(\frac{P_h - P_0}{h} \right) P_t,$$

where taking the limit $h \downarrow 0$ shows $\partial_t P_t = \mathcal{A} P_t = P_t \mathcal{A}$. □

- Earlier: $\mathcal{A}$ characterizes $\mathbf{P}$. Indeed, since $\partial_t P_t = \mathcal{A} P_t$, we formally have $P_t = e^{t\mathcal{A}}$.

Properties of the generator (cont.)

Proposition

Let $\mathbf{P}$ *be a Markov semigroup with invariant distribution* μ. *Let* $\mathcal{A}$ *be its generator. Then for all* $f \in L^1(\mu)$,

$$\mathop{\mathbb{E}}_{X\sim\mu}[\mathcal{A}f(X)] = \int_{\mathbb{R}^d} \mathcal{A}f(x)\,\mu(dx) = 0.$$

Proof.

Since $\mathcal{A}f = \partial_t P_t f\big|_{t=0}$, by interchanging limits and applying invariance, we have:

$$\mathop{\mathbb{E}}_{X\sim\mu}[\mathcal{A}f(X)] = \frac{\partial}{\partial t}\int_{\mathbb{R}^d} P_t f(x)\,\mu(dy)\bigg|_{t=0} = \frac{\partial}{\partial t}\int_{\mathbb{R}^d} f(x)\,\mu(dy) = 0. \qquad \square$$

- Indeed, if $X_0 \sim \mu$ is invariant, then $\mathbb{E}[f(X_t)]$ is a constant over all time.

Kolmogorov backward equation

Theorem (Hille-Yosida, Brezis (2010))

Let $f \in \mathcal{D}_{\mathcal{A}}$. Define the statistic $u(t,x) = E[f(X_t) \mid X_0 = x]$. For all times, $u(t,\cdot) \in \mathcal{D}_{\mathcal{A}}$. Further, u is uniquely defined the partial differential equation:

$$\frac{\partial u}{\partial t} = \mathcal{A}u, \qquad t > 0, x \in \mathbb{R}^d$$
$$u(0,x) = f(x); \qquad x \in \mathbb{R}^d,$$

where $\mathcal{A}$ is applied to the function $x \mapsto u(t,x)$.

- This PDE (Kolmogorov backward equation) describes the evolution of statistics u.
- C.f. Picard's theorem for existence and uniqueness for the ODE: if $F : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz, then for all initial point $x_0 \in \mathbb{R}^d$, there exists unique $x(t)$ satisfying:

$$\frac{dx(t)}{dt} = F(x(t)) \quad \text{and} \quad x(0) = x_0$$

Dynkin's formula

Corollary (Dynkin's formula)

If $f \in \mathcal{D}_\mathcal{A}$, then: $\quad P_t f(x) = f(x) + E\left[\int_0^t \mathcal{A} f(X_s^x)\, ds\right].$

Proof.

Integrating the Kolmogorov backward equation, we obtain:

$$\begin{aligned} P_t f(x) &= f(x) + \int_0^t \mathcal{A} P_s f(x)\, ds \\ &= f(x) + \int_0^t P_s \mathcal{A} f(x)\, ds && (P_s \text{ and } \mathcal{A} \text{ commute}) \\ &= f(x) + \int_0^t E[\mathcal{A} f(X_s^x)]\, ds && (\text{definition of } P_s) \end{aligned}$$

Applying Fubini's to interchange limits proves the result. □

Duality and the Kolmogorov forward equation

Let $(X_t^x)_{t\geq 0}$ be a Feller process where $p_t(x,\,\cdot\,)$ describes distribution of X_t^x.

(i) Let $f \in \mathcal{D}_\mathcal{A}$. Then $E[f(X_t^x)] = P_t f(x) = \int_{\mathbb{R}^d} f(y)\, p_t(x, dy)$.

(ii) Therefore, the Kolmogorov backward equation states:

$$\partial_t P_t f(x) = \mathcal{A} P_t f(x) = P_t \mathcal{A} f(x) = \int_{\mathbb{R}^d} \mathcal{A} f(y)\, p_t(x, y)\, dy.$$

(iii) Apply duality to RHS to obtain $\partial_t P_t f(x) = \int_{\mathbb{R}^d} f(y)\, \mathcal{A}^* p_t(x, dy)$.

