Peking University, CS buzhidao

Notes of Stochastic Analysis

Narsil Zhang

Spring 2019

Contents

1 Stochastic Analysis

1.1 Ito's Formula

To handle Ito's formula one only need to remember

$$dB_t^2 = dt, \quad dB_t dt = 0, \quad (dt)^2 = 0$$

To derive Ito's formula for an Ito process X_t satisfying $dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$ (σ is matrix when X and B are vectors) and arbitrary function $f(t, x)$, remember to perform Taylor's expansion to 2 order:

$$\begin{aligned} df(t, X_t) &= \nabla_t f(t, X_t)dt + \nabla_x f(t, X_t) \cdot dX_t + \frac{1}{2}(dX_t)^T \cdot \nabla_x^2 f(t, X_t) \cdot (dX_t) && (1.1) \\ &= \nabla_t f dt + \nabla_x f \cdot (bdt + \sigma dB_t) + \frac{1}{2}(bdt + \sigma dB_t)^T \cdot \nabla_x^2 f \cdot (bdt + \sigma dB_t) && (1.2) \\ &= \nabla_t f dt + \nabla_x f \cdot (bdt + \sigma dB_t) + \frac{1}{2}(\sigma dB_t)^T \cdot \nabla_x^2 f \cdot \sigma dB_t && (1.3) \end{aligned}$$

Denote $\boldsymbol{A} : \boldsymbol{B} = \sum_{ij} a_{ij} b_{ji}$ we have

$$\begin{aligned} (\boldsymbol{\sigma} d\boldsymbol{B}_t)^T \cdot \nabla^2 f \cdot (\boldsymbol{\sigma} d\boldsymbol{B}_t) &= \sum_{l,k,i,j} dB_t^l \sigma_{il} \partial_{ij}^2 f \sigma_{jk} dB_t^k \\ &= \sum_{k,i,j} \sigma_{ik}\sigma_{jk} \partial_{ij}^2 f dt = \boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f dt \end{aligned}$$

then we have Ito's formula

$$df(t, X_t) = \left(\nabla_t f + \nabla_x f \cdot \boldsymbol{b} + \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f\right) dt + (\nabla_x f)^T \cdot \boldsymbol{\sigma} \cdot d\boldsymbol{B}_t \quad (1.4)$$

For an SDE, sometimes we define an infinite small generator $\mathcal{L}$ as

$$\mathcal{L} f = \nabla_x f \cdot \boldsymbol{b} + \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f$$

1.2 Forward Equation

Here we wish to derive a PDE to describe $p(t, x)$, where for $s < t$

$$p(\boldsymbol{x}, t|\boldsymbol{y}, s)d\boldsymbol{x} = \mathbb{P}\left\{\boldsymbol{X}_t \in [\boldsymbol{x}, \boldsymbol{x} + d\boldsymbol{x}) | \boldsymbol{X}_s = \boldsymbol{y}\right\}$$

For arbitrary $f(x)$, from Ito's formula 1.4 we have

$$\begin{aligned} f(\boldsymbol{X}_t) - f(\boldsymbol{X}_s) &= \int_s^t \nabla f(\boldsymbol{X}_\tau) \cdot \{\boldsymbol{b}(\boldsymbol{X}_\tau, \tau)\, d\tau + \boldsymbol{\sigma}(\boldsymbol{X}_\tau, \tau)\, d\boldsymbol{B}_\tau\} \\ &+ \frac{1}{2}\int_s^t \sum_{i,j} \partial_{ij}^2 f(\boldsymbol{X}_\tau)\, a_{ij}(\boldsymbol{X}_\tau, \tau)\, d\tau \end{aligned}$$

where the diffusion matrix $\boldsymbol{a}(\boldsymbol{x}, t) := \boldsymbol{\sigma}(\boldsymbol{x}, t)\boldsymbol{\sigma}^T(\boldsymbol{x}, t)$.

