CS886: Diffusion Models

Lec 01: Stochastic Integral

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

January 16, 2024

An Inherent Difficulty

How do we make sense of the stochastic differential equation?

$$\mathrm{d}\mathsf{X}_t = \mathbf{f}_t(\mathsf{X}_t)\,\mathrm{d}t + G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$$

- Brownian motion is nowhere differentiable: $\mathrm{d}\mathsf{B}_t$ does not exist!
- Interpret as an integral:

$$\mathsf{X}_T = \mathsf{X}_0 + \int_0^T \mathbf{f}_t(\mathsf{X}_t)\,\mathrm{d}t + \int_0^T G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$$

  - the 1st integral is the familiar one (Riemann or Lebesgue)
  - the 2nd integral?

"There are in this world optimists who feel that any symbol that starts off with an integral sign must necessarily denote something that will have every property that they should like an integral to possess. This of course is quite annoying to us rigorous mathematicians; what is even more annoying is that by doing so they often come up with the right answer."

— *Edward J. McShane*

E. J. McShane. "Integrals devised for special purposes". *Bulletin of the American Mathematical Society*, vol. 69 (1963), pp. 597–627.

What Is An Integral Anyway?

$$\int_0^T G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$$

- Integrand: $G_t(\mathsf{X}_t)$, living in some space
- Integrator: B_t, living in some possibly different space
- An integral is simply some notation that pairs an integrand and an integrator
  - often also written as a dual pairing $\langle G(\mathsf{X}); \mathsf{B}\rangle$
- We have yet to specify the pairing, i.e., integral

What Makes An Integral Useful?

- The properties that it enjoys, ideally conform to our “intuition” or tradition
- e.g., a bilinear form of integrand and integrator:
  - $\int (\alpha f + \beta g)\,\mathrm{dB} = \alpha \int f\,\mathrm{dB} + \beta \int g\,\mathrm{dB}$, i.e., $\langle \alpha f + \beta g; \mathrm{B}\rangle = \alpha \langle f; \mathrm{B}\rangle + \beta \langle g; \mathrm{B}\rangle$
  - $\int f\,\mathrm{d}(\alpha \mathrm{B} + \beta \mathrm{M}) = \alpha \int f\,\mathrm{dB} + \beta \int f\,\mathrm{dM}$, i.e., $\langle f; \alpha \mathrm{B} + \beta \mathrm{M}\rangle = \alpha \langle f; \mathrm{B}\rangle + \beta \langle f; \mathrm{M}\rangle$
- e.g., some continuity w.r.t. integrand and integrator
- Generality: can accommodate a large class of integrands and integrators
- Change of variable formula
- Possible to compute numerically, at least approximately

Wiener Integral

- Let $g : [0, T] \to \mathbb{R}$ be of bounded variation (e.g., continuously differentiable)
- $g(0) = g(T) = 0$
- Define the integral through integration by parts:

$$\int_0^T g(t)\,\mathrm{dB}_t = -\int_0^T \mathrm{B}_t\,\mathrm{d}g(t)$$

  - the rhs exists if $t \mapsto \mathrm{B}_t$ is continuous, a.k.a. Riemann-Stieltjes integral

- For more general g, apply approximation and define $\int g\,\mathrm{dB} = \lim_{n\to\infty} \int g_n\,\mathrm{dB}$
- What about $\int_0^T \mathrm{B}_t\,\mathrm{dB}_t$?

A Conundrum

- Let $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$, with $\delta_n := \max_{0 \le k \le n} |t_{k+1} - t_k| \to 0$
- Let $\tau_k := (1-\lambda)t_k + \lambda t_{k+1}$ for some $\lambda \in [0, 1]$
- Riemann-Stieltjes approximation of $\int_0^1 \mathsf{B}_t \, d\mathsf{B}_t$:

$$R := R(\delta_n, \lambda) = \sum_{k=0}^{n} \mathsf{B}_{\tau_k} [\mathsf{B}_{t_{k+1}} - \mathsf{B}_{t_k}]$$

$$= \frac{\mathsf{B}_1^2}{2} - \frac{1}{2} \underbrace{\sum_{k=0}^{n} [\mathsf{B}_{t_{k+1}} - \mathsf{B}_{t_k}]^2}_{S_n(1)} + \underbrace{\sum_{k=0}^{n} [\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k}]^2}_{S_n(\lambda)} + \sum_{k=0}^{n} \underbrace{[\mathsf{B}_{t_{k+1}} - \mathsf{B}_{\tau_k}]}_{\sqrt{(1-\lambda)\Delta t_k} N_k} \underbrace{[\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k}]}_{\sqrt{\lambda \Delta t_k} M_k}$$

