CS886: Diffusion Models

Lec 05: Reverse Stochastic Differential Equations

Yaoliang Yu

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

February 13, 2024

"I can illustrate the ... approach with the ... image of a nut to be opened. The first analogy that came to my mind is of immersing the nut in some softening liquid, and why not simply water? From time to time you rub so the liquid penetrates better, and otherwise, you let time pass. The shell becomes more flexible through weeks and months — when the time is ripe, hand pressure is enough, the shell opens like a perfectly ripened avocado! A different image came to me a few weeks ago. The unknown thing to be known appeared to me as some stretch of earth or hard marble, resisting penetration ... the sea advances insensibly in silence, nothing seems to happen, nothing moves, the water is so far off you hardly hear it ... yet finally it surrounds the resistant substance."

— *Alexandre Grothendieck*

Notation for divergence

$$\langle \nabla, \mathbf{b} \rangle := \sum_i \nabla_i b_i, \quad \text{in particular} \quad \langle \nabla, \nabla p \rangle = \sum_i \nabla_i (\nabla_i p) = \sum_i \nabla_i^2 p =: \Delta p.$$

- $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ and $p : \mathbb{R}^d \to \mathbb{R}$
- Not to confuse the scalar $\langle \nabla, \nabla p \rangle$ with the matrix $\nabla^2 p$

$$\boxed{\int \langle \nabla p(\mathbf{x}), \mathbf{b}(\mathbf{x}) \rangle \, \mathrm{d}\mathbf{x} = - \int p(\mathbf{x}) \langle \nabla, \mathbf{b}(\mathbf{x}) \rangle \, \mathrm{d}\mathbf{x}}$$

- Assuming each pb_i vanishes at the boundary
- "Just" pull the scale-valued function p out of the inner product and negate the sign

$$\left\langle \nabla^2, A \right\rangle = \sum_{i,j} \nabla^2_{ij} A_{ij}, \quad \text{in particular} \quad \left\langle \nabla^2, pI \right\rangle = \Delta p.$$

- $A : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $p : \mathbb{R}^d \to \mathbb{R}$
- Again, not to confuse the scalar $\langle \nabla^2, A \rangle$ with the tensor $\nabla^2 A$
- We omit arguments of functions whenever no confusion will result in
  - for instance, we will rewrite integration by parts simply as

$$\int \langle \nabla p, \mathbf{b} \rangle = -\int p \langle \nabla, \mathbf{b} \rangle \iff \int \langle \nabla p + p\nabla, \mathbf{b} \rangle = \int \langle \nabla, p\mathbf{b} \rangle = 0$$

  - it should be clear that p and $\mathbf{b}$ are functions, whose arguments are being integrated over

Fokker-Planck-Kolmogorov Equation

$$\partial_t p = \mathsf{L}_t^* p = -\langle \nabla, p\mathbf{b}\rangle + \frac{1}{2}\left\langle \nabla^2, pA\right\rangle \qquad \text{(forward)}$$

$$-\partial_s p = \mathsf{L}_s p = \langle \mathbf{b}, \nabla p\rangle + \frac{1}{2}\left\langle A, \nabla^2 p\right\rangle \qquad \text{(backward)}$$

FPK for SDE

Consider the stochastic differential equation

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

and its generator

$$\mathsf{L}_t f(\mathbf{x}) := \langle \mathbf{b}_t(\mathbf{x}), \nabla f(\mathbf{x}) \rangle + \frac{1}{2} \left\langle G_t(\mathbf{x}) G_t(\mathbf{x})^\top, \nabla^2 f(\mathbf{x}) \right\rangle$$

Then,

$$\boxed{\partial_t p = \mathsf{L}_t^* p = -\langle \nabla_{\mathbf{y}}, p\mathbf{b} \rangle + \frac{1}{2} \left\langle \nabla^2, pGG^\top \right\rangle}$$

$$\boxed{\partial_s p = \mathsf{L}_s p = \langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \left\langle GG^\top, \nabla^2 p \right\rangle}$$

- Continuity equation when $G = 0$

Wasserstein Gradient

- Wasserstein inner product on two functions h_1 and h_2 with $\int h_1 = \int h_2 = 0$:

$$\langle h_1, h_2 \rangle_p := \int \langle \nabla\varphi_1, \nabla\varphi_2 \rangle \cdot p \, \mathrm{d}\mathbf{x}, \quad \text{where} \quad \varphi_i \quad \text{solves} \ \langle \nabla, p\nabla\varphi_i \rangle = -h_i$$

