Langevin Dynamics

A Continuous and Discrete Analysis

D. Cortild, F. Voronine

Faculty of Science and Engineering
University of Groningen

BB 5161.0165, January 23rd, 2024

university of groningen / faculty of science and engineering / mathematics and applied mathematics

Reminder of Probability Theory

Let X be a random variable on a probability space $(\mathbb{R}^n, \mathcal{B}, \mathbb{P})$.

Reminder of Probability Theory

Let X be a random variable on a probability space $(\mathbb{R}^n, \mathcal{B}, \mathbb{P})$.

- The **probability density** of the random variable $X \sim \mu$ is a function $\rho : \mathbb{R}^n \to \mathbb{R}$ such that

$$\int_A \rho(x)dx = \mathbb{P}(X \in A) = \mu(A).$$

Reminder of Probability Theory

Let X be a random variable on a probability space $(\mathbb{R}^n, \mathcal{B}, \mathbb{P})$.

- The **probability density** of the random variable $X \sim \mu$ is a function $\rho : \mathbb{R}^n \to \mathbb{R}$ such that

$$\int_A \rho(x)dx = \mathbb{P}(X \in A) = \mu(A).$$

- Various notions of **statistical distance** can be defined on the space of probability measures. Examples include the Wasserstein 2-norm, the χ^2-divergence and the KL-divergence

Reminder of Probability Theory

Let X be a random variable on a probability space $(\mathbb{R}^n, \mathcal{B}, \mathbb{P})$.

- The **probability density** of the random variable $X \sim \mu$ is a function $\rho : \mathbb{R}^n \to \mathbb{R}$ such that

$$\int_A \rho(x)dx = \mathbb{P}(X \in A) = \mu(A).$$

- Various notions of **statistical distance** can be defined on the space of probability measures. Examples include the Wasserstein 2-norm, the χ^2-divergence and the KL-divergence

$$W_2(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi \|x-y\|^2\right)^{\frac{1}{2}}, \quad \chi^2(\mu,\nu) = \mathrm{Var}_\nu\left(\frac{\rho_\mu}{\rho_\nu}\right),$$

$$\mathrm{KL}(\mu,\nu) = \mathbb{E}_\mu \log\left(\frac{\rho_\mu}{\rho_\nu}\right).$$

Reminder of Stochastic Processes

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.
- **Brownian motion** is a stochastic process $(B_t)_{t \geq 0}$ such that

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.
- **Brownian motion** is a stochastic process $(B_t)_{t \geq 0}$ such that $B_0 = 0$,

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.
- **Brownian motion** is a stochastic process $(B_t)_{t\geq 0}$ such that $B_0 = 0$, (B_t) is almost surely continuous,

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.
- **Brownian motion** is a stochastic process $(B_t)_{t \geq 0}$ such that $B_0 = 0$, (B_t) is almost surely continuous, (B_t) has independent increments

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.
- **Brownian motion** is a stochastic process $(B_t)_{t\geq 0}$ such that $B_0 = 0$, (B_t) is almost surely continuous, (B_t) has independent increments and $B_{t+s} - B_t \sim \mathcal{N}(0, s)$.

Reminder of Stochastic Processes

- A **stochastic process** is a sequence (possibly continuous in time) of random variables. We denote these sequences by $(X_t)_t$, where t is the time coordinate.
- **Brownian motion** is a stochastic process $(B_t)_{t \geq 0}$ such that $B_0 = 0$, (B_t) is almost surely continuous, (B_t) has independent increments and $B_{t+s} - B_t \sim \mathcal{N}(0, s)$.

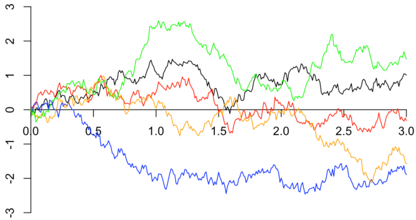

Figure: Brownian Motion. Image from the internet.

Langevin Dynamics

Langevin Dynamics

Langevin dynamics are governed by the stochastic differential equation

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t,$$

where $U : \mathbb{R}^n \to \mathbb{R}$ is a differentiable potential function.

Langevin Dynamics

Langevin dynamics are governed by the stochastic differential equation

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t,$$

where $U : \mathbb{R}^n \to \mathbb{R}$ is a differentiable potential function.

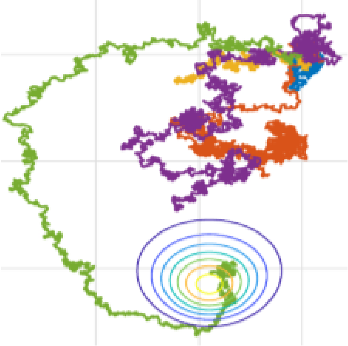

Figure: Particule following the Langevin Dynamics. Image from the internet.