(iv) Combine (ii) and (iv), and interchange limits to obtain **Fokker-Planck**:

$$\partial_t p_t(x,\,\cdot\,) = \mathcal{A}^* p_t(x,\,\cdot\,) \qquad \text{(Kolmogorov forward equation)}$$

which describes the evolution of the distribution of X_t.

Focus: Itô diffusion

Let's reduce the level of generality and apply our results to Itô diffusion:

Definition (Itô diffusion, Øksendal (2003))

Let $(B_t)_{t\geq 0}$ be an m-dimensional Brownian motion. A (time-homogeneous) ***Itô diffusion*** *$(X_t)_{t\geq 0}$ is a solution to the Itô stochastic differential equation*

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t,$$

where $b(x) \in \mathbb{R}^n$ is the ***drift coefficient*** *and $\sigma(x) \in \mathbb{R}^{n\times m}$ (or sometimes $\frac{1}{2}\sigma\sigma^\top$) is the* ***diffusion coefficient****. Furthermore, $b(\,\cdot\,)$ and $\sigma(\,\cdot\,)$ are Lipschitz continuous,*[2]

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq K|x - y|; \qquad x, y \in \mathbb{R}^n.$$

[2] Recall the Lipschitz condition ensures existence and uniqueness of solution.

Example: deterministic flow

Let $(x_t)_{t\geq 0}$ be the solution to $dx_t = b(x_t)\,dt$ and $x_0 = x$.

- If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then by chain rule:

$$\mathcal{A}f(x) = \lim_{t\downarrow 0} \frac{f(x_t) - f(x)}{t} = \left(\frac{dx_t}{dt}\right)^\top \nabla f(x)$$

For deterministic flows, $\mathcal{A}$ is a first-order partial differential operator on $C^1(\mathbb{R}^n; \mathbb{R})$,

$$\mathcal{A}f(x) = \sum_{i=1}^{n} b_i(x) \cdot \partial_i f(x).$$

Review: one-dimensional Itô's formula

Recall that dB_t can be thought of as an infinitesimal of order $1/2$, with $dB_t^2 = dt$.

- Taylor expansion of $df(X_t)$ looks like:

$$\begin{aligned} df(X_t) &= f(X_t + dX_t) - f(X_t) \\ &= f(X_t)dX_t + \frac{1}{2}f''(X_t)dX_t^2 + \cdots \\ &= f'(X_t)\{b\,dt + \sigma\,dB_t\} + \frac{1}{2}\sigma^2 f''(X_t)\,dB_t^2, \end{aligned}$$

where all higher-order terms have been discarded. Apply $dB_t^2 = dt$ to obtain:

$$df(X_t) = \left\{ b f' + \frac{1}{2}\sigma^2 f'' \right\} dt + \sigma f'\,dB_t.$$

One-dimensional generator

Let $(X_t^x)_{t\geq 0}$ be a one-dimensional Itô diffusion. Let $f \in C_0^2(\mathbb{R})$ (i.e. twice differentiable with compact support). Apply Itô's formula:

$$E^x[f(X_t^x)] = f(x) + \int_0^t \left\{ b f' + \frac{1}{2}\sigma^2 f'' \right\} ds + E\left[\int_0^t \sigma f' \, dB_s \right].$$

- Condition on f implies $E\left[\int_0^t \sigma f' \, dB_s \right] = 0$.
- Fundamental theorem of calculus implies:

$$\mathcal{A}f(x) = b(x) \cdot \partial f(x) + \frac{1}{2}\sigma(x)^2 \, \partial^2 f(x).$$

Thus, $\mathcal{A}$ is a second-order differential operator on $C_0^2(\mathbb{R};\mathbb{R})$.