- Wasserstein gradient represents derivative w.r.t. Wasserstein inner product:

$$\langle \nabla_{\mathrm{W}_2} f(p), \partial_t p_t \restriction_{t=0} \rangle_p = \frac{\mathrm{d} f(p_t)}{\mathrm{d}t} \restriction_{t=0},$$

where $p_t : (-\epsilon, \epsilon) \to \mathcal{P}_2$ is any smooth curve with $p_0 = p$

- Explicit formula through L_2 gradient:

$$\boxed{\nabla_{\mathrm{W}_2} f(p) = - \langle \nabla, p \nabla \nabla_{L_2} f(p) \rangle}$$

FPK as Gradient Flow

$$\mathrm{d}\mathsf{X}_t = -\nabla\varphi(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2\beta}\,\mathrm{d}\mathsf{B}_t$$
$$\partial_t p = \langle \nabla, p\nabla\varphi\rangle + \beta\Delta p = \langle \nabla, p(\nabla\varphi + \beta\mathbf{s}_p)\rangle$$

- If φ does not grow too fast, unique solution of FPK, a.k.a. Boltzmann-Gibbs:

$$\mathbf{s}_p = -\nabla\varphi/\beta \iff p \propto \exp(-\varphi/\beta)$$

- Lyapunov function:

$$f(p) = \int p\varphi + \beta p\log p - \beta p = \beta\mathsf{KL}(p\|q) + f_\star, \;\; q \propto \exp(-\varphi/\beta), \;\; f_\star := \inf f = c(\beta)$$

- FPK equation becomes the Wasserstein gradient flow:

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = -\nabla_{\mathrm{W}_2} f(p_t)$$

Assuming φ is λ-convex, we have

$$\frac{\mathrm{d}f(p_t)}{\mathrm{d}t} = \langle \nabla_{\mathbb{W}_2} f(p_t), \partial_t p_t \rangle_{p_t} = -\langle \nabla_{\mathbb{W}_2} f(p_t), \nabla_{\mathbb{W}_2} f(p_t) \rangle_{p_t} \leq -2\lambda[f(p_t) - f_\star]$$

$$\boxed{f(p_t) - f_\star \leq e^{-2\lambda t}[f(p_0) - f_\star]}$$

$$\mathbb{W}_2^2(p_t, q) \leq \tfrac{2}{\lambda}[f(p_t) - f_\star] \leq \tfrac{2}{\lambda} e^{-2\lambda t}[f(p_0) - f_\star]$$

$$\tfrac{1}{2}\|p_t - q\|_1^2 \leq \mathsf{KL}(p_t \| q) = [f(p_t) - f_\star]/\beta \leq e^{-2\lambda t} \cdot \tfrac{f(p_0) - f_\star}{\beta}$$

Log-Sobolev Inequality

Consider the Boltzmann-Gibbs density $q \propto \exp(-\varphi/\beta)$ for some λ-convex φ. Then,

$$\beta\mathsf{KL}(p\|q) = f(p) - f_\star \leq \frac{1}{2\lambda}\left\langle \nabla_{\mathrm{W}_2} f(p), \nabla_{\mathrm{W}_2} f(p)\right\rangle_p = \frac{1}{2\lambda}\int \|\nabla\varphi + \beta\mathbf{s}_p\|_2^2 \cdot p\,\mathrm{d}\mathbf{x}$$
$$\leq \frac{\beta^2}{2\lambda}\int \|\mathbf{s}_q - \mathbf{s}_p\|_2^2 \cdot p\,\mathrm{d}\mathbf{x}$$

To put in a more succinct and familiar form:

$$\boxed{\mathsf{KL}(p\|q) \leq \frac{\beta}{2\lambda}\mathsf{F}(p\|q)}, \quad \text{where} \quad q \propto \exp(-\varphi/\beta) \text{ for some } \lambda\text{-convex } \varphi$$

Reverse-time SDE

$$\mathrm{d}X_t = \mathbf{b}_t(X_t)\,\mathrm{d}t + G_t(X_t)\,\mathrm{d}B_t \qquad \text{(forward-SDE)}$$

$$\mathrm{d}\overleftarrow{X}_t = \overleftarrow{\mathbf{b}}_t(\overleftarrow{X}_t)\,\mathrm{d}t + \overleftarrow{G}_t(\overleftarrow{X}_t)\,\mathrm{d}\overleftarrow{B}_t \qquad \text{(reverse-SDE)}$$

- FPK to reverse-SDE (negation due to time reversal: $\overleftarrow{X}_t = X_{1-t}$):

$$-\partial_s \overleftarrow{p}(s, \mathbf{x}, t, \mathbf{y}) = -\left\langle \nabla, \overleftarrow{p}\overleftarrow{\mathbf{b}} \right\rangle + \tfrac{1}{2}\left\langle \nabla^2, \overleftarrow{p}\overleftarrow{A} \right\rangle, \quad \text{where} \quad \overleftarrow{A} := \overleftarrow{G}\overleftarrow{G}^\top$$

