

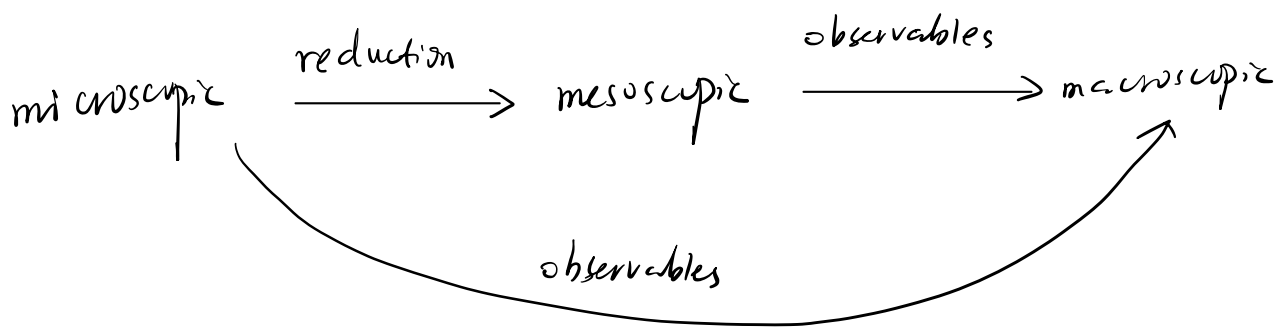
An introduction to Stochastic Dynamics

Zhongwei Shen zhongwei@ualberta.ca

What is "stochastic dynamics"?

- complex behaviors over microscopic scales
- behaviors over mesoscopic scales
- simple behaviors over macroscopic scales

stochastic dynamics



Tools :

probability
stochastic process
stochastic analysis

PDE

dynamical system

Overdamped Langevin equation

Langevin equation

$$\dot{x}_t = -\nabla V(x_t) + \sqrt{2\beta^{-1}} \dot{W}_t \quad \text{in } \mathbb{R}^d$$

$$m\ddot{x}_t = -\nabla V(x_t) - \gamma \dot{x}_t + \sqrt{2\gamma\beta^{-1}} \dot{W}_t$$

scaling limit with $m=1$
 $\gamma \gg 1$

$$dx_t = -\nabla V(x_t) dt + \sqrt{2\beta^{-1}} dW_t$$

$$x_t = x_0 - \int_0^t \nabla V(x_s) ds + \sqrt{2\beta^{-1}} W_t$$

It describes the motion of a particle that moves according to Newton's second law $ma=F$ and is in contact with a thermal reservoir with temperature β^{-1}

W_t is a Brownian motion, or Wiener process

- $W_0 = 0$

- independent increment: for $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_n < t_n < \infty$

$$W_{t_1} - W_{s_1}, W_{t_2} - W_{s_2}, \dots, W_{t_n} - W_{s_n}$$

are jointly independent.

- Gaussian increment: for $t > s \geq 0$, $W_t - W_s \sim \mathcal{N}(0, (t-s)I_{d \times d})$

• stationary distribution π :

$$x_0 \sim \mu \Rightarrow x_t \sim \mu \text{ for all } t > 0.$$

it has a unique SD with density π :

$$\pi(x) = \frac{e^{-\beta V(x)}}{\int_{\mathbb{R}^d} e^{-\beta V(x)} dx}$$

growth condition on V is required to ensure the integrability

How to find it? Need to solve the stationary Fokker-Planck equation

Transition probability density:

$$\mathbb{P}\{x_t \in B \mid x_0 = x\} = \int_B p(t, x, y) dy, \quad \forall B \in \mathcal{B}(\mathbb{R}^d)$$

and it satisfies the Fokker-Planck equation

$$\begin{cases} \partial_t p(t, x, y) = \frac{1}{\beta} \Delta_y p(t, x, y) + \nabla_y \cdot (p(t, x, y) \nabla_y V(y)) \\ p(0, x, y) = \delta(y - x) \end{cases}$$

