CS480/680: Introduction to Machine Learning

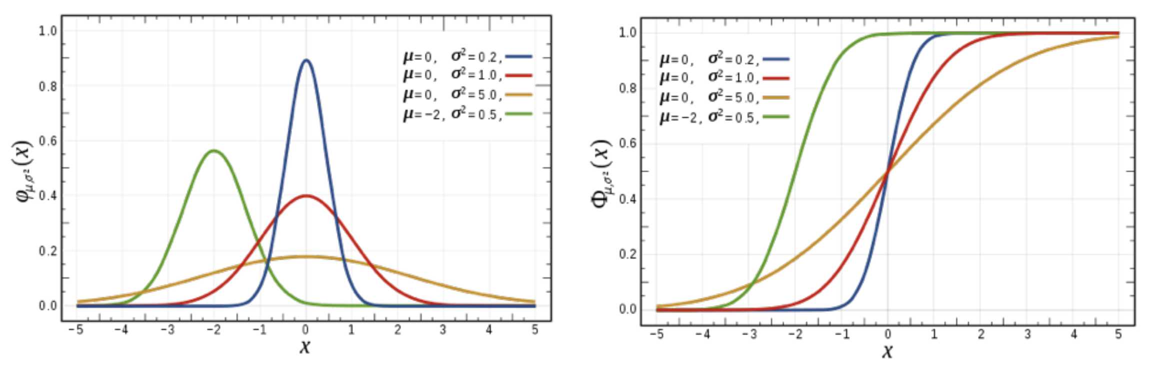
Lec 17: Variational Auto-Encoders

Yaoliang Yu

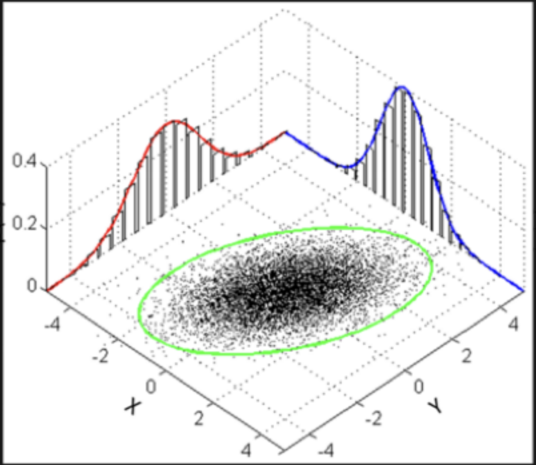
UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

July 08, 2024

Recap: Gaussian Distribution



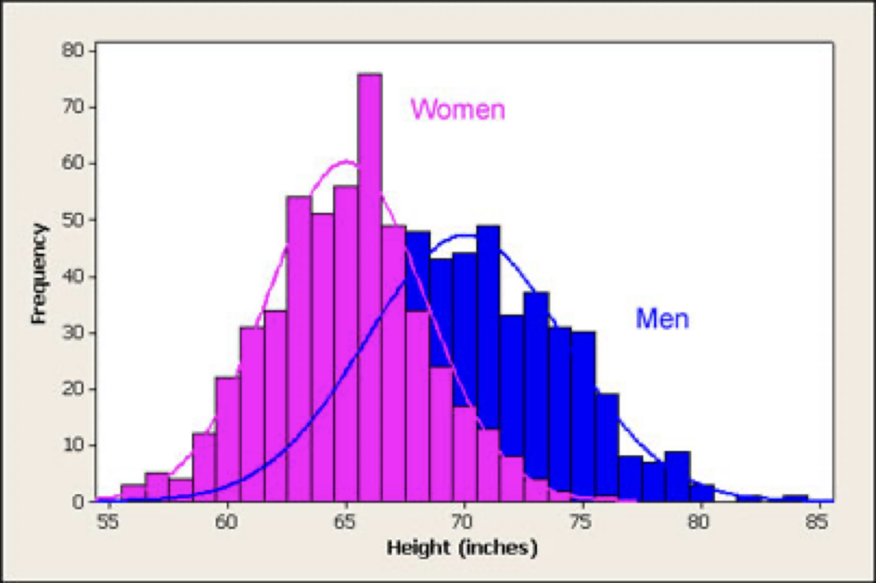
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



$$p(\mathbf{x}) = (2\pi)^{-d/2}[\det(S)]^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- Linear transformation of Gaussian is Gaussian

Multi-modality



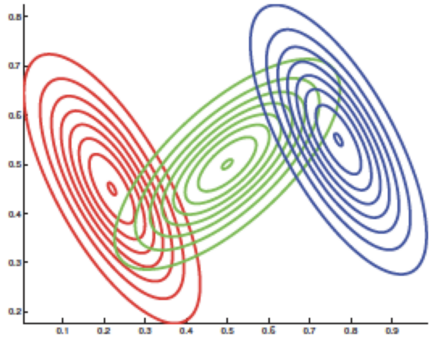
Mixture Models

$$p(\mathbf{x}|\boldsymbol{\theta}) := \sum_{k=1}^{K} \underbrace{\overbrace{p(z=k)}^{:=\lambda_k} \cdot \overbrace{p(\mathbf{x}|z=k,\boldsymbol{\theta})}^{p_k(\mathbf{x}|\boldsymbol{\theta})}}_{p(\mathbf{x},z=k|\boldsymbol{\theta})}$$