Because $\int [\cdot] dB_t = 0$ and denote $X_s = y$, take expectation we get

$$\mathbb{E} f(\boldsymbol{X}_t) - f(\boldsymbol{y}) = \mathbb{E}\int_s^t \mathcal{L} f(\boldsymbol{X}_\tau, \tau)\, d\tau$$

Take derivative of time on both side

$$\int_{\mathbb{R}^d} f(\boldsymbol{x}) \cdot \partial_t p(\boldsymbol{x}, t|\boldsymbol{y}, s) d\boldsymbol{x} = \int_{\mathbb{R}^d} \mathcal{L} f(\boldsymbol{x}, \tau) \cdot p(\boldsymbol{x}, \tau|\boldsymbol{y}, s) d\boldsymbol{x}$$

that is

$$(f, \partial_t p)_{L^2} = (\mathcal{L} f, p)_{L^2} = (f, \mathcal{L}^* p)_{L^2}$$

we get

$$\partial_t p = \mathcal{L}^* p$$

This is forward equation (or Fokker-Planck equation), where "forward" means taking derivative of future time t. We can calculate that the adjoint of $\mathcal{L}$ is

$$\mathcal{L}^* f(\boldsymbol{x}, t) = -\nabla_{\boldsymbol{x}} \cdot (\boldsymbol{b}(\boldsymbol{x}, t) f(\boldsymbol{x})) + \frac{1}{2}\nabla_{\boldsymbol{x}}^2 : (\boldsymbol{a}(\boldsymbol{x}, t) f(\boldsymbol{x}))$$

where $\nabla_{\boldsymbol{x}}^2 : (\boldsymbol{a} f) = \sum_{ij} \partial_{ij} \left(a_{ij} f\right)$.

1.3 Backward Equation

For arbitrary $f(x)$ define

$$u(\boldsymbol{y}, s) = \mathbb{E}^{\boldsymbol{y}, s} f(\boldsymbol{X}_t) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) p(\boldsymbol{x}, t|\boldsymbol{y}, s) d\boldsymbol{x}, \quad s \leq t$$

$$du(\boldsymbol{X}_\tau, \tau) = (\partial_\tau u + \mathcal{L} u)(\boldsymbol{X}_\tau, \tau)\, d\tau + \nabla u \cdot \boldsymbol{\sigma} \cdot d\boldsymbol{W}_\tau$$

Taking expectation

$$\begin{aligned} \lim_{t\to s} \frac{1}{t-s}\left(\mathbb{E}^{\boldsymbol{y}, s} u(\boldsymbol{X}_t, t) - u(\boldsymbol{y}, s)\right) &= \lim_{t\to s} \frac{1}{t-s}\int_s^t \mathbb{E}^{\boldsymbol{y}, s} (\partial_\tau u + \mathcal{L} u)(\boldsymbol{X}_\tau, \tau)\, d\tau \\ &= \partial_s u(\boldsymbol{y}, s) + \mathcal{L} u(\boldsymbol{y}, s) \end{aligned}$$

Based on

$$\mathbb{E}^{\boldsymbol{y},s} u\left(\boldsymbol{X}_t, t\right) = \mathbb{E}^{\boldsymbol{y},s} f\left(\boldsymbol{X}_t\right) = u(\boldsymbol{y}, s)$$

we get

$$\partial_s u(\boldsymbol{y}, s) + \mathcal{L} u(\boldsymbol{y}, s) = 0$$

and then

$$\partial_s p(\boldsymbol{x}, t | \boldsymbol{y}, s) + \mathcal{L}_{\boldsymbol{y}} p(\boldsymbol{x}, t | \boldsymbol{y}, s) = 0$$

This is backward equation. Backward means taking derivative of previous time s.

2 SDE & MCMC

2.1 SGLD

Let's say we want to sample from $p(\theta|X) = p(\theta)p(X|\theta)/Z$. We hope $p(\theta|X)$ is a stationary distribution of SDE $d\theta_t = b(\theta_t)dt + \sigma(\theta_t)dW_t$. The forward equation is

$$\frac{\partial \mu_t}{\partial t} = -\nabla_\theta \cdot (\mu_t b(\theta_t)) + \frac{1}{2}\nabla_\theta^2 : (\sigma\sigma^T \mu_t)$$

If $\mu_t = \mu_t(\theta_t)$ is the stationary distribution, then $\frac{\partial \mu_t}{\partial t} = 0$.