If $x_0 \sim \mu$, then $p_\mu(t, y) \triangleq \int_{\mathbb{R}^d} p(t, x, y) d\mu(x)$ is the density of the distribution of x_t , and satisfies

$$\begin{cases} \partial_t p(t, y) = \frac{1}{\beta} \Delta p(t, y) + \nabla \cdot (p(t, y) \nabla V(y)) \\ p(0, y) = \mu \end{cases}$$

So, if μ is a SD, then $p(t, y)$ would be independent of t , and therefore, $p(t, y) = p(y)$ solves

$$\frac{1}{\beta} \Delta p + \nabla \cdot (p \nabla V) = 0. \quad (\text{or } \nabla \cdot (\frac{1}{\beta} \nabla p + p \nabla V) = 0).$$

Note solving $\frac{1}{\beta} \nabla p + p \nabla V = 0$ suffices to find a solution.

$$\nabla \ln p = -\nabla V$$

$$\Rightarrow p \approx e^{-\beta V}$$

$$\Rightarrow \pi = \frac{e^{-\beta V}}{\int_{\mathbb{R}^d} e^{-\beta V}}$$

normalization

Therefore, π defines a SD of $(*)$.

Fundamental theory guarantees the uniqueness in this case.

- exponential convergence to π

Let $p(t, x)$ be the solution of

$$\begin{cases} \partial_t p = \frac{1}{\beta} \Delta p + \nabla \cdot (p \nabla V) \\ p(0, x) \sim \mu. \end{cases}$$

Recall π satisfies $\frac{1}{\beta} \nabla \pi + \pi \nabla V = 0$. Then,

$$\begin{aligned} \partial_t p &= \nabla \cdot \left(\frac{1}{\beta} \nabla p + p \nabla V \right) \\ &= \nabla \cdot \left[p \left(\frac{1}{\beta} \nabla \ln p + \nabla V \right) \right] \\ &= \frac{1}{\beta} \nabla \cdot \left[p (\nabla \ln p - \nabla \ln \pi) \right] \\ &= \frac{1}{\beta} \nabla \cdot \left[p \nabla \ln \frac{p}{\pi} \right] \end{aligned}$$

$L^2(\pi^{-1})$
||

Let compute the distance between $p(t, \cdot)$ and π in $L^2(\mathbb{R}^d, \pi^{-1}(x) dx)$

$$\frac{d}{dt} \|p(t, \cdot) - \pi\|_{L^2(\pi^{-1})}^2 = \frac{d}{dt} \int_{\mathbb{R}^d} |p(t, x) - \pi(x)|^2 \pi^{-1}(x) dx$$

$$= 2 \int_{\mathbb{R}^d} \frac{p(t, x) - \pi(x)}{\pi(x)} p_t(t, x) dx$$

$$= \frac{2}{\beta} \int_{\mathbb{R}^d} \frac{p(t, x) - \pi(x)}{\pi(x)} \nabla \cdot \left[p(t, x) \nabla \ln \frac{p(t, x)}{\pi(x)} \right] dx$$

$$= -\frac{2}{\beta} \int_{\mathbb{R}^d} \nabla \frac{p(t, x)}{\pi(x)} \cdot p(t, x) \nabla \ln \frac{p(t, x)}{\pi(x)} dx$$

$$= -\frac{2}{\beta} \int_{\mathbb{R}^d} \left| \nabla \frac{p(t, x)}{\pi(x)} \right|^2 \pi(x) dx.$$

That is, we obtained

$$\frac{d}{dt} \left\| p(t, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2 = -\frac{2}{\beta} \int_{\mathbb{R}^d} \left| \nabla \frac{p(t, x)}{\pi(x)} \right|^2 \pi(x) dx.$$

Poincaré inequality of π : there's $\lambda = \lambda(\beta) > 0$ s.t.

$$\int_{\mathbb{R}^d} |f(x) - 1|^2 \pi(x) dx \leq \frac{1}{\lambda} \int_{\mathbb{R}^d} |\nabla f(x)|^2 \pi(x) dx$$

for all $f: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $f\pi$ is a prob density function.