Quadratic Variation

$$
\begin{aligned}
\mathbb{E}\Big(\sum_{k=0}^{n}(\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k})^2 - \lambda\Big)^2 &= \mathbb{E}\sum_{k,l=0}^{n}[(\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k})^2 - (\tau_k - t_k)] \cdot [(\mathsf{B}_{\tau_l} - \mathsf{B}_{t_l})^2 - (\tau_l - t_l)] \\
&= \sum_{k=0}^{n}\mathbb{E}[(\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k})^2 - (\tau_k - t_k)]^2 \\
&= \sum_{k=0}^{n}(\tau_k - t_k)^2 \cdot \mathbb{E}[\chi_1^2 - 1]^2 \\
&\le \lambda^2\delta_n \cdot \mathbb{E}[\chi_1^2 - 1]^2 \to 0
\end{aligned}
$$

- $\lambda = 0$: $R = (\mathsf{B}_1^2 - 1)/2$ — Itô's integral
- $\lambda = \frac{1}{2}$: $R = \mathsf{B}_1^2/2$ — Stratonovich's integral
- $\lambda = 1$: $R = (\mathsf{B}_1^2 + 1)/2$

How to Build Things?

- Start with very simple primitives, e.g., a rectangle
  - for an interval $(a, b]$ let $\mu(a, b] = b - a$
  - for any n, $\mathsf{X}_{t_1}, \mathsf{X}_{t_2}, \ldots, \mathsf{X}_{t_n}$ exists
- Extend by "obvious" desires, e.g., by linearity
- Extend by less "obvious" desires, e.g., by some form of continuity
- Extend by more technical means

Building Itô's Integral

- Recall that $X_t = X(t, \omega)$ is a function of two variables
- Indicator: $X(t, \omega) = \mathbf{1}_{(\varsigma, \tau]}(t) \cdot \mathbf{1}_A(\omega)$ for some $A \subseteq \Omega$
  - A can only depend on $\{B_t : t \in [0, \varsigma]\}$, i.e., $A \in \mathcal{F}_\varsigma$ (information up to time ς)
  - $\int X_t \, dB_t := [B_\tau - B_\varsigma] \cdot \mathbf{1}_A$ is a function of ω but not t (integrated out)
  - linear in the integrator by definition
- Also want linearity in the integrand
  - for $X(t, \omega) = \sum_k c_k \mathbf{1}_{(\varsigma_k, \tau_k]} \cdot \mathbf{1}_{A_k}$, define $\int X_t \, dB_t = \sum_k c_k [B_{\tau_k} - B_{\varsigma_k}] \cdot \mathbf{1}_{A_k}$
  - indeed well-defined

Approximation

- Adapted (non-anticipating): $\mathsf{X}_t \in \mathcal{F}_t = \sigma(\{\mathsf{B}_s : 0 \le s \le t\})$, information up to t
- Left continuous (l.c.): $t \mapsto \mathsf{X}_t$ is continuous from the left

$$\mathsf{X}_t^n := \mathsf{X}_{(\lceil t2^n \rceil - 1)/2^n} = \sum_{k=0}^{\infty} \mathsf{X}_{k/2^n} \,[\![k/2^n < t \le (k+1)/2^n]\!]$$

  - can approximate $\mathsf{X}_{k/2^n}(\omega) \approx \sum_i \frac{i}{2^m} \mathbf{1}_{A_i}(\omega)$, $A_i := \{(i-1)/2^m < \mathsf{X}_{k/2^n} \le i/2^m\}$

- Define $\int \mathsf{X}_t \, \mathrm{d}\mathsf{B}_t = \lim_n \int \mathsf{X}_t^n \, \mathrm{d}\mathsf{B}_t$?