Now we set

$$U(\theta) = logp(X|\theta) + logp(\theta)$$

(which means $p(\theta|X) = exp(U(\theta))/Z$),
$b(\theta) = \frac{1}{2}\nabla_\theta U(\theta)$ and $\sigma(\theta) = I$, then $\frac{\partial \mu_t}{\partial t} = -\nabla_\theta \cdot (\frac{1}{2}\mu_t \nabla U(\theta_t)) + \frac{1}{2}\nabla_\theta \cdot (\nabla \mu_t) = 0 \Rightarrow$

$$\nabla_\theta \cdot (\mu_t \nabla_\theta U(\theta)) = \nabla_\theta \cdot (\nabla_\theta \mu_t) \tag{2.1}$$

$$\mu_t \nabla_\theta U(\theta) = \nabla_\theta \mu_t \tag{2.2}$$

$$\mu_t \propto exp(U(\theta)) \propto p(\theta|X) \tag{2.3}$$

Now we can discrete the SDE ($d\theta_t = \frac{1}{2}\nabla_\theta U(\theta_t)dt + dW_t$) to simulate $p(\theta|X)$! (And we only need to use the score function $\nabla_\theta U(\theta) = \nabla_\theta logp(\theta|X)$.)

2.2 Stochastic Gradient Hamiltonian MC

Still we define $U(\theta) = logp(X|\theta) + logp(\theta)$ and $H(\theta, r) = U(\theta) + \frac{1}{2}r^T M^{-1} r$. Introduce an augment variable r and, define a joint distribution $\pi(\theta, r) \propto \exp(-H(\theta, r))$.

Now let's dive into this 2-order SDE:

$$d\begin{bmatrix} \theta \\ r \end{bmatrix} = -\begin{bmatrix} 0 & -I \\ I & B \end{bmatrix}\begin{bmatrix} \nabla U(\theta) \\ M^{-1}r \end{bmatrix} dt + \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2B} \end{bmatrix} dW_t \tag{2.4}$$

$$= -[D + G]\nabla H(\theta, r)dt + \mathcal{N}(0, 2Ddt) \tag{2.5}$$

where $G = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix}$.

Its Fokker-Planck equation is

$$\partial_t p_t(\theta, r) = \nabla^T \{[D+G][\nabla H(\theta, r) p_t(\theta, r)]\} + \nabla^T [D \nabla p_t(\theta, r)] \tag{2.6}$$

$$= \nabla^T \{[D+G][p_t(\theta, r)\nabla H(\theta, r) + \nabla p_t(\theta, r)]\} \tag{2.7}$$

Let $z = (\theta, r)$, here we use

$$\begin{aligned} &\nabla^T [D \nabla p_t(\theta, r)] \\ &\quad = \sum_{ij} \partial_{z_i} \left[D_{ij}(z) \partial_{z_j} p_t(z)\right] \\ &\quad = \sum_{ij} \partial_{z_i} \left[D_{ij}(z) \partial_{z_j} p_t(z)\right] + \sum_{ij} \partial_{z_i} \left[\left(\partial_{z_j} D_{ij}(z)\right) p_t(z)\right] \\ &= \sum_{ij} \partial_{z_i} \partial_{z_j} \left[D_{ij}(z) p_t(z)\right] \end{aligned} \tag{2.8}$$

because $\partial_{z_j} D_{ij}(z) = \partial_{r_j} B_{ij}(\theta) = 0$

and $\nabla^T [G \nabla p_t(\theta, r)] = -\partial_\theta \partial_r p_t(\theta, r) + \partial_r \partial_\theta p_t(\theta, r) = 0$.

Because

$$e^{-H(\theta, r)} \nabla H(\theta, r) + \nabla e^{-H(\theta, r)} = 0$$

so it's easy to see $\pi(\theta, r)$ is the stationary distribution of the 2-order SDE.

3 Wasserstein

See section 4 of Convergence of Langevin MCMC in KL-divergence.

Wasserstein distance is

$$W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int \left(\|x - y\|_2^2\right) d\gamma^*(x, y)$$

, where $\gamma^* = \left(Id, T_{opt}\right)_\# \mu$ and T_{opt} is the optimal transport map satisfying $T_{opt\#}\boldsymbol{\mu} = \boldsymbol{\nu}$. The optimal displacement map is defined as $T_{opt} - Identity$. (Monge formulation)

The constant-speed-geodesic $\boldsymbol{\mu}_t$ between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ satisfies:

- $\boldsymbol{\mu}_0 = \boldsymbol{\nu}, \boldsymbol{\mu}_1 = \boldsymbol{\pi}$
- $W_2(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t) = (s - t) W_2(\boldsymbol{\nu}, \boldsymbol{\pi})$
- $\boldsymbol{\mu}_t = (Id + t v_{\boldsymbol{\nu}}^{\pi})_\# \boldsymbol{\nu}, \quad$ where $v_{\boldsymbol{\nu}}^{\pi}$ is the optimal displacement map