- A sufficient condition for the validity of Poincaré inequality is that there are $a \in (0, 1)$, $c > 0$ and $R > 0$ such that $a\beta |\nabla V(x)|^2 - \Delta V(x) \geq c$ for all $|x| \geq R$.

What is the optimal/maximal λ ?

Applying Poincaré inequality with $f(x) = \frac{p(t, x)}{\pi(x)}$, we arrive at

$$\int_{\mathbb{R}^d} \left| \frac{p(t, x)}{\pi(x)} - 1 \right|^2 \pi(x) dx \leq \frac{1}{\lambda} \int_{\mathbb{R}^d} \left| \nabla \frac{p(t, x)}{\pi(x)} \right|^2 \pi(x) dx.$$

It follows that

$$\begin{aligned} \frac{d}{dt} \left\| p(t, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2 &\leq -\frac{2\lambda}{\beta} \int_{\mathbb{R}^d} \left| \frac{p(t, x)}{\pi(x)} - 1 \right|^2 \pi(x) dx \\ &= -\frac{2\lambda}{\beta} \int_{\mathbb{R}^d} |p(t, x) - \pi(x)|^2 \pi^{-1}(x) dx \\ &= -\frac{2\lambda}{\beta} \left\| p(t, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2 \end{aligned}$$

Namely,

$$\frac{d}{dt} \left\| p(t, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2 \leq -\frac{2\lambda}{\beta} \left\| p(t, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2.$$

Hence,

$$\left\| p(t, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2 \leq e^{-\frac{2\lambda}{\beta} t} \left\| p(0, \cdot) - \pi \right\|_{L^2(\pi^{-1})}^2, \quad \forall t \geq 0.$$

- Euler-Maruyama numerical scheme

time step Δt : $0 \quad \Delta t \quad 2\Delta t \quad 3\Delta t \quad \dots$

$$x_{(j+1)\Delta t} = x_{j\Delta t} - \underbrace{\int_{j\Delta t}^{(j+1)\Delta t} \nabla V(x_s) ds}_{\nabla V(x_{j\Delta t}) \Delta t} + \underbrace{\sqrt{2\beta^{-1}} (W_{(j+1)\Delta t} - W_{j\Delta t})}_2$$

$$\mathcal{N}(0, \Delta t I_{d \times d})$$

$$\parallel \sqrt{\Delta t} \xi \text{ with } \xi \sim \mathcal{N}(0, I_{d \times d})$$

Let $X_j = x_{j\Delta t}$. Then,

$$X_{j+1} = X_j - \nabla V(X_j) \Delta t + \sqrt{2\beta^{-1}} \sqrt{\Delta t} \xi_j, \quad j = 0, 1, 2, \dots$$

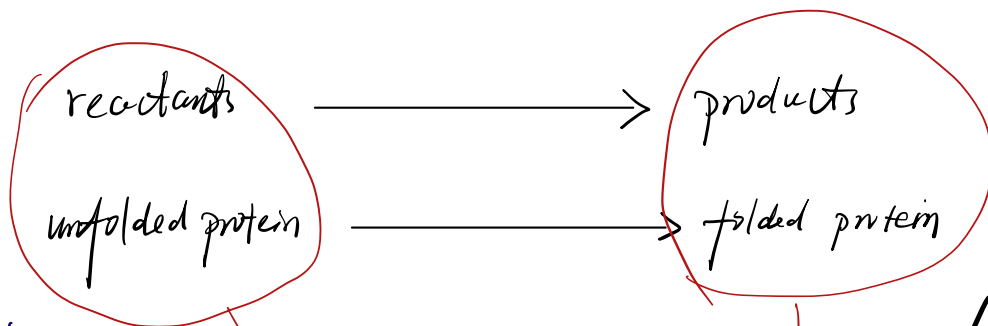
where $\{\xi_j\}_j$ are i.i.d $\mathcal{N}(0, I_{d \times d})$.