- K: number of components; chosen beforehand
- $\lambda_k := p(z=k)$: mixing distribution, a.k.a.prior over the latent variable z
- $\boldsymbol{\theta}$: parameters; for convenience, we lump all parameters, including $\boldsymbol{\lambda}$, into $\boldsymbol{\theta}$
- $p_k(\mathbf{x}|\boldsymbol{\theta}) := p(\mathbf{x}|z=k,\boldsymbol{\theta})$: k-th component density, a.k.a. conditional
- $p(\mathbf{x},z=k|\boldsymbol{\theta})$: joint density between $\mathbf{x}$ and z
- $p(\mathbf{x})$: marginal over $\mathbf{x}$, the observed variable
- $p(z=k|\mathbf{x},\boldsymbol{\theta})$: posterior; given observation $\mathbf{x}$, infer latent z

Gaussian Mixture Models (GMM)

$$p(\mathbf{x}|\{\lambda_k, \boldsymbol{\mu}_k, S_k\}) = \sum_{k=1}^{K} \lambda_k \cdot (2\pi)^{-d/2}[\det(S_k)]^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top S_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right]$$



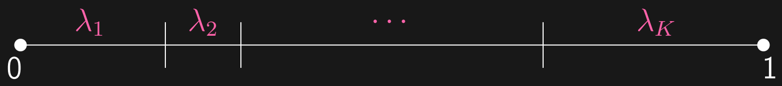
(a)


(b)

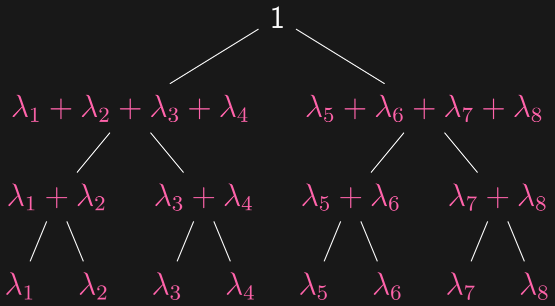
Sampling from Mixture Models

$$p(\mathbf{x}|\boldsymbol{\theta}) := \sum_{k=1}^{K} \underbrace{\overbrace{p(z=k)}^{:=\lambda_k} \cdot \overbrace{p(\mathbf{x}|z=k,\boldsymbol{\theta})}^{p_k(\mathbf{x}|\boldsymbol{\theta})}}_{p(\mathbf{x},z=k|\boldsymbol{\theta})}$$

- Sample Z according to the mixing distribution $\lambda_k := p(z=k)$
- Sample X according to the conditional $p(\mathbf{x}|\mathsf{Z}=z,\boldsymbol{\theta})$
- (X, Z) form a sample from the joint density $p(\mathbf{x}, z)$
- Discard Z, X alone forms a sample from the marginal $p(\mathbf{x})$



- Draw $\mathsf{U} \sim \mathrm{Uniform}(0,1)$
- Find which interval U lies in



Estimation From Data

- Given i.i.d. sample $X_1, X_2, \ldots, X_n \sim q(\mathbf{x})$, the unknown true data density
  - often replace $q(\mathbf{x})$ with $\hat{q}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\mathbf{x})$, i.e. a delta mass on each X_i w.p. $\frac{1}{n}$
- Model density that we choose (may or may not be correct):

$$\boxed{p_{\boldsymbol{\theta}}(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}) := \int p_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\, \mathrm{d}\mathbf{z}}$$

- The variables $Z_1, \ldots, Z_n$ are missing (unobserved)
- Estimate the parameters $\boldsymbol{\theta}$
- Different methods under different constraints and parameterizations

Minimizing KL = Maximum Likelihood Estimation (MLE)

$$\boxed{\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \,\|\, p_{\boldsymbol{\theta}}) \qquad \equiv \qquad \max_{\boldsymbol{\theta}} \ \mathop{\mathbb{E}}_{\mathsf{X}\sim q} \log p_{\boldsymbol{\theta}}(\mathsf{X})}$$

- KL divergence: $\mathsf{KL}(q\|p) := \mathop{\mathbb{E}}_{\mathsf{X}\sim q} \log \frac{q(\mathsf{X})}{p(\mathsf{X})} = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x}$
  - ≥ 0, with equality iff $q = p$
  - asymmetric: $\mathsf{KL}(q\|p) \neq \mathsf{KL}(p\|q)$
- For Gaussian Mixture Models (GMM):

$$\max_{\lambda_k, \boldsymbol{\mu}_k, S_k} \frac{1}{n} \sum_{i=1}^{n} \log \left[\sum_k \lambda_k (2\pi)^{-d/2} [\det(S_k)]^{-1/2} \exp\left[-\tfrac{1}{2} (\mathsf{X}_i - \boldsymbol{\mu}_k)^\top S_k^{-1} (\mathsf{X}_i - \boldsymbol{\mu}_k) \right] \right]$$