- FPK to forward-SDE for $p(s, \mathbf{x}, t, \mathbf{y})$ and $q(s, \mathbf{x})$:

$$-\partial_s \log p = \frac{\langle \mathbf{b}, \nabla p\rangle + \frac{1}{2}\langle A, \nabla^2 p\rangle}{p}, \quad -\partial_s \log q = \frac{\langle \nabla, q\mathbf{b}\rangle - \frac{1}{2}\langle \nabla^2, qA\rangle}{q}, \quad A := GG^\top$$

B. D. O. Anderson. "Reverse-time diffusion equation models". *Stochastic Processes and their Applications*, vol. 12, no. 3 (1982), pp. 313–326.

Let $r = pq$ (the joint density of X_s and X_t). We add the above two equations:

$$
\begin{aligned}
-\partial_s \log r &= \frac{\langle \mathbf{b}, \nabla p \rangle + \frac{1}{2}\langle A, \nabla^2 p \rangle}{p} + \frac{\langle \nabla, q\mathbf{b} \rangle - \frac{1}{2}\langle \nabla^2, qA \rangle}{q} \\
&= \langle \mathbf{b}, \nabla \log p \rangle + \frac{1}{2p}\left\langle A, \nabla^2 p \right\rangle + \langle \nabla \log q, \mathbf{b} \rangle + \langle \nabla, \mathbf{b} \rangle - \frac{1}{2q}\left\langle \nabla^2, qA \right\rangle \\
&= \frac{1}{r}[\langle \nabla, r\mathbf{b} \rangle + \frac{q}{2}\left\langle A, \nabla^2 p \right\rangle - \frac{p}{2}\left\langle \nabla^2, qA \right\rangle] \\
&= \frac{1}{r}[\langle \nabla, r\mathbf{b} \rangle + \frac{1}{2}\left\langle qA, \nabla^2 p \right\rangle - \frac{1}{2}\langle p\nabla, \nabla \cdot (qA) \rangle] \\
&= \frac{1}{r}[\langle \nabla, r\mathbf{b} \rangle + \frac{1}{2}\left\langle qA, \nabla^2 p \right\rangle + \frac{1}{2}\langle \nabla p, \nabla \cdot (qA) \rangle - \frac{1}{2}\langle \nabla p, \nabla \cdot (qA) \rangle - \\
&\qquad\qquad - \frac{1}{2}\langle p\nabla, \nabla \cdot (qA) \rangle] \\
&= \frac{1}{r}[\langle \nabla, r\mathbf{b} \rangle + \frac{1}{2}\langle \nabla, (\nabla p) \cdot (qA) \rangle - \frac{1}{2}\langle \nabla, p\nabla \cdot (qA) \rangle] \\
&= \frac{1}{r}[\langle \nabla, r\mathbf{b} \rangle + \frac{1}{2}\langle \nabla, \nabla \cdot (pqA) \rangle - \langle \nabla, p\nabla \cdot (qA) \rangle] \\
&= \frac{1}{r}\left[\left\langle \nabla, r(\mathbf{b} - \frac{1}{q}\nabla \cdot (qA)) \right\rangle + \frac{1}{2}\left\langle \nabla^2, rA \right\rangle\right]
\end{aligned}
$$

$$\boxed{-\partial_s r = \left\langle \nabla, r(\mathbf{b} - \tfrac{1}{q}\nabla \cdot (qA)) \right\rangle + \tfrac{1}{2}\left\langle \nabla^2, rA \right\rangle}$$

- Dividing both sides by $q(t, \mathbf{y})$ (and noting that ∇ and ∇^2 are w.r.t. $\mathbf{x}$):

$$\begin{aligned} -\partial_s \overleftarrow{p}(s, \mathbf{x}, t, \mathbf{y}) &= \left\langle \nabla, \overleftarrow{p}(\mathbf{b} - \tfrac{1}{q}\nabla \cdot (qA)) \right\rangle + \tfrac{1}{2}\left\langle \nabla^2, \overleftarrow{p}A \right\rangle \\ &= \left\langle \nabla, \overleftarrow{p}(\mathbf{b} - A\mathbf{s}_q + \tfrac{1}{2}A\mathbf{s}_{\overleftarrow{p}} - \tfrac{1}{2}\nabla \cdot A) \right\rangle \end{aligned}$$