Isometry

- Recall $\mathsf{X}(t,\omega) : \mathbb{R}_+ \times \Omega \to \mathbb{R}$
- Define the Doléans measure $\lambda = \lambda_{\mathsf{B}^2}$ as follows:

$$\boxed{\lambda\{(s,t] \times \mathbf{1}_A\} := \mathbb{E}[(\mathsf{B}_t^2 - \mathsf{B}_s^2)\mathbf{1}_A] = \mathbb{E}[(\mathsf{B}_t - \mathsf{B}_s)^2\mathbf{1}_A]}$$
$$= (t-s)\cdot\mu(A) = (\text{Lebesgue} \times \mu)\{(s,t] \times \mathbf{1}_A\}$$

  - i.e., $\lambda = \lambda_{\mathsf{B}^2} = \text{Lebesgue} \times \mu$
- Can now treat $\mathsf{X}(t,\omega)$ as a r.v. from $\mathcal{L}^2(\mathbb{R}_+ \times \Omega, \mathcal{P}, \lambda)$ to $\mathbb{R}$
- The integral $\int \mathsf{X}_t \, \mathrm{d}\mathsf{B}_t$ is a linear map that sends $\mathsf{X} \in \mathcal{L}^2(\mathbb{R}_+ \times \Omega, \lambda)$ to $\mathcal{L}^2(\Omega, \mu)$

$$\boxed{\|\mathsf{X}\|^2_{\mathcal{L}^2(\mathbb{R}_+\times\Omega,\lambda)} = \left\| \int \mathsf{X}_t \, \mathrm{d}\mathsf{B}_t \right\|^2_{\mathcal{L}^2(\Omega,\mu)}}$$

For $X(t,\omega) = \sum_k \mathbf{1}_{(\varsigma_k,\tau_k]}(t) \cdot Y_k(\omega)$, where $Y_k \in \mathcal{F}_{\varsigma_k}$, we have

$$\int X_t \, dB_t := \sum_k [B_{\tau_k} - B_{\varsigma_k}] \cdot Y_k$$

$$\|X\|^2_{\mathcal{L}^2(\mathbb{R}_+ \times \Omega, \lambda)} = \int \Big(\sum_k \mathbf{1}_{(\varsigma_k,\tau_k]}(t) \cdot Y_k(\omega) \Big)^2 d\lambda = \sum_k \int \mathbf{1}_{(\varsigma_k,\tau_k]}(t) \cdot Y_k^2(\omega) \, d\lambda(t,\omega)$$

$$= \sum_k \mathbb{E}[(B_{\tau_k} - B_{\varsigma_k})^2 \cdot Y_k^2]$$

$$\left\| \int X_t \, dB_t \right\|^2_{\mathcal{L}^2(\Omega,\mu)} = \sum_k \sum_l \int \int [B_{\tau_k} - B_{\varsigma_k}] \cdot Y_k \cdot [B_{\tau_l} - B_{\varsigma_l}] \cdot Y_l \, d\mu(\omega)$$

$$= \sum_k \mathbb{E}\left[(B_{\tau_k} - B_{\varsigma_k})^2 \cdot Y_k^2 \right]$$

$$
\begin{array}{ccc}
\mathsf{X}^n(t,\omega) & \xrightarrow{\quad L_2(\lambda) \quad} & \mathsf{X}(t,\omega) \\
\Big\downarrow {\scriptstyle isometry} & & \Big\downarrow {\scriptstyle def} \\
\int \mathsf{X}^n(t,\omega)\,\mathrm{d}\mathsf{B}_t & \xrightarrow{\quad L_2(\mu) \quad} & \int \mathsf{X}(t,\omega)\,\mathrm{d}\mathsf{B}_t
\end{array}
$$

Calculus of Itô's Integral

- Linear in integrand (and integrator)
- Zero mean (as some kind of average of Brownian motion):

$$\mathbb{E}\left(\int \mathsf{X}_t(\omega)\,\mathrm{d}\mathsf{B}_t(\omega)\right) = \int \left[\int \mathsf{X}_t(\omega)\,\mathrm{d}\mathsf{B}_t(\omega)\right] \mathrm{d}\mu(\omega) = 0$$