Generator of Itô diffusion

Theorem (Form of generator, Øksendal (2003))

Let X_t be the Itô diffusion $dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t$. If $f \in C_0^2(\mathbb{R}^n)$, define the second-order partial differential operator L*:*

$$\mathrm{L}f(x) = \sum_{i=1}^{n} b_i(x)\,\partial_i f(x) + \frac{1}{2}\sum_{i,j}\left(\sigma\sigma^\top\right)_{ij}(x)\,\partial_{ij} f(x).$$

Then, $C_0^2(\mathbb{R}^n) \subset \mathcal{D}_{\mathcal{A}}$ and if for all $f \in C_0^2(\mathbb{R}^n)$:

$$\mathcal{A}f = \mathrm{L}f.$$

- We sometimes call L the **Markov generator** of $(X_t)_{t\geq 0}$.
- Itô's formula shows: $f(X_t^x) = f(x) + \int_0^t \mathrm{L}f(X_s^x)\,ds + \int_0^t \sigma(X_s^x)^\top \nabla f(X_s^x)\,dB s$.

Example: Brownian motion

Let X_t solve $dX_t = \sqrt{2}dB_t$. Then the generator of B_t is:

$$\mathrm{L}f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2},$$

when $f \in C_0^2(\mathbb{R}^n)$. That is, $\mathrm{L} = \Delta$, where Δ is the Laplace operator.

- If $u(t,x) = P_t f(x)$, then u solves the heat equation:

$$\partial_t u = \Delta u \quad \text{and} \quad u(0,x) = f(x).$$

Adjoints

- Let $\langle \cdot\, ,\, \cdot \rangle$ be the inner product in $L^2(dy)$, so that $\langle \phi, \psi \rangle = \int_{\mathbb{R}^n} \phi(y)\psi(y)\, dy$.
- Given a bounded linear operator T, the adjoint T^* is defined:

$$\langle T\phi, \psi \rangle = \langle \phi, T^*\psi \rangle.$$

- Let D be the differential operator. If $T\phi(x) = \sum_{k=0}^{n} a_k(x) D^k \phi(x)$, then:

$$T^*\psi = \sum_{k=0}^{n} (-1)^k D^k \left(a_k \psi \right).$$

Kolmogorov's forward equation (or, the Fokker-Planck equation)

The evolution of a statistic $u(t,x) = E[f(X_t^x)]$ from Kolmogorov backward equation:

$$P_t f(x) = f(x) + \int_0^t \mathcal{A}u(s,x)\,ds = f(x) + \int_0^t \int_{\mathbb{R}^d} \mathrm{L}f(y)\,p_s(x,dy)\,ds,$$

where $p_s(x,\cdot)$ describe the distribution of X_s^x. For short, let $p_s^x = p_s(x,\cdot)$.

- The Kolmogorov forward equation describes the dual—how does the probability density evolve over time. Taking the time derivative:

$$\langle f, \partial_t p_t^x \rangle = \langle \mathrm{L}f, p_t^x \rangle = \langle f, \mathrm{L}^* p_t^x \rangle; \qquad f \in C_0^2.$$

- Let $\mathbf{D} = \frac{1}{2}\sigma\sigma^\top$ be the diffusion matrix. This implies:

$$\frac{\partial p_t^x}{\partial t} = \sum_{i,j} \frac{\partial^2}{\partial_i \partial_j}\left(\mathbf{D}_{ij} p_t^x\right) - \sum_{i=1}^n \partial_i\left(b_i \cdot p_t^x\right) \qquad \text{(Fokker-Planck equation)}$$

Markov semigroups and generators: summary

- If $(X_t)_{t\geq 0}$ is a diffusion process, it is characterized by its Markov semigroup $(P_t)_{t\geq 0}$,

$$P_t f(x) = E[f(X_t^x)] = \int_{\mathbb{R}^d} f(y)\, p_t(x, dy).$$

- The generator $\mathcal{A}$, seen as the derivative of $t \mapsto P_t$, also characterizes $(P_t)_{t\geq 0}$,

$$\mathcal{A}f = \lim_{t\downarrow 0} \frac{P_t f - P_0 f}{t},$$

  which computes $E[f(X_t^x)]$ via the Kolmogorov backward equation.
- If $(X_t)_{t\geq 0}$ is Itô diffusion, then $\mathcal{A}$ has the form of the Markov generator L.
- If $(X_t)_{t\geq 0}$ has an invariant distribution μ, then for all $f \in \mathcal{D}_{\mathcal{A}}$:

$$\mathop{\mathbb{E}}_{X\sim\mu}[\mathcal{A}f(X)] = \int_{\mathbb{R}^d} \mathcal{A}f(x)\, \mu(dx) = 0.$$

Interlude II: the Poisson equation and mean ergodic theorem[3]

[3] This is an informal version of Mattingly et al. (2010) mainly for intuition; see paper for rigor.