- Comparing with the FPK for reverse-SDE, we may identify

$$\overleftarrow{G}_{1-t} = G_t, \ \overleftarrow{\mathbf{b}}_{1-t} = -\mathbf{b}_t + \tfrac{1}{q}\nabla \cdot (qG_tG_t^\top), \quad \text{or},$$

$$\overleftarrow{G}_{1-t} = 0, \ \overleftarrow{\mathbf{b}}_{1-t} = -\mathbf{b}_t + G_tG_t^\top\mathbf{s}_q - \tfrac{1}{2}G_tG_t^\top\mathbf{s}_{\overleftarrow{p}} + \tfrac{1}{2}\nabla \cdot (G_tG_t^\top)$$

Expectation-Maxmization

- Given training data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \sim q(\mathbf{x})$, the data density
- Parameterize $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$, the joint model density, e.g. Gaussian mixture
- Estimate $\boldsymbol{\theta}$ by minimizing some "distance" between q (the unknown data density) and $p_{\boldsymbol{\theta}}$ (the chosen model density):

$$\min_{\boldsymbol{\theta}} \min_{q(\mathbf{z}|\mathbf{x})} \mathsf{KL}\big(q(\mathbf{x})q(\mathbf{z}|\mathbf{x}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\big) \approx -\frac{1}{n}\sum_{i=1}^{n} \int [\log q(\mathbf{z}|\mathbf{x}_i) - \log p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{z})] \cdot q(\mathbf{z}|\mathbf{x}_i)\, \mathrm{d}\mathbf{z}$$

$$\boxed{q(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}$$

- After training, can generate new data $\mathsf{X} \sim p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ (by discarding Z)
- Need a training sample from $q(\mathbf{x})$, an explicit form of $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ and $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$
  - Monte Carlo EM: can sample from $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

Variational Inference

$$\min_{\boldsymbol{\theta}} \min_{\boldsymbol{\phi}} \ \mathsf{KL}\Big(q(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\Big)$$

- Parameterize $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \cdot p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, with $p(\mathbf{z})$ standard Gaussian (say)
- Parameterize $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, in case the optimal solution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is hard to compute
- Encoder: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, from latent $\mathbf{z}$ to observation $\mathbf{x}$
- Decoder: $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, from observation $\mathbf{x}$ to latent $\mathbf{z}$
- After training, can generate new data $\mathsf{X} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\mathsf{Z})$, where $\mathsf{Z} \sim p(\mathbf{z})$
- With only a training sample from $q(\mathbf{x})$, $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$

VAE as Triangular Flow

- Consider reference densities $s(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \cdot q(\mathbf{x})$ and $r(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \cdot \mathcal{N}(\mathbf{x}; \mathbf{0}, I)$
  - recall that q is the (unknown) data density and p is say standard Gaussian

Theorem: Uniqueness for increasing triangular maps

For any two densities r and p on $\mathbb{R}^d$, there exists a unique (up to permutation) increasing triangular map $\mathbf{T}$ so that $p = \mathbf{T}_{\#} r$.

- It follows that $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = (\mathbf{T}_{\boldsymbol{\theta}} \times \mathrm{Id})_{\#} r$, where $\mathbf{T}_{\boldsymbol{\theta}} : \mathbb{R}^{z+x} \to \mathbb{R}^x$
- Similarly, $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}) = q(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = (\mathrm{Id} \times \mathbf{S}_{\boldsymbol{\phi}})_{\#} s$, where $\mathbf{S}_{\boldsymbol{\phi}} : \mathbb{R}^{z+x} \to \mathbb{R}^z$

A Trivial Look

$$\mathsf{KL}\big(q(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\big) = \mathsf{KL}\big((\mathrm{Id} \times \mathbf{S}_{\boldsymbol{\phi}})_{\#}s \| (\mathbf{T}_{\boldsymbol{\theta}} \times \mathrm{Id})_{\#}r\big)$$

- Can apply change-of-variable to compute density of $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = (\mathbf{T}_{\boldsymbol{\theta}} \times \mathrm{Id})_{\#}r$
- Can sample from $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}) = (\mathrm{Id} \times \mathbf{S}_{\boldsymbol{\phi}})_{\#}s$; recall $s(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \cdot q(\mathbf{x})$
  - e.g. $S_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}) = \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}) + \sigma_{\boldsymbol{\phi}}(\mathbf{x}) \odot \mathbf{z}$

$$\mathsf{KL}\big(q(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\big) \equiv \underbrace{-\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})} \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}_{\text{reconstruction}} + \underbrace{\mathbb{E}_{q(\mathbf{x})}\Big[\mathsf{KL}\big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}), p(\mathbf{z})\big)\Big]}_{\text{regularization}}$$

Euler-Maruyama

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

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

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

  - $\Delta B_{t_k} \overset{i.i.d.}{\simeq} \mathcal{N}(0, \Delta t_k)$

Score Matching

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

- 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:

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

- 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.