A random-number generator can be used to generate i.i.d Gaussian random variables.

Exercise: try and look at the histogram at time $j\Delta t$, $j=1, 2, 3, \dots$ to see whether SD can be observed.

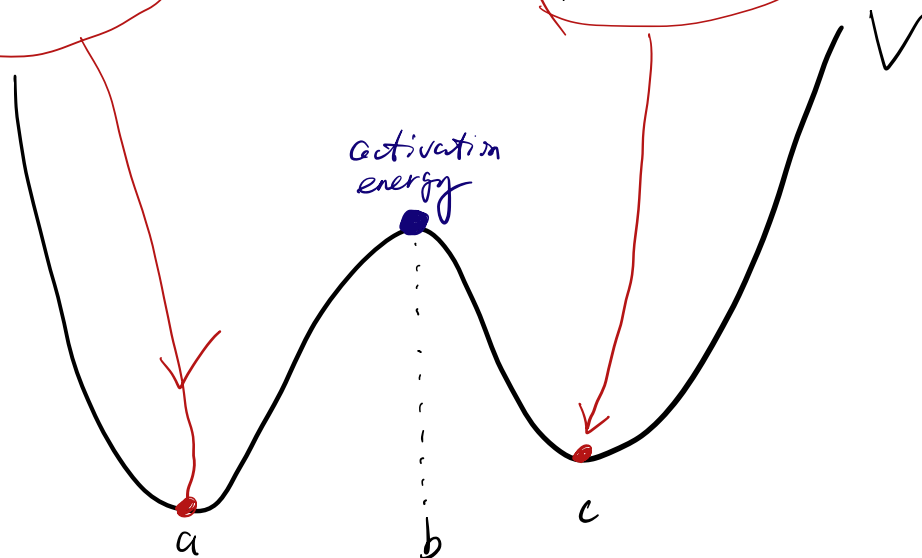
Reaction rate theory and rare events

chemical reaction:



the dynamical process
is very complicated
as a large number of atoms
and molecules are involved

representation by potential
energy functions



reaction rate: how often the chemical reaction happens

$$\frac{1}{\text{reaction rate}} = \text{mean exit time } (a \rightarrow b)$$

Dynamical equation:

$$\dot{x}_t^\beta = -V'(x_t^\beta) + \sqrt{2\beta^{-1}} \dot{W}_t, \quad x \in \mathbb{R}, \quad \beta \gg 1$$

For $x_0^\beta = x \in (-\infty, b)$, x_t^β exits $(-\infty, b)$ with probability 1.

Let τ^β be the first exit time from $(-\infty, b)$. Then, its mean $\tau^\beta(x) := \mathbb{E}_x[\tau^\beta]$ satisfies

$$\begin{cases} \frac{1}{\beta} (\tau^\beta)'' - V'(\tau^\beta)' = -1 \\ \tau^\beta(-\infty) = 0, \quad \tau^\beta(b) = 0. \end{cases}$$

$$(e^{-\beta V} (\tau^\beta)')' = -\beta e^{-\beta V}$$

$$\Rightarrow e^{-\beta V(y)} (\tau^\beta)'(y) = -\beta \int_{-\infty}^y e^{-\beta V(z)} dz$$

$$\Rightarrow (\tau^\beta)'(y) = -\beta e^{\beta V(y)} \int_{-\infty}^y e^{-\beta V(z)} dz$$

integration over $[a, b]$

$$\Rightarrow \tau^\beta(x) = \beta \int_x^b e^{\beta V(y)} \left[\int_{-\infty}^y e^{-\beta V(z)} dz \right] dy, \quad x \in (-\infty, b).$$

Consider $x=a$, namely,

$$\tau^\beta(a) = \beta \int_a^b e^{\beta V(y)} \left[\int_{-\infty}^y e^{-\beta V(z)} dz \right] dy$$

As $\beta \gg 1$, the integral w.r.t z is dominated by V near a .