Estimating the invariant measure

Consider the following Itô diffusion:

$$dX_t = b(X_t)\, dt + \sigma(X_t)\, dB_t,$$

and assume that it has a unique stationary measure μ.

- **Goal:** estimate $\mathbb{E}_{X\sim\mu}[f(X)]$ for wide array of test functions f.

The Poisson equation

- Let $\mathcal{A}$ be the infinitesimal generator of the SDE.
- Let $f : \mathbb{R}^d \to \mathbb{R}$ be sufficiently smooth.
- Define $\overline{f}$ as the space average w.r.t. the invariant distribution μ,

$$\overline{f} = \int_{\mathbb{R}^d} f(x)\, \mu(dx) = \mathop{\mathbb{E}}_{X\sim\mu}[f(X)].$$

- The **Poisson equation** is the following PDE:

$$\mathcal{A}u_f = f - \overline{f}.$$

  - Under certain conditions, a unique (up to a constant term) and smooth u_f exists.

Solution to the Poisson equation

The formal solution to the Poisson equation is $u_f(x) = -\int_0^\infty P_t(f - \bar{f})(x)\, dt$.

Proof (informal).

Since μ is the stationary distribution, define the operator:

$$P_\infty g(x) = \int_{\mathbb{R}^d} g(y)\, \mu(dy) = \mathop{\mathbb{E}}_{X\sim\mu}[g(X)],$$

so that P_t converges to P_∞ as $t \to \infty$. In particular, $P_\infty(f - \bar{f}) = 0$.

- By Dynkin's formula, $P_\infty(f - \bar{f}) = (f - \bar{f}) + \int_0^\infty \mathcal{A} P_t(f - \bar{f})\, dt$.
- Substitute $P_\infty(f - \bar{f}) = 0$ and $(f - \bar{f}) = \mathcal{A} u_f$ to obtain

$$0 = \mathcal{A}\left[u_f + \int_0^\infty P_t(f - \bar{f})\, dt\right]. \qquad \square$$

Solution to the Poisson equation (cont.)

$$u_f(x) = -\int_0^\infty P_t(f - \overline{f})(x)\, dt$$

- **Interpretation:** $u_f(x)$ measures the *net fluctuation over all time of* $P_t f$ *from* $\overline{f}$.
  - Note: $P_t f$ converges to $P_\infty f = \overline{f}$.
  - Under certain assumptions, u_f is bounded (essentially, the diffusion X_t remains in a bounded region of the space).

Mean ergodic theorem

Recall Itô's formula:

$$u_f(X_t^x) = u_f(x) + \int_0^t \mathcal{A}u_f(X_s^x)\, ds + \int_0^t \sigma(X_s^x)^\top \nabla u_f(X_s)\, dB_s.$$

- Replace $\mathcal{A}u_f = f - \bar{f}$ and rearrange:

$$\frac{1}{t}\int_0^t f(X_s^x)\, ds - \bar{f} = \frac{u_f(X_t^x) - u_f(x)}{t} - \frac{1}{t}\int_0^t \sigma^\top \nabla u_f\, dB_s.$$

  - Assuming that u_f, $\|\sigma\|$, $\|\nabla u_f\|$ are bounded, then Itô isometry implies:

$$E\left[\left(\frac{1}{T}\int_0^T f(X_t)\, dt - \bar{f}\right)^2\right] \leq \frac{K}{T}.$$

Mean ergodic theorem (cont.)

It follows that we can view $\frac{1}{T}\int_0^T f(X_t)\,dt$ as an estimator for $\mu(f) = \mathbb{E}_{X\sim\mu}[f(X)]$.