Discretising the Langevin Dynamics

Consider the **Exact Langevin Dynamics** with initial condition $X_0 \sim \mu_0$

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t.$$

Discretising the Langevin Dynamics

Consider the **Exact Langevin Dynamics** with initial condition $X_0 \sim \mu_0$

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t.$$

Its Euler-Murayama discretization **Discretized Langevin Dynamics** with initial condition $Y_0 \sim \mu_0$ and step size δ is given by

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Discretising the Langevin Dynamics

Consider the **Exact Langevin Dynamics** with initial condition $X_0 \sim \mu_0$

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t.$$

Its Euler-Murayama discretization **Discretized Langevin Dynamics** with initial condition $Y_0 \sim \mu_0$ and step size δ is given by

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

The fully iterative **Langevin MCMC Algorithm** with initial condition $Z^0 \sim \mu_0$ and step size δ is then written as

$$Z^{k+1} = Z^k - \delta\nabla U(Z^k)dt + \sqrt{2\delta}\xi^k \quad \text{where } \xi^k \sim \mathcal{N}(0,1).$$

Discretising the Langevin Dynamics

Consider the **Exact Langevin Dynamics** with initial condition $X_0 \sim \mu_0$

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t.$$

Its Euler-Murayama discretization **Discretized Langevin Dynamics** with initial condition $Y_0 \sim \mu_0$ and step size δ is given by

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

The fully iterative **Langevin MCMC Algorithm** with initial condition $Z^0 \sim \mu_0$ and step size δ is then written as

$$Z^{k+1} = Z^k - \delta\nabla U(Z^k)dt + \sqrt{2\delta}\xi^k \quad \text{where } \xi^k \sim \mathcal{N}(0,1).$$

Since $B_{(k+1)\delta} - B_{t\delta} \sim \sqrt{\delta}\mathcal{N}(0,1)$, it holds that Z^k and $Y_{k\delta}$ are equivalent.

Convergence Result in Discrete Setting

Discretized Langevin Dynamics:

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Convergence Result in Discrete Setting

Discretized Langevin Dynamics:

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Theorem (Strong Convexity Result[a])

[a] Xiang Cheng and Peter Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory*. PMLR. 2018, pp. 186–211.

Suppose $U : \mathbb{R}^d \to \mathbb{R}$ *is strongly convex and smooth.*

Convergence Result in Discrete Setting

Discretized Langevin Dynamics:

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Theorem (Strong Convexity Result[a])

[a] Xiang Cheng and Peter Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory*. PMLR. 2018, pp. 186–211.

Suppose $U: \mathbb{R}^d \to \mathbb{R}$ is strongly convex and smooth. Under mild conditions on μ_0, δ, k and $\varepsilon > 0$, $\mathrm{KL}(\mu_{k\delta}, \mu^*) \leq \varepsilon$,

Convergence Result in Discrete Setting

Discretized Langevin Dynamics:

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Theorem (Strong Convexity Result[a])

[a] Xiang Cheng and Peter Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory*. PMLR. 2018, pp. 186–211.

Suppose $U: \mathbb{R}^d \to \mathbb{R}$ is strongly convex and smooth. Under mild conditions on μ_0, δ, k and $\varepsilon > 0$, $\mathrm{KL}(\mu_{k\delta}, \mu^*) \leq \varepsilon$, where μ^* is the stationary distribution of the **Exact Langevin Dynamics**

Convergence Result in Discrete Setting

Discretized Langevin Dynamics:

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Theorem (Strong Convexity Result[a])

[a] Xiang Cheng and Peter Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory*. PMLR. 2018, pp. 186–211.

Suppose $U : \mathbb{R}^d \to \mathbb{R}$ is strongly convex and smooth. Under mild conditions on μ_0, δ, k and $\varepsilon > 0$, $\mathrm{KL}(\mu_{k\delta}, \mu^*) \leq \varepsilon$, where μ^* is the stationary distribution of the **Exact Langevin Dynamics** *and $\rho_{k\delta}$ is the distribution of $Y_{k\delta}$ given by the* **Discretized Langevin Dynamics**.

Convergence Result in Discrete Setting

Discretized Langevin Dynamics:

$$dY_t = -\nabla U(Y_{k\delta})dt + \sqrt{2}dB_t \quad \text{for } t \in [k\delta, (k+1)\delta].$$

Theorem (Strong Convexity Result[a])

[a] Xiang Cheng and Peter Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory*. PMLR. 2018, pp. 186–211.

Suppose $U : \mathbb{R}^d \to \mathbb{R}$ *is strongly convex and smooth. Under mild conditions on* μ_0, δ, k *and* $\varepsilon > 0$, $\mathrm{KL}(\mu_{k\delta}, \mu^*) \leq \varepsilon$, *where* μ^* *is the stationary distribution of the* **Exact Langevin Dynamics** *and* $\rho_{k\delta}$ *is the distribution of* $Y_{k\delta}$ *given by the* **Discretized Langevin Dynamics**.

Note that since Z^k and $Y_{k\delta}$ are equivalent, the above also proves convergence in distribution of the **Langevin MCMC Algorithm**.

