CS886: Diffusion Models

Lec 02: Stochastic Differential Equations

Yaoliang Yu

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

January 23, 2024

Itô Process

$$\mathrm{d}\mathsf{X}(t,\omega) = \mathsf{F}(t,\omega)\,\mathrm{d}t + \mathsf{G}(t,\omega)\,\mathrm{d}\mathsf{B}(t,\omega)$$

$$\mathsf{X}(t,\omega) = \mathsf{X}(0,\omega) + \int_0^t \mathsf{F}(s,\omega)\,\mathrm{d}s + \int_0^t \mathsf{G}(s,\omega)\,\mathrm{d}\mathsf{B}(s,\omega)$$

- The 1st integral is the familiar Riemann or Lebesgue integral
- The 2nd integral is Itô's stochastic integral

Itô's formula:

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

- $[\mathsf{X}]_t := \lim_{n\to\infty}\sum_{k=0}^{n}(\mathsf{X}_{t_{k+1}} - \mathsf{X}_{t_k})^2 = \mathsf{G}^\top\mathsf{G}\cdot[\mathsf{B}]_t, \quad [\mathsf{B}]_t = t$

Martingale

$(M_t, \mathcal{F}_t)$ is a martingale iff

- $M_t \in \mathcal{F}_t$, i.e., M_t determined by information $\mathcal{F}_t$ up to t
  - e.g., $\mathcal{F}_t = \sigma(M_s : s \leq t)$, i.e., all (measurable) functions of $\{M_s : s \leq t\}$
- for all t, $\mathbb{E}|M_t| < \infty$
- for all $t \geq s$, $\mathbb{E}[M_t|\mathcal{F}_s] = M_s$
  - $\mathbb{E}[M_t|M_s] = \mathbb{E}[(M_t - M_s) + M_s|\mathcal{F}_s] = M_s \iff \mathbb{E}[M_t - M_s|\mathcal{F}_s] = 0$
  - given what we know ($\mathcal{F}_s$) at time s, changes in the future ($M_t - M_s$) are 0 in expectation
  - e.g. cumsum of independent r.v., $S_n := \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$, $\mathcal{F}_n := \sigma(X_i, i = 1, \ldots, n)$

Itô's Integral is a Square Integrable Martingale

$$X_t = \int_0^t G_s \, dB_s$$

- If $t \mapsto G_t$ is continuous, so is $t \mapsto X_t$
- $X_0 := 0$, $\mathbb{E}[X_t] = \mathbb{E}[X_0] = 0$, $X_t \in \mathcal{F}_t$
- $\mathbb{E}[X_t^2 - X_s^2] = \mathbb{E}[X_t - X_s]^2$

Theorem: Martingale representation

Let $(\mathsf{M}_t, \mathcal{F}_t)$ be a square integrable martingale. Then, there exists a unique process G_t such that

$$\mathsf{M}_t = \mathbb{E}[\mathsf{M}_0] + \int_0^t \mathsf{G}_s \,\mathrm{d}\mathsf{B}_s.$$

Itô's Stochastic Differential Equation (SDE)

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

$$\mathsf{X}_t = \mathsf{X}_0 + \int_0^t \mathbf{b}_s(\mathsf{X}_s)\,\mathrm{d}s + \int_0^t G_s(\mathsf{X}_s)\,\mathrm{d}\mathsf{B}_s$$

- We will see that $t \mapsto \mathsf{X}_t$ is continuous
- We assume $\mathbf{b}$ and G are continuous
- The 1st integral is the usual Riemann integral
- The 2nd integral is Itô's stochastic integral

But, does such X_t actually exist?

Geometric Brownian Motion

$$\frac{\mathrm{d}S_t}{S_t} = b\,\mathrm{d}t + \sigma\,\mathrm{dB}_t, \text{ or equivalently } \boxed{\mathrm{d}S_t = bS_t\,\mathrm{d}t + \sigma S_t\,\mathrm{dB}_t}$$