Taylor's expansion: $V(z) = V(a) + \frac{V''(a)}{2} (z-a)^2 + \dots$

Then,

$$\int_{-\infty}^y e^{-\beta V(z)} dz \approx e^{-\beta V(a)} \int_{-\infty}^{+\infty} e^{-\frac{\beta V''(a)}{2} (z-a)^2} dz = e^{-\beta V(a)} \sqrt{\frac{2\pi}{\beta V''(a)}}$$

Similarly, the integral w.r.t y is dominated by V near b .

Taylor expansion: $V(y) = V(b) + \frac{V''(b)}{2} (y-b)^2 + \dots$

Then,

$$\int_a^b e^{\beta V(y)} dy \approx e^{\beta V(b)} \int_{-\infty}^b e^{\frac{\beta V''(b)}{2} (y-b)^2} dy = \frac{1}{2} e^{\beta V(b)} \sqrt{\frac{2\pi}{-\beta V''(b)}}$$

Hence,

$$\tau^\beta(a) \approx \frac{\pi}{\sqrt{-V''(a)V''(b)}} e^{\beta [V(b) - V(a)]}$$

MCMC and accelerating algorithm

How to sample from a higher-dimensional probability distribution?

The idea of MCMC (Monte-Carlo Markov chain) is to construct a Markov chain (process) whose unique SD is the given prob distribution, then simulate the transition prob to approximate the given prob distribution.

Here, we focus on the diffusion process setting. Discretizing the diffusion process leads to a Markov chain.

Given a prob distribution with a sufficiently smooth ^{and positive} density π .

We can construct a diffusion process (actually, an overdamped Langevin equation)

$$\dot{x}_t = -\nabla V(x_t) + \sqrt{2} \dot{W}_t \quad \text{in } \mathbb{R}^d \quad (*)$$

with $e^{-V} = \pi$, or $V = -\ln \pi$. Then, $\pi = e^{-V}$ defines the unique SD of the above equation. Simulating its transition probability approximates the SD.

The idea of MCMC does not work very well when V has deep wells, resulting in the "very slow" convergence to the SD. This more or less motivates the accelerating algorithm. The idea is to "stir" (*) by adding a vector field $b: \mathbb{R}^d \rightarrow \mathbb{R}^d$, resulting in

explain the meaning!

$$\dot{x}_t = -\nabla V(x_t) + \boxed{b(x_t)} + \sqrt{2} \dot{w}_t \quad (**)$$

The vector field b is not arbitrary. At least, it should not change the SD. That is, $\pi = e^{-V}$ should be the SD of $(**)$ as well.

The stationary Fokker-Planck equation of $(**)$ reads

$$\Delta u + \nabla \cdot (u \nabla V) - \nabla \cdot (u b) = 0$$

Inserting $u = \pi = e^{-V}$ yields

$$\nabla \cdot (e^{-V} b) = 0$$

This is a generalized divergence-free condition. The usual one is

$$\nabla \cdot b = 0$$

Such a vector field b always exists.

For example, $b = S \nabla V$, where

S is an anti-symmetric dxd matrix $S^T = -S$.

This idea has been used in numerical simulations. But, the mathematical theory is not complete.

- (C.-R. Hwang, S.-Y. Hwang-Ma, S.-J. Sheu
Accelerating Diffusions, AAP, 2005)

Let λ_b be the spectral gap of $\mathcal{L}_b$. Then,

$$\lambda_b \geq \lambda_0 \quad (\lambda_b > \lambda_0 ? \text{ not known})$$

when λ_b is an eigenvalue of $-\mathcal{L}_b$, there's a necessary and sufficient condition for $\lambda_b = \lambda_0$.