- Isometry, and hence continuity

From Definite to Indefinite

For any (left) continuous and bounded X_t, obtain

$$M_t = \int_0^t X_\tau \, dB_\tau = \int \mathbf{1}_{(0,t]}(\tau) \cdot X_\tau \, dB_\tau, \qquad \text{now as a function of both } t \text{ and } \omega$$

- If $t \mapsto X_t$ is continuous, so is $t \mapsto M_t$
- $M_0 := 0$, $\mathbb{E}[M_t] = 0$, $M_t \in \mathcal{F}_t$
- $\int_s^t X_\tau \, dB_\tau = \int_0^t X_\tau \, dB_\tau - \int_0^s X_\tau \, dB_\tau$
- $\mathbb{E}[M_t | \mathcal{F}_s] = \mathbb{E}[(M_t - M_s) + M_s | \mathcal{F}_s] = M_s$, a martingale
- $\mathbb{E}[M_t^2 - M_s^2] = \mathbb{E}[M_t - M_s]^2$

The Power of Abstraction

We have in fact defined the integral

$$\int \mathsf{Z}_t \, \mathrm{d}\mathsf{M}_t$$

- If $t \mapsto \mathsf{Z}_t$ is (left) continuous and adapted
- If M_t is a (continuous) martingale

$$\boxed{\int_0^t \mathsf{X}\mathsf{Z} \, \mathrm{d}\mathsf{B} = \int_0^t \mathsf{Z} \, \mathrm{d}\mathsf{M}} \quad \text{where} \quad \mathsf{M}_t = \int_0^t \mathsf{X} \, \mathrm{d}\mathsf{B}$$

"There are in this world optimists who feel that any symbol that starts off with an integral sign must necessarily denote something that will have every property that they should like an integral to possess. This of course is quite annoying to us rigorous mathematicians; what is even more annoying is that by doing so they often come up with the right answer."

— *Edward J. McShane*

E. J. McShane. "Integrals devised for special purposes". *Bulletin of the American Mathematical Society*, vol. 69 (1963), pp. 597–627.

Quadratic Variation, again

$$\mathsf{S}_t^n := \sum_{k=0}^{n} [\mathsf{M}_{t_{k+1}} - \mathsf{M}_{t_k}]^2$$

- $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = t$ a partition of $[0, t]$
- M_t is a continuous martingale (e.g., Brownian motion)
- As $\delta_n := \max_k t_{k+1} - t_k \to 0$,

$$\mathsf{S}_t^n \to [\mathsf{M}]_t := \mathsf{M}_t^2 - 2\int_0^t \mathsf{M}\,\mathrm{d}\mathsf{M} - \mathsf{M}_0^2$$

- For Brownian motion, $[\mathsf{B}]_t = t$
- Quadratic variation $[\mathsf{M}]_t$ is increasing, continuous with $[\mathsf{M}]_0 = 0$

Another Integral

$$\int_0^t \mathsf{X}_\tau \,\mathrm{d}[\mathsf{M}]_\tau$$

- Fix ω, reduce to Riemann-Stieltjes integral
- Essentially a conditional measure of the Doléans measure

$$\boxed{\lambda_{\mathsf{M}^2}(\mathrm{d}t, \mathrm{d}\omega) = \mu(\mathrm{d}\omega) \cdot [\mathsf{M}](\mathrm{d}t, \omega)}$$

- For Brownian motion:

$$\lambda(\mathrm{d}t, \mathrm{d}\omega) = \mu(\mathrm{d}\omega) \times \mathsf{Lebesgue}(\mathrm{d}t), \quad [\mathsf{B}](\mathrm{d}t, \omega) = \mathrm{d}t$$

$$\boxed{\int \mathsf{X}^2(t, \omega) \,\mathrm{d}\lambda(\mathrm{d}t, \mathrm{d}\omega) = \int \left(\int \mathsf{X}^2(t, \omega) \,\mathrm{d}[\mathsf{M}]_t \right) \mathrm{d}\mu(\mathrm{d}\omega)}$$