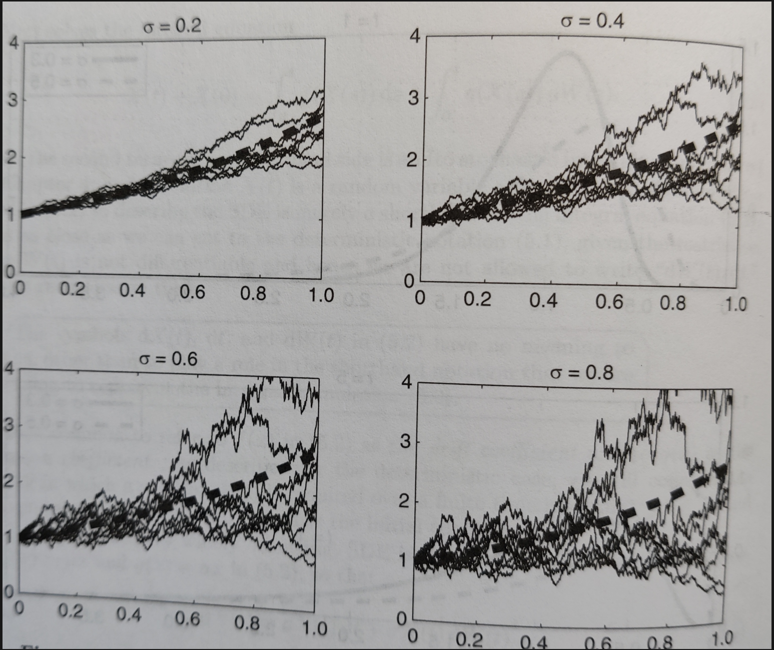
- $b(S_t) = bS_t$
- $G(S_t) = \sigma S_t$
- Apply Itô's formula to $f(S_t) = \ln(S_t)$:

$$\mathrm{d}\ln(S_t) = \frac{1}{S_t}\,\mathrm{d}S_t - \frac{1}{2S_t^2}\,\mathrm{d}[S]_t = \frac{1}{S_t}\,\mathrm{d}S_t - \frac{\sigma^2}{2}\,\mathrm{d}t = (b - \frac{\sigma^2}{2})\,\mathrm{d}t + \sigma\,\mathrm{dB}_t$$

$$\boxed{S_t = S_0 \cdot \exp\left[(b - \frac{\sigma^2}{2})t + \sigma \mathrm{B}_t\right]}$$

- Take expectation:

$$\mathrm{d}\mathbb{E}[S_t] = b\mathbb{E}[S_t] \cdot \mathrm{d}t \implies \mathbb{E}[S_t] = \mathbb{E}[S_0] \cdot \exp(bt)$$


σ = 0.2
σ = 0.4
σ = 0.6
σ = 0.8
0 1 2 3 4
0 0.2 0.4 0.6 0.8 1.0

Langevin's Equation

$$\mathrm{d}X_t = -bX_t\,\mathrm{d}t + \sigma\,\mathrm{d}B_t$$

- X_t is the velocity of a particle
- b is the friction coefficient
- σ models random perturbation

$$X_t = \exp(-bt)\cdot X_0 + \sigma\int_0^t \exp[-b(t-s)]\,\mathrm{d}B_s$$

$$\mathbb{E}[X_t] = \exp(-bt)\cdot\mathbb{E}[X_0],\quad \mathbb{E}[X_t^2] = \exp(-2bt)\cdot\mathbb{E}[X_0^2] + \frac{\sigma^2}{2b}[1-\exp(-2bt)]$$

- As $t\to\infty$, $X_t \to \mathcal{N}(0, \frac{\sigma^2}{2b})$

Brownian Motion on a Riemannian Manifold

$$\mathsf{X}(t) = (\cos \mathsf{B}_t,\ \sin \mathsf{B}_t)$$

$$\begin{cases} d\mathsf{X}_1 &= -\frac{1}{2}\mathsf{X}_1\, dt - \mathsf{X}_2\, d\mathsf{B}_t \\ d\mathsf{X}_2 &= -\frac{1}{2}\mathsf{X}_2\, dt + \mathsf{X}_1\, d\mathsf{B}_t \end{cases} \iff d\mathsf{X} = -\frac{1}{2}\mathsf{X}\, dt + J\mathsf{X}\, d\mathsf{B}_t, \quad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$d\mathbb{E}[\mathsf{X}] = -\frac{1}{2}\mathbb{E}[\mathsf{X}]\, dt \implies \mathbb{E}[\mathsf{X}] = \mathbb{E}[\mathsf{X}(0)] \cdot \exp(-\tfrac{1}{2}t)$$

Brownian Bridge

$$\mathrm{d}\mathsf{X}_t = \frac{-\mathsf{X}_t}{1-t}\,\mathrm{d}t + \mathrm{d}\mathsf{B}_t, \quad 0 \le t < 1, \quad \mathsf{X}_0 = 0$$

$$\mathsf{X}_t = \int_0^t \frac{1-t}{1-s}\,\mathrm{d}\mathsf{B}_s, \qquad \mathbb{E}[\mathsf{X}_t] \equiv 0$$