We define the metric derivative for a curve $\boldsymbol{\mu}_t$ as

$$\left|\boldsymbol{\mu}_t'\right| \triangleq \lim_{s \to t} \frac{W_2\left(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t\right)}{|s - t|}$$

If $\boldsymbol{\mu}_t$ is constant speed geodesic between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$, then

$$\begin{aligned}
\frac{W_2^2(\boldsymbol{\mu}_t, \boldsymbol{\nu})}{t^2} &= \frac{\int \|x - y\|_2^2 d\gamma_t^*(x, y)}{t^2} \\
&= \frac{\int \|x - (Id + t v_{\boldsymbol{\nu}}^{\pi})(x)\|_2^2 d\boldsymbol{\nu}(x)}{t^2} \\
&= \int \|v_{\boldsymbol{\nu}}^{\pi}(x)\|_2^2 d\boldsymbol{\nu}(x)
\end{aligned}$$

Thus $\left|\boldsymbol{\mu}_t'\right| = \sqrt{\int \|v_{\boldsymbol{\nu}}^{\pi}(x)\|_2^2 d\boldsymbol{\nu}(x)}$. It is also equal to $W_2(\boldsymbol{\nu}, \boldsymbol{\pi})$ from the above bullet 2.

4 Gradient Flow

4.1 General Case

$$\partial_t \mu_t = -\nabla \cdot (\mathbf{v}_t \mu_t) = \nabla \cdot \left(\mu_t \nabla \left(\frac{\delta E}{\delta \mu_t} (\mu_t) \right) \right) \tag{4.1}$$

4.2 SGLD

Define energy function

$$E(\mu) \triangleq \underbrace{-\int U(\boldsymbol{\theta})\mu(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}_{E_1} + \underbrace{\int \mu(\boldsymbol{\theta}) \log \mu(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}_{E_2}$$

If $U(\theta) = \log p_\theta$, then $E(\mu) = KL(\mu || p_\theta)$.

we have

$$\frac{\delta E_1}{\delta \mu} = -U, \quad \frac{\delta E_2}{\delta \mu} = \log \mu + 1$$

substitute this in Eq 4.1 we get

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla(-U + \log \mu_t)) = -\nabla_\theta \cdot (\frac{1}{2}\mu_t \nabla U(\theta_t)) + \frac{1}{2}\nabla_\theta \cdot (\nabla \mu_t)$$

This is same as the FPE in 2.1. So that FPE is also a gradient flow.

5 Generator of Stochastic Process

5.1 Definition

Let's say $\{x_t\}_t$ is a continuous time stochastic process, then its generator is

$$\mathcal{G}f := \frac{d}{dt}\mathbb{E}[f(x_t)|x_0]|_{t=0} = \lim_{t\to 0}\mathbb{E}[\frac{f(x_t)-f(x_0)}{t}]$$

The generator of SVGD has not been studied yet due to the interaction of particles, while that of SGLD is straight forward:

5.2 Generator of SGLD

SGLD is

$$x_\epsilon \leftarrow x_0 + \frac{\epsilon}{2}\nabla \log p(x_0) + \sqrt{\epsilon}\eta$$

where $\eta \sim N(0,1)$. Then its generator is

$$\begin{aligned}
\mathcal{G}f(x_0) &= \lim_{t\to 0}\frac{\mathbb{E}[f(x_t)-f(x_0)]}{t} && (5.1)\\
&= \lim_{t\to 0}\frac{\mathbb{E}[\nabla f(x_0)^T(x_t-x_0)+\frac{1}{2}(x_t-x_0)^T\nabla^2 f(x_0)(x_t-x_0)]+\mathrm{O}(t^2)}{t} && (5.2)\\
&= \frac{1}{2}\nabla f(x_0)^T\nabla\log p(x_0)+\frac{1}{2}\mathbb{E}[\eta^T\nabla^2 f\eta] && (5.3)\\
&= \frac{1}{2}\nabla f^T\nabla\log p+\frac{1}{2}\mathrm{Trace}(\nabla^2 f)|_{x=x_0} && (5.4)
\end{aligned}$$