Continuous Results

In the deterministic case (zero noise), the Langevin Dynamics boil down to

$$\frac{dX(t)}{dt} = -\nabla U(X(t)).$$

Continuous Results

In the deterministic case (zero noise), the Langevin Dynamics boil down to

$$\frac{dX(t)}{dt} = -\nabla U(X(t)).$$

Let ρ_t denote the density of X_t. The Langevin Dynamics may be rewritten using the Fokker-Planck equation as

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla U) + \Delta \rho_t.$$

Continuous Results

In the deterministic case (zero noise), the Langevin Dynamics boil down to

$$\frac{dX(t)}{dt} = -\nabla U(X(t)).$$

Let ρ_t denote the density of X_t. The Langevin Dynamics may be rewritten using the Fokker-Planck equation as

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla U) + \Delta \rho_t.$$

Theorem (Convergence in Continuous Time[a])

[a] Stephen Tu. "On the exponential convergence of Langevin diffusions: from deterministic to stochastic stability". In: *GitHub* (2022).

If there exists a suitable Lyapunov function for the deterministic system, then, for any initial measure, we achieve linear convergence towards the stationary distribution (in the χ^2-divergence).

JKO Scheme

JKO Scheme

Jordan, Kinderlehrer and Otto established the following algorithm

$$\rho_\delta^{k+1} = \underset{\rho}{\operatorname{argmin}} \left\{ \delta F(\rho) + \frac{1}{2} W_2(\rho_\delta^k, \rho)^2 \right\},$$

where $\delta > 0$ is a fixed step size, F is the free energy functional depending on the potential U.

JKO Scheme

Jordan, Kinderlehrer and Otto established the following algorithm

$$\rho_\delta^{k+1} = \underset{\rho}{\operatorname{argmin}} \left\{ \delta F(\rho) + \frac{1}{2} W_2(\rho_\delta^k, \rho)^2 \right\},$$

where $\delta > 0$ is a fixed step size, F is the free energy functional depending on the potential U.

Theorem (Strong Convergence to the True Solution[a])

[a] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998).

As $\delta \to 0$, the JKO iterations converge strongly to the unique true solution of the Fokker-Planck equation in L^1-norm .

JKO Scheme

Jordan, Kinderlehrer and Otto established the following algorithm

$$\rho_\delta^{k+1} = \operatorname*{argmin}_\rho \left\{ \delta F(\rho) + \frac{1}{2} W_2(\rho_\delta^k, \rho)^2 \right\},$$

where $\delta > 0$ is a fixed step size, F is the free energy functional depending on the potential U.

Theorem (Strong Convergence to the True Solution[a])

[a] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998).

As $\delta \to 0$, the JKO iterations converge strongly to the unique true solution of the Fokker-Planck equation in L^1-norm .

Note that under the assumption on the previous slide, the true solution converges to the stationary distribution.

Wasserstein Gradient Flows

The space of probability measures can be equipped with a differentiable structure through **Otto calculus**.

[1] Marc Lambert et al. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14434–14447.

Wasserstein Gradient Flows

The space of probability measures can be equipped with a differentiable structure through **Otto calculus**.

Example

For $F(\rho_t) = \mathrm{KL}(\rho_t, \rho)$, we have $\nabla_W F(\rho_t) = \nabla \log \frac{\rho_t}{\rho}(\cdot)$.
For $F(\rho_t) = \chi^2(\rho_t, \rho)$, we have $\nabla_W F(\rho_t) = 2\nabla \frac{\rho_t}{\rho}(\cdot)$.

[1] Marc Lambert et al. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14434–14447.

Wasserstein Gradient Flows

The space of probability measures can be equipped with a differentiable structure through **Otto calculus**.

Example

For $F(\rho_t) = \mathrm{KL}(\rho_t, \rho)$, we have $\nabla_W F(\rho_t) = \nabla \log \frac{\rho_t}{\rho}(\cdot)$.
For $F(\rho_t) = \chi^2(\rho_t, \rho)$, we have $\nabla_W F(\rho_t) = 2\nabla \frac{\rho_t}{\rho}(\cdot)$.

The JKO scheme can be viewed as a discretisation of the following deterministic gradient flow in Wasserstein space[1]

$$\dot{X}_t = -\nabla_W F(\rho_t)(X_t).$$

[1] Marc Lambert et al. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14434–14447.

Wasserstein Gradient Flows

The space of probability measures can be equipped with a differentiable structure through **Otto calculus**.

Example

For $F(\rho_t) = \mathrm{KL}(\rho_t, \rho)$, we have $\nabla_W F(\rho_t) = \nabla \log \frac{\rho_t}{\rho}(\cdot)$.
For $F(\rho_t) = \chi^2(\rho_t, \rho)$, we have $\nabla_W F(\rho_t) = 2\nabla \frac{\rho_t}{\rho}(\cdot)$.

The JKO scheme can be viewed as a discretisation of the following deterministic gradient flow in Wasserstein space[1]

$$\dot{X}_t = -\nabla_W F(\rho_t)(X_t).$$

In this case, standard optimisation techniques such as Gradient Descent, Newton Method, Proximal Method, etc. apply.

[1] Marc Lambert et al. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14434–14447.

Thank you for your attention !

Any questions ?