Itô's Formula

$$f(\mathsf{M}_t, \mathsf{V}_t) - f(\mathsf{M}_0, \mathsf{V}_0) = \int_0^t f_x(\mathsf{M}_s, \mathsf{V}_s)\, \mathrm{d}\mathsf{M}_s + \int_0^t f_y(\mathsf{M}_s, \mathsf{V}_s)\, \mathrm{d}\mathsf{V}_s + \frac{1}{2} \int_0^t f_{xx}(\mathsf{M}_s, \mathsf{V}_s)\, \mathrm{d}[\mathsf{M}]_s$$

- M_t is a continuous martingale, e.g., Brownian motion
- V_t is continuously differentiable in t
- Continuous partial derivatives f_x, f_y, f_{xx} exist
- Recall that for Brownian motion, $[\mathsf{B}]_t = t$

$$f(\mathsf{M}_t, \mathsf{V}_t) - f(\mathsf{M}_0, \mathsf{V}_0) = \sum_k [f(\mathsf{M}_{t_{k+1}}, \mathsf{V}_{t_{k+1}}) - f(\mathsf{M}_{t_{k+1}}, \mathsf{V}_{t_k}) + f(\mathsf{M}_{t_{k+1}}, \mathsf{V}_{t_k}) - f(\mathsf{M}_{t_k}, \mathsf{V}_{t_k})]$$

$$\approx \sum_k f_y(\mathsf{M}_{t_k}, \mathsf{V}_{t_k}) \Delta \mathsf{V}_{t_k} + f_x(\mathsf{M}_{t_k}, \mathsf{V}_{t_k}) \Delta \mathsf{M}_{t_k} + \frac{1}{2} f_{xx}(\mathsf{M}_{t_k}, \mathsf{V}_{t_k}) (\Delta \mathsf{M}_{t_k})^2$$

As $\delta_n := \max_k t_{t+1} - t_k \to 0$, apply continuity to obtain the limit:

$$\int_0^t f_y(\mathsf{M}_s, \mathsf{V}_s) \, \mathrm{d}\mathsf{V}_s + \int_0^t f_x(\mathsf{M}_s, \mathsf{V}_s) \, \mathrm{d}\mathsf{M}_s + \frac{1}{2} \int_0^t f_{xx}(\mathsf{M}_s, \mathsf{V}_s) \, \mathrm{d}[\mathsf{M}]_s$$

From Integral to Differential

$$\boxed{\mathrm{d}f(\mathsf{M}_t, \mathsf{V}_t) = f_x(\mathsf{M}_t, \mathsf{V}_t)\,\mathrm{d}\mathsf{M}_t + f_y(\mathsf{M}_t, \mathsf{V}_t)\,\mathrm{d}\mathsf{V}_t + \tfrac{1}{2} f_{xx}(\mathsf{M}_t, \mathsf{V}_t)\,\mathrm{d}[\mathsf{M}]_t}$$

- Recall also

$$\int_0^t \mathsf{X}\mathsf{Z}\,\mathrm{d}\mathsf{B} = \int_0^t \mathsf{Z}\,\mathrm{d}\mathsf{M} \quad \text{where} \quad \mathsf{M}_t = \int_0^t \mathsf{X}\,\mathrm{d}\mathsf{B}$$

- Rewritten in terms of differential

$$\mathsf{Z}\,\mathrm{d}\mathsf{M} = \mathsf{X}\mathsf{Z}\,\mathrm{d}\mathsf{B} \quad \text{where} \quad \mathrm{d}\mathsf{M} = \mathsf{X}\,\mathrm{d}\mathsf{B}$$

- Let $\langle \mathsf{X}; \mathsf{M} \rangle := \int_0^t \mathsf{X}\,\mathrm{d}\mathsf{M}$, then

$$\boxed{[\langle \mathsf{X}; \mathsf{M} \rangle] = \langle \mathsf{X}^2; [\mathsf{M}] \rangle}$$

Example

$$\mathrm{d}\mathsf{X} = \mathsf{F}\,\mathrm{d}t + \mathsf{G}\,\mathrm{d}\mathsf{M}_t$$