- The *time average* of $f(X_t)$ converges in L^2 to the *space average* $\mu(f)$ with respect to the stationary distribution μ.
- The *rate of convergence* depends on the bounds on $\|\nabla u_f\|$ where u_f is solution to the Poisson equation,

$$\mathcal{A}u_f = f - \mu(f).$$

The bounds on $\|\nabla u_f\|$ are called **Stein factors**, hidden in the constant K.

Interlude: connection to Stein's method[4]

[4]This section follows Gorham et al. (2019).

Stein's method

Goal: bound distance between two distributions ν_t and μ, e.g. give non-asymptotic rates of convergence. In particular, quantify how well $\mathbb{E}_{\nu_t}$ approximates $\mathbb{E}_{\mu}$.

- Developed by Charles Stein to provide alternate proof of the central limit theorem.

General framework

Let ν_t and μ be distributions on $\mathbb{R}^d$. Let $\mathcal{F}$ be a family of test functions $f : \mathbb{R}^d \to \mathbb{R}$. Define the measure $d_{\mathcal{F}}(\nu_t, \mu)$ by:

$$d_{\mathcal{F}}(\nu_t, \mu) = \sup_{f \in \mathcal{F}} \left| \mu(f) - \nu_t(f) \right|.$$

- $d_{\mathcal{F}}$ is **convergence determining** if it is an integral probability metric (IPM) and $d_{\mathcal{F}}(\nu_t, \mu)$ converges to zero only if ν_t converges in distribution to μ.
- In general, it may be intractable to evaluate $\mathbb{E}_\mu[f(Z)]$.
  - **Idea:** it suffices to replace each f with $f - \mu(f)$, so $\mu(f) = 0$ for all $f \in \mathcal{F}$.

Stein's method

1. Identify an operator $\mathcal{T}$ acting on functions $g : \mathbb{R}^d \to \mathbb{R}^d$ in $\mathcal{G}$ to mean-zero functions under μ,
$$\mu(\mathcal{T}g) = 0 \quad \forall g \in \mathcal{G}.$$
2. Define the **Stein discrepancy**,
$$\mathcal{S}(\nu_t, \mathcal{T}, \mathcal{G}) = \sup_{g \in \mathcal{G}} \left|\nu_t(\mathcal{T}g)\right| = d_{\mathcal{T}\mathcal{G}}(Q_n, P),$$
where $\mathcal{T}\mathcal{G} = \{\mathcal{T}g : g \in \mathcal{G}\}$.
3. Prove that for each $f \in \mathcal{F}$ there exists some $u_f \in \mathcal{G}$ solving the **Stein equation**,
$$\mathcal{T}u_f = f - \mu(f).$$
Thus, $\mathcal{F} \subset \mathcal{T}\mathcal{G}$ so that $d_{\mathcal{F}}(\nu_t, \mu) \leq d_{\mathcal{T}\mathcal{G}}(\nu_t, \mu)$.
4. Prove upper bounds the Stein discrepancy so that $\mathcal{S}(\nu_t, \mathcal{T}, \mathcal{G}) \to 0$.

Stein's method (cont.)

Summary: in order to show ν_t converges in distribution to μ,

- If $\mathcal{F}$ is convergence determining, we need to show $d_{\mathcal{F}}(\nu_t, \mu) \to 0$.
- Show (step 3) that if $d_{\mathcal{TG}}(\nu_t, \mu)$ converges to zero, then so does $d_{\mathcal{F}}(\nu_t, \mu)$.
- Show (step 4) that $d_{\mathcal{TG}}(\nu_t, \mu)$ converges to zero.