- (C.-R. Hwang, R. Normand, S.-J. Wu
Variance reduction for diffusions,
Stochastic Processes and their Applications, 2015)

$$\text{LLN: } \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f(x_s^b) ds = \int_{\mathbb{R}^d} f d\pi$$

$$\text{CLT: } \lim_{t \rightarrow \infty} \sqrt{t} \left(\frac{1}{t} \int_0^t f(x_s^b) ds - \int_{\mathbb{R}^d} f d\pi \right) \stackrel{(d)}{=} \mathcal{N}(0, \sigma_b^2(f))$$

$$\text{Then, } \sigma_b^2(f) \leq \sigma_o^2(f).$$

Gradient descent and stochastic gradient descent (GD)

Optimization problem

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

assume $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is continuous and $f(x) \rightarrow \infty$ as $|x| \rightarrow \infty$.

Then, there exists $x_* \in \mathbb{R}^d$, not unique in general, such that

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

Practical problem

How to search for x_* or similar?

Gradient descent (GD)

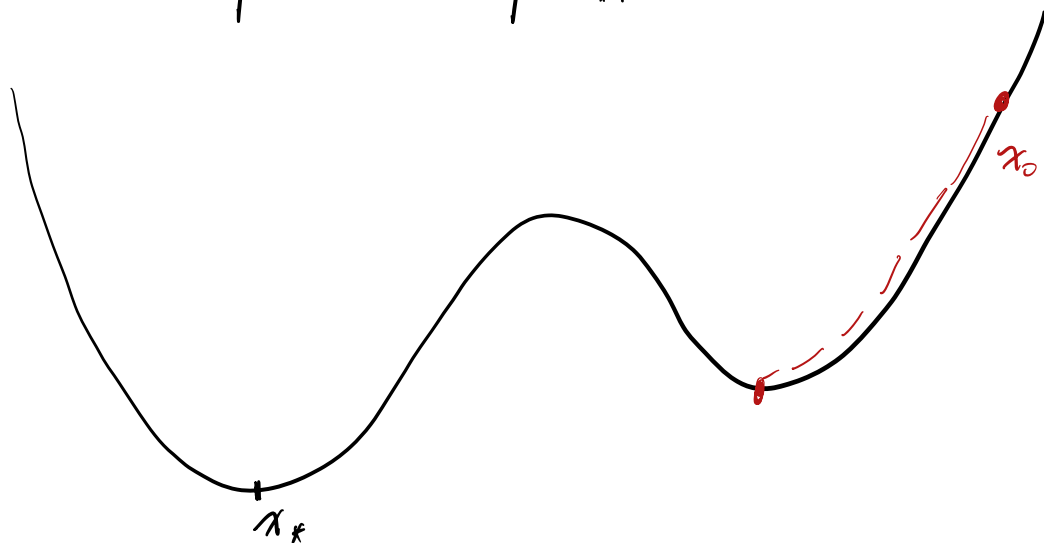
$$\begin{cases} \text{choose } x_0 \\ \text{iterate } x_{j+1} = x_j - s \nabla f(x_j), \quad j=0, 1, 2, \dots \end{cases}$$

When $s = \Delta t$ is a small constant, this algorithm is a discrete-time simulation for the ODE $\dot{X}_t = -\nabla f(X_t)$.

When s is sufficiently small, x_j converges to a local minimum of f .
Therefore, when f is convex, x_* is unique, and GD helps to find x_* .

When f is non-convex, an x_* can not be found in general

unless x_0 is in the basin of attraction of x_* .



Stochastic Gradient Descent

choose x_0

$$\text{iterate } x_{j+1} = x_j - s \nabla f(x_j) + \sqrt{2S} \eta_j$$

where $\{\eta_j\}_j$ i.i.d. $\mathcal{N}(0, I_{d \times d})$

When $s = \Delta t$ and $S = \beta^{-1} \Delta t$, this algorithm becomes a discrete-time simulation of the overdamped Langevin equation

$$X_t = -\nabla f(X_t) + \sqrt{2\beta^{-1}} \dot{W}_t$$