- Plug in the marginal density of mixtures:

$$\max_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{\mathsf{X}\sim q} \log \underbrace{\int p_{\boldsymbol{\theta}}(\mathsf{X}, \mathbf{z}) \, \mathrm{d}\mathbf{z}}_{p_{\boldsymbol{\theta}}(\mathsf{X})} \implies \frac{\partial}{\partial \boldsymbol{\theta}} = \int \int \frac{\partial_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{z}$$

$$= \int \int \frac{\partial_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x}) \cdot p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})} \cdot q(\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) \cdot \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{z}$$

$$= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Z})\sim q(\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})} \left[\partial_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathsf{X}, \mathsf{Z})\right]$$

- Hard to solve in general: gradient ascent often converges to poor local maxima

The Power of Lifting

$$\min_{\mathbf{w}} \ \|X\mathbf{w} - \mathbf{y}\|_1, \quad \text{where} \quad \|\mathbf{a}\|_1 := \sum_j |a_j|$$

- A nice trick:

$$|t| = \tfrac{1}{2} \cdot \min_s \ [t^2/s^2 + s^2]$$

- Apply component-wise:

$$\min_{\mathbf{w}} \min_{\mathbf{s}} \ \tfrac{1}{2}\|\tfrac{1}{\mathbf{s}} \odot (X\mathbf{w} - \mathbf{y})\|_2^2 + \tfrac{1}{2} \cdot \mathbf{1}^\top \mathbf{s}^2$$

- Fix $\mathbf{w}$, find $\mathbf{s}^2 = |X\mathbf{w} - \mathbf{y}|$
- Fix $\mathbf{s}$, find $\mathbf{w}$ by $\operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\tfrac{1}{\mathbf{s}} \odot (X\mathbf{w} - \mathbf{y})\|_2^2 + \tfrac{1}{2} \cdot \mathbf{1}^\top \mathbf{s}^2$

Expectation-Maximization (EM)

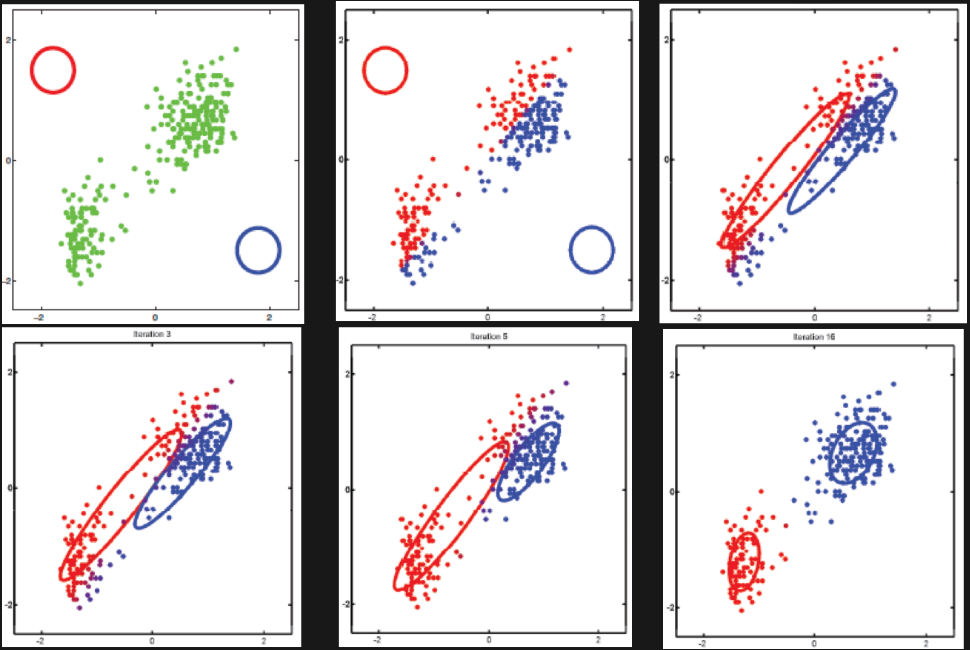
$$\mathsf{KL}\big(q(\mathbf{x},\mathbf{z}) \,\|\, p(\mathbf{x},\mathbf{z})\big) = \mathsf{KL}\big(q(\mathbf{x}) \,\|\, p(\mathbf{x})\big) + \mathop{\mathbb{E}}_{\mathbf{x}\sim q}\Big[\mathsf{KL}\big(q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}|\mathbf{x})\big)\Big]$$

- Expectation-Maximization (EM):

$$\boxed{\min_{q(\mathbf{z}|\mathbf{x})} \min_{\boldsymbol{\theta}} \; \mathsf{KL}\big(q(\mathbf{x}) \cdot q(\mathbf{z}|\mathbf{x}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\big)}$$

- Fix $\boldsymbol{\theta}$, $\boxed{q(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}$: assuming conditional density $p_\theta(\mathbf{z}|\mathbf{x})$ easy to compute
- Fix $q(\mathbf{z}|\mathbf{x})$, find $\boldsymbol{\theta}$ by $\boxed{\mathop{\mathrm{argmax}}_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Z})\sim q(\mathbf{x})q(\