Stein's method for diffusion processes

Recall: if $(X_t)_{t\geq 0}$ is a Feller process with generator $\mathcal{A}$ and invariant measure μ, then:

$$\mu(\mathcal{A}u) = 0 \quad \forall u \in \mathcal{D}_{\mathcal{A}}.$$

It follows that we just need to be able to solve the Stein equation,

$$\mathcal{A}u_f = f - \mu(f),$$

which is the familiar Poisson equation. In our discussion of the mean ergodic theorem,

- ν_t is the distribution of X_t^x,
- we used Itô's formula to bound $E[f(X_t)] - \mu(f)$ using $E[u_f(X_t)]$ and $\|\nabla u_f\|$,
- ν_t converges to μ, so we were able to show a rate:

$$\frac{1}{T}\int_0^T f(X_t)\,dt \to \mu(f).$$

Optimization using discretized diffusion[5]

[5] We now return to the main paper, Erdogdu et al. (2018).

Recap: goal

Solve the unconstrained optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.

- **Idea:**
  - Construct a diffusion process $(Z_t)_{t \geq 0}$ with stationary distribution p concentrated around the minima of f, so that $p(f) \approx \min_{x \in \mathbb{R}^d} f(x)$.
  - Show that the time-average quickly converges:

$$\frac{1}{T} \int_0^T f(Z_t)\, dt \to p(f),$$

    where the rate depends on the regularity (Stein factors) of u_f solving Poisson equation:

$$\mathcal{A} u_f = f - p(f).$$

  - Show that discretized dynamics $(X_n)_{n=0}^{\infty}$ behaves similarly to $(Z_t)_{t \geq 0}$.

This paper: results

$$\min_{m\in[M]} \mathbb{E}[f(X_m)] - \min_x f(x) \leq \underbrace{\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[f(X_m) - p(f)]}_{\text{integration error}} + \underbrace{p(f) - \min_x f(x)}_{\text{expected suboptimality}} .$$

- Provides bounds on integration error
  - Rate controlled by Stein factors; they provide general bounds for Stein factors
- Provides bounds on suboptimality gap
- Gives examples of optimizing ‘heavy-tailed’ objectives with general diffusion
  - Cases that fail with standard Langevin dynamics with constant diffusion coefficients

Constructing an invariant distribution

Let $(Z_t)_{t\geq 0}$ be Itô diffusion for the SDE: $dZ_t = b(Z_t)\,dt + \sigma(Z_t)\,dB_t$.

Theorem (Invariant measure, Gorham et al. (2019))

A density p is an invariant measure of the diffusion $(Z_t)_{t\geq 0}$ if and only if:

$$b(x) = \frac{1}{2p(x)} \left\langle \nabla, p(x)\big(\mathbf{D}(x) + \mathbf{C}(x)\big) \right\rangle,$$

where $\mathbf{D}(x) = \sigma(x)\sigma(x)^\top$ is the covariance coefficient, $\mathbf{C}(x) \in \mathbb{R}^{d\times d}$ is a differentiable skew-symmetric stream coefficient, and $\langle \nabla, \mathbf{M} \rangle$ is a row-wise divergence operator.

Example: Gibbs measure

Recall the Gibbs measure: $p_\gamma(x) \propto \exp\big(-\gamma f(x)\big)$.

- Set $\sigma(x) = \sqrt{\frac{2}{\gamma}} I$ and $c(x) = 0$.
- Solve for $b(x)$. Note that $\mathbf{D} + \mathbf{C}$ is diagonal. So,

$$b(x)_i = \frac{1}{2p(x)} \operatorname{div}\Big(p(x)\big(\mathbf{D}(x) + \mathbf{C}(x)\big)_i\Big)$$
$$= \frac{1}{2p_\gamma(x)} \cdot \left(\frac{2}{\gamma} \frac{\partial p_\gamma(x)}{\partial x_j}\right) = \frac{1}{\gamma p_\gamma(x)} \frac{\partial p_\gamma(x)}{\partial x_j}.$$

- This implies $b(x) = -\nabla f(x)$ and $dZ_t = -\nabla f(x) + \sqrt{\frac{2}{\gamma}}\, dB_t$.

Assumption I: existence and uniqueness of SDE

Again, we consider the SDE: $dZ_t = b(Z_t)\,dt + \sigma(Z_t)\,dB_t$.

Assumption (Polynomial growth of coefficients)

The drift and diffusion coefficients have growth bounded above:

$$\|b(x)\|_2 + \|\sigma(x)\|_2 \le C(1 + \|x\|_2) \quad \textit{and} \quad \|\mathbf{D}(x)\|_{\mathrm{op}} \le C'(1 + \|x\|_2^2).$$

- This ensures existence and uniqueness of the SDE.