Derive the differential of $f(\mathsf{X}, t)$:

$$\begin{aligned}
\mathrm{d}f(\mathsf{X},t) &= f_x(\mathsf{X},t)\,\mathrm{d}\mathsf{X} + f_y(\mathsf{X},t)\,\mathrm{d}t + \tfrac{1}{2}f_{xx}(\mathsf{X},t)\,\mathrm{d}[\mathsf{X}]_t \\
&= f_x(\mathsf{X},t)[\mathsf{F}\,\mathrm{d}t + \mathsf{G}\,\mathrm{d}\mathsf{M}_t] + f_y(\mathsf{X},t)\,\mathrm{d}t + \tfrac{1}{2}f_{xx}(\mathsf{X},t)\mathsf{G}^2\,\mathrm{d}[\mathsf{M}]_t \\
&= [f_x(\mathsf{X},t)\cdot\mathsf{F} + f_y(\mathsf{X},t)]\,\mathrm{d}t + [f_x(\mathsf{X},t)\cdot\mathsf{G}]\,\mathrm{d}\mathsf{M}_t + \tfrac{1}{2}f_{xx}(\mathsf{X},t)\mathsf{G}^2\,\mathrm{d}[\mathsf{M}]_t
\end{aligned}$$

For the Brownian motion, we have

$$\mathrm{d}f(\mathsf{X},t) = [f_x(\mathsf{X},t)\cdot\mathsf{F} + f_y(\mathsf{X},t) + \tfrac{1}{2}f_{xx}(\mathsf{X},t)\cdot\mathsf{G}^2]\,\mathrm{d}t + [f_x(\mathsf{X},t)\cdot\mathsf{G}]\,\mathrm{d}\mathsf{M}_t$$

- Rule of thumb: $(\mathrm{d}t)^2 = 0, \mathrm{d}t\,\mathrm{d}\mathsf{B}_t = 0, (\mathrm{d}\mathsf{B}_t)^2 = \mathrm{d}t,\ \mathrm{d}\mathsf{B}_t\,\mathrm{d}\tilde{\mathsf{B}}_t = 0$

More Examples

$$\mathrm{d}(\mathsf{B}_t^m) = m\mathsf{B}_t^{m-1}\,\mathrm{d}\mathsf{B}_t + \binom{m}{2}\mathsf{B}_t^{m-2}\,\mathrm{d}t$$

- For $m = 2$, $\mathrm{d}(\mathsf{B}_t^2) = 2\mathsf{B}_t\,\mathrm{d}\mathsf{B}_t + \mathrm{d}t$

$$\mathrm{d}\left(\exp(\lambda\mathsf{B}_t - \tfrac{\lambda^2 t}{2})\right) = \lambda\exp(\lambda\mathsf{B}_t - \tfrac{\lambda^2 t}{2})\,\mathrm{d}\mathsf{B}_t$$

- Let $\mathsf{Y}_t = \exp(\lambda\mathsf{B}_t - \frac{\lambda^2 t}{2})$, then

$$\mathrm{d}\mathsf{Y}_t = \lambda\mathsf{Y}_t\,\mathrm{d}\mathsf{B}_t, \qquad \mathsf{Y}_0 = 1$$

Score Matching

$$\begin{aligned}\mathbb{F}(p\|q) &:= \tfrac{1}{2}\mathbb{E}_{\mathsf{X}\sim q}\|\partial_{\mathbf{x}}\log p(\mathsf{X}) - \partial_{\mathbf{x}}\log q(\mathsf{X})\|_2^2 \\ &= \mathbb{E}_{\mathsf{X}\sim q}\left[\tfrac{1}{2}\|\mathsf{s}_p(\mathsf{X})\|_2^2 + \langle\partial_{\mathbf{x}}, \mathsf{s}_p(\mathsf{X})\rangle + \tfrac{1}{2}\|\mathsf{s}_q(\mathsf{X})\|_2^2\right] \\ &\approx \hat{\mathbb{E}}_{\mathsf{X}\sim q}\left[\tfrac{1}{2}\|\mathsf{s}_p(\mathsf{X})\|_2^2 + \langle\partial_{\mathbf{x}}, \mathsf{s}_p(\mathsf{X})\rangle\right]\end{aligned}$$

- Under mild conditions, $\mathbb{F}(p\|q) = 0 \iff p \propto q$
- A Convenient way to estimate the score s_q and hence the density q
- The model score function s_p can be chosen as any NN

A. Hyvärinen. "Estimation of Non-Normalized Statistical Models by Score Matching". *Journal of Machine Learning Research*, vol. 6, no. 24 (2005), pp. 695–709.