Constructing the Solution

$$\mathsf{X}_t = \mathsf{X}_0 + \int_0^t \mathbf{b}_s(\mathsf{X}_s)\,\mathrm{d}s + \int_0^t G_s(\mathsf{X}_s)\,\mathrm{d}\mathsf{B}_s$$

$$\mathsf{X}^{k+1} := \mathscr{T}(\mathsf{X}^k) := \mathsf{X}_0^k + \int_0^t \mathbf{b}_s(\mathsf{X}_s^k)\,\mathrm{d}s + \int_0^t G_s(\mathsf{X}_s^k)\,\mathrm{d}\mathsf{B}_s$$

- A solution X_t is simply a fixed point of the mapping $\mathscr{T}$
- Perhaps the iterated sequence X^k converges to a solution?

$$\mathbb{E}\left[\sup_{0\le s\le t}|\mathscr{T}(\mathsf{Y}_s)-\mathscr{T}(\mathsf{Z}_s)|^2\right]\lesssim \mathbb{E}[|\mathsf{Y}_0-\mathsf{Z}_0|^2]+\mathbb{E}\left[\int_0^t|\mathsf{Y}_s-\mathsf{Z}_s|^2\,\mathrm{d}s\right]$$

$$|\mathscr{T}(\mathsf{Y}_s)-\mathscr{T}(\mathsf{Z}_s)|^2\lesssim|\mathsf{Y}_0-\mathsf{Z}_0|^2+|\int_0^t[G_s(\mathsf{Y}_s)-G_s(\mathsf{Z}_s)]\,\mathrm{d}\mathsf{B}_s|^2+|\int_0^t[b_s(\mathsf{Y}_s)-b_s(\mathsf{Z}_s)]\,\mathrm{d}s|^2$$

$$\mathbb{E}|\int_0^t[G_s(\mathsf{Y}_s)-G_s(\mathsf{Z}_s)]\,\mathrm{d}\mathsf{B}_s|^2\lesssim\mathbb{E}[\int_0^t|G_s(\mathsf{Y}_s)-G_s(\mathsf{Z}_s)|^2\,\mathrm{d}s]\lesssim\mathbb{E}[\int_0^t|\mathsf{Y}_s-\mathsf{Z}_s|^2\,\mathrm{d}s]$$

$$\mathbb{E}|\int_0^t[b_s(\mathsf{Y}_s)-b_s(\mathsf{Z}_s)]\,\mathrm{d}s|^2\lesssim t\mathbb{E}[\int_0^t|b_s(\mathsf{Y}_s)-b_s(\mathsf{Z}_s)|^2]\lesssim\mathbb{E}[\int_0^t|\mathsf{Y}_s-\mathsf{Z}_s|^2\,\mathrm{d}s]$$

- Existence

$$\mathbb{E}\Big[\underbrace{\sup_{0\le s\le t}|\mathscr{T}(X^{n+1}_s)-\mathscr{T}(X^n_s)|^2}_{D^{n+1}_t}\Big]\lesssim\mathbb{E}\Big[\int_0^t\underbrace{|X^{n+1}_s-X^n_s|^2}_{\le D^n_s}\,\mathrm{d}s\Big]$$

  - apply recursion to obtain $D^n_t\lesssim\frac{t^n}{n!}+\frac{t^{n+1}}{(n+1)!}$
  - it follows that $X^n t X$, which is indeed a solution

- Uniqueness

$$\mathbb{E}\left[\sup_{0\le s\le t}|Y_s-Z_s|^2\right]\lesssim\mathbb{E}\left[\int_0^t|Y_s-Z_s|^2\,\mathrm{d}s\right]$$

  - apply Gronwall's inequality: $D(t)\lesssim\int_0^t D(s)\,\mathrm{d}s\implies D(t)\le 0$

Existence and Uniqueness

$$\mathsf{X}_t = \mathsf{X}_0 + \int_0^t \mathbf{b}_s(\mathsf{X}_s)\,\mathrm{d}s + \int_0^t G_s(\mathsf{X}_s)\,\mathrm{d}\mathsf{B}_s$$

There exists a unique continuous solution of SDE, if

- b and G are locally Lipschitz continuous (e.g., continuously differentiable)
  - i.e., $|b(\mathbf{x}) - b(\mathbf{y})| \leq L \cdot |\mathbf{x} - \mathbf{y}|$ for all $\mathbf{x}, \mathbf{y} \in B(0, r)$
- b and G have linear growth
  - i.e., $|b(\mathbf{x})|^2 \leq K \cdot (1 + |\mathbf{x}|^2)$
  - trivially holds if b is linear in $\mathbf{x}$