Assumption II: diffusion does not diverge

Assumption (Dissipativity)

There exists $\alpha, \beta > 0$ *such that:*

$$\mathcal{A}\|x\|_2^2 \leq -\alpha\|x\|_2^2 + \beta.$$

- This ensures the diffusion travels inward when far from the origin.
  - Recall that the infinitesimal expected change in $\|X_t\|^2$ at time t is $\mathcal{A}\|x\|^2$.
  - If X_t is large, then $\|X_{t+h}\|^2$ is expected to take a large step inward:

$$E[\|X_{t+h}\|^2|X_t] \approx X_t - \alpha h\|X_t\|^2 + \beta h.$$

  - Dissipativity relaxes the condition that f is strongly convex (c.f. mixture of Gaussians).

Assumption III: rate of convergence bounded

Assumption (Finite Stein factors)

The function u_f that solves the Poisson equation:

$$\mathcal{A}u_f = f - p(f)$$

Lipschitz and has higher order derivatives with polynomial growth,

$$\|\nabla^i u_f(x)\|_{\text{op}} \leq C_i(1 + \|x\|^n) \quad i = 1, 2, 3, 4.$$

Applications of assumption

Recall from our discussion on the mean ergodic theorem:

$$\frac{1}{t}\int_0^t f(Z_s)\,ds - p(f) = \frac{u_f(Z_t) - u_f(Z_0)}{t} - \frac{1}{t}\int_0^t (\sigma^\top \nabla u_f)(Z_s)\,dB_s.$$

- **Dissipativity**: ensures that $\dfrac{u_f(Z_t) - u_f(Z_0)}{t} \to 0$ as $t \to \infty$
- **Stein factors + dissipativity**: ensures that $\|(\sigma^\top \nabla u_f)(Z_s)\| \leq M$, so that:

$$\mathbb{E}\left[\left(\frac{1}{T}\int_0^T f(Z_t)\,dt - p(f)\right)^2\right] \sim \frac{M^2 \mathrm{Var}(B_t)}{t^2} = O\left(\frac{1}{t}\right).$$

Discretization of diffusion

Given the SDE: $dZ_t = b(Z_t)\,dt + \sigma(Z_t)\,dB_t$.

- The **Euler discretization** of $(Z_t)_{t\geq 0}$ corresponds to the process:

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

  where $0 < \eta < 1$ is the **step size** and $W_m \sim \mathcal{N}(0, I)$ is Gaussian noise.

Integration error bound

Theorem (Integration error of discretization)

Let Assumptions I, II, III hold. Then:

$$\left| \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[f(X_m)] - p(f) \right| = O\left(\frac{1}{\eta M} + \eta \right).$$

- The constants depend on the Stein factors and Lipschitz coefficients.
- To reach ε-closeness, need $O(\varepsilon^{-2})$ steps.

Expected suboptimality bound

Theorem (Suboptimality gap)

Suppose p is the stationary density of a dissipative diffusion with global maximizer x^*. If p is of the form $p_{\gamma,\theta} \propto \exp\big(-\gamma(f(x) - f(x^*))^\theta\big)$, and $\nabla f(x^*) = 0$, then:

$$p(f) - f(x^*) = O\left(\frac{1}{\theta}\frac{d}{\gamma}\log\frac{\gamma}{d}\right)^{1/\theta}$$

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Springer Science & Business Media, 2010.

Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. *arXiv preprint arXiv:1810.12361*, 2018.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, Lester Mackey, et al. Measuring sample quality with diffusions. *Annals of Applied Probability*, 29(5):2884–2928, 2019.

Jonathan C Mattingly, Andrew M Stuart, and Michael V Tretyakov. Convergence of numerical time-averaging and stationary measures via poisson equations. *SIAM Journal on Numerical Analysis*, 48 (2):552–577, 2010.

Bernt Øksendal. *Stochastic differential equations: an introduction with applications.* Springer, 2003.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.