Score Matching for Exponential Family

$$\min_{\boldsymbol{\theta}} \ \hat{\mathbb{E}}_{\mathsf{X}\sim q}\left[\tfrac{1}{2}\|\mathbf{s}(\mathsf{X};\boldsymbol{\theta})\|_2^2 + \langle \partial_{\mathbf{x}}, \mathbf{s}(\mathsf{X};\boldsymbol{\theta})\rangle\right]$$

- If the model density p is in the exponential family:

$$\mathbf{s}(\mathbf{x};\boldsymbol{\theta}) = \partial_{\mathbf{x}}\left\langle \mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right\rangle = [\partial_{\mathbf{x}}\mathbf{T}(\mathbf{x})]^\top \boldsymbol{\theta}$$

$$\langle \partial_{\mathbf{x}}, \mathbf{s}(\mathbf{x};\boldsymbol{\theta})\rangle = \langle \partial_{\mathbf{x}}, \partial_{\mathbf{x}}\left\langle \mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right\rangle\rangle = \left\langle \partial_{\mathbf{x}}^2 \mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right\rangle$$

- Can solve $\boldsymbol{\theta}$ in closed-form by simply setting the derivative w.r.t. $\boldsymbol{\theta}$ to $\mathbf{0}$:

$$\boldsymbol{\theta} = -\left\{\hat{\mathbb{E}}_{\mathsf{X}\sim q}[\partial_{\mathbf{x}}\mathbf{T}(\mathbf{x})]^\top[\partial_{\mathbf{x}}\mathbf{T}(\mathbf{x})]\right\}^{-1} \cdot \hat{\mathbb{E}}_{\mathsf{X}\sim q}[\partial_{\mathbf{x}}^2\mathbf{T}(\mathbf{x})]$$

- For multivariate Gaussian, $\boldsymbol{\theta} = (S^{-1}, S^{-1}\boldsymbol{\mu})$, $\mathbf{T}(\mathbf{x}) = (-\tfrac{1}{2}\mathbf{x}\mathbf{x}^\top, \mathbf{x})$ and

$$\min_{\boldsymbol{\mu}, S} \ \hat{\mathbb{E}}_{\mathsf{X}\sim q} \tfrac{1}{2}\|S^{-1}(\mathbf{x}-\boldsymbol{\mu})\|_2^2 - \mathrm{tr}(S^{-1})$$

Denoising Auto-Encoder

- Suppose also have a latent variable Z with joint density $q(\mathbf{x}, \mathbf{z})$
- Exchage differentiation with integration we obtain:

$$
\begin{aligned}
\mathbb{F}(p\|q) &:= \tfrac{1}{2}\mathbb{E}_{\mathsf{X}\sim q}\|\partial_{\mathbf{x}} \log p(\mathsf{X}) - \partial_{\mathbf{x}} \log q(\mathsf{X})\|_2^2 \\
&= \tfrac{1}{2}\mathbb{E}_{(\mathsf{X},\mathsf{Z})\sim q}[\|\mathbf{s}_p(\mathsf{X}) - \partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})\|_2^2 + \|\mathbf{s}_q(\mathsf{X})\|_2^2 - \|\partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})\|_2^2] \\
&\approx \tfrac{1}{2}\hat{\mathbb{E}}_{(\mathsf{X},\mathsf{Z})\sim q}\|\mathbf{s}_p(\mathsf{X}) - \partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})\|_2^2
\end{aligned}
$$

- Useful when the conditional density $\partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})$ is easy to obtain

---

P. Vincent. "A Connection Between Score Matching and Denoising Autoencoders". *Neural Computation*, vol. 23, no. 7 (2011), pp. 1661–1674.