Markov Process

$$\mathsf{X}_t = \mathsf{X}_0 + \int_0^t \mathbf{b}_s(\mathsf{X}_s)\,\mathrm{d}s + \int_0^t G_s(\mathsf{X}_s)\,\mathrm{d}\mathsf{B}_s$$

For any bounded and continuous function f:

$$\mathbb{E}^{\mathbf{x}}[f(\mathsf{X}_{t+s})|\mathcal{F}_s] = \mathbb{E}^{\mathsf{X}_s}[f(\mathsf{X}_t)]$$

- $\mathbb{E}^{\mathbf{x}}$: when we start with $\mathsf{X}_0 = \mathbf{x}$

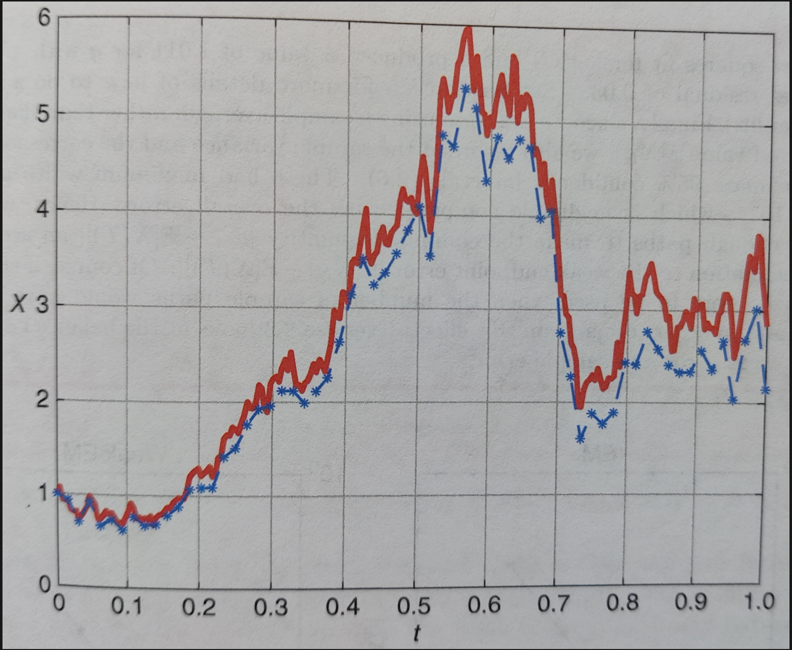
Euler-Maruyama

$$\mathsf{X}_t = \mathsf{X}_s + \int_s^t \mathbf{b}_\tau(\mathsf{X}_\tau)\,\mathrm{d}\tau + \int_s^t G_\tau(\mathsf{X}_\tau)\,\mathrm{d}\mathsf{B}_\tau$$
$$\approx \mathsf{X}_s + \mathbf{b}_s(\mathsf{X}_s)\cdot[t-s] + G_s(\mathsf{X}_s)[\mathsf{B}_t - \mathsf{B}_s]$$

- Divide $0 := t_0 < t_1 < \cdots < t_n < t_{n+1} = t$
- For $k = 1, \ldots, n$, compute

$$\mathsf{X}_{t_{k+1}} = \mathsf{X}_{t_k} + b_{t_k}(X_{t_k})\cdot\Delta t_k + G_{t_k}(\mathsf{X}_{t_k})\cdot\Delta B_{t_k}$$

  - $\Delta B_{t_k} \overset{i.i.d.}{\simeq} \mathcal{N}(0, \Delta t_k)$
- Interpolate
  - $\hat{\mathsf{X}}_t := \sum_k \mathsf{X}_{t_k}\cdot[\![t_k \le t < t_{k+1}]\!]$
  - $\bar{\mathsf{X}}_t := \mathscr{T}(\hat{\mathsf{X}}_t)$



6
5
4
X 3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
t

"I did not hear any reaction from other mathematicians about my paper until long after the end of the war, when Gisiro Maruyama told me in person that he had been drafted into the army and had read my paper in the barracks in 1942. Perhaps he and I were the only researchers then interested in the problem of sample paths."

— *Kiyosi Itô*

Langevin again

$$\mathrm{d}\mathsf{X}_t = -b(\mathsf{X}_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}\mathsf{B}_t$$

- Euler-Maruyama:

$$\mathsf{X}_{k+1} = \mathsf{X}_k - b(\mathsf{X}_k)\Delta t_k + \mathcal{N}(0, \sigma_k^2 \Delta t_k)$$

- Let $\Delta t_k =: \eta_k$ and $b(\mathsf{X}_k) =: \nabla f(\mathsf{X}_t)$

$$\mathsf{X}_{k+1} = \mathsf{X}_k - \eta_k[\nabla f(\mathsf{X}_k) + \mathcal{N}(0, \sigma_k^2/\eta_k)]$$

- In typical SGD, $\eta_k = O(1/\sqrt{k})$ and $\sigma_k^2/\eta_k = O(1)$, meaning $\sigma_k = k^{-1/4}$

- In Langevin gradient, $\sigma_k \equiv 1$

Complexity

It can be shown that the Euler-Muruyama scheme has

- weak convergence of order $O(\Delta t)$
- strong convergence of order $O(\sqrt{\Delta t})$

More generally, for a scheme with convergence order $O(\Delta t)^p$, to achieve ϵ accuracy:

- Need $\Delta t = \epsilon^{1/p}$
- Need to evaluates b and G roughly $1/\Delta t$ times
- For $p = 1/2$, obtain the familiar rate $1/\epsilon^2$ for SGD

Implicit methods

$$\mathsf{X}_t = \mathsf{X}_s + \int_s^t \mathbf{b}_\tau(\mathsf{X}_\tau)\,\mathrm{d}\tau + \int_s^t G_\tau(\mathsf{X}_\tau)\,\mathrm{d}\mathsf{B}_\tau$$

$$\mathsf{X}_{t_{k+1}} = \mathsf{X}_{t_k} + [(1-\theta)b_{t_k}(X_{t_k}) + \theta b_{t_{k+1}}(X_{t_{k+1}})] \cdot \Delta t_k + G_{t_k}(\mathsf{X}_{t_k}) \cdot \Delta B_{t_k}$$

- $\theta = 0$: Euler-Muruyama
- $\theta = 1$: implicit method
- $\theta = \frac{1}{2}$: Trapezoid

Richardson Extrapolation

Suppose

$$\mathbb{E}[\mathsf{X}_t^{\Delta}] - \mathbb{E}[\mathsf{X}_t] = A\Delta + B\Delta^2 + \cdots$$

Then, with half of Δ:

$$\mathbb{E}[\mathsf{X}_t^{\Delta/2}] - \mathbb{E}[\mathsf{X}_t] = A\Delta/2 + B\Delta^2/4 + \cdots$$

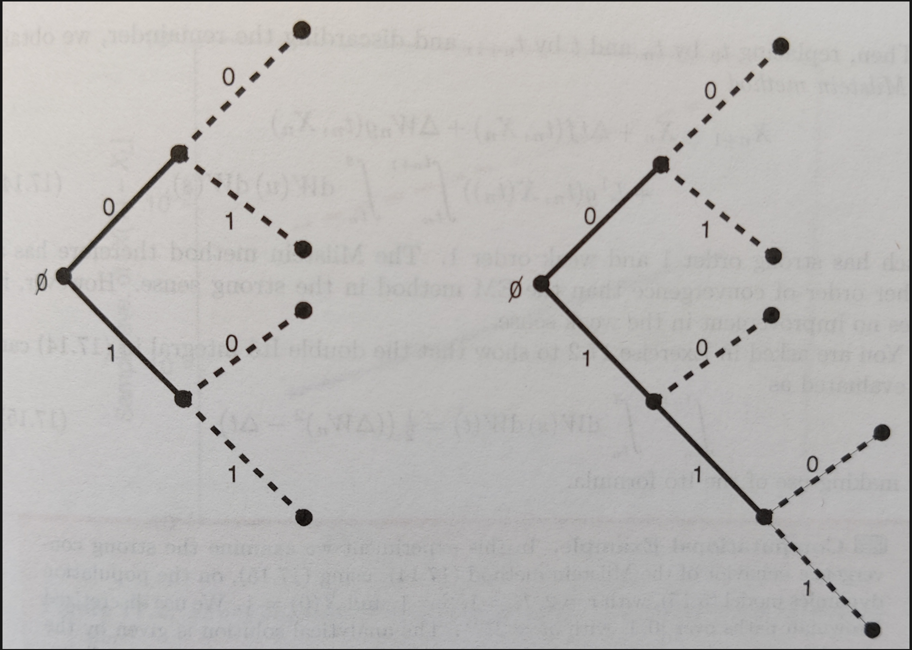
Thus,

$$\mathbb{E}[2\mathsf{X}_t^{\Delta/2} - \mathsf{X}_t^{\Delta}] - \mathbb{E}[\mathsf{X}_t] = -B\Delta^2/2 + \cdots$$

Extrapolation eliminates lower order terms!

Multi-level Monte Carlo

High order methods



ø
0
1
0
1
0
1
ø
0
1
0
1
0
1
0
1

Score Matching

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

- 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:

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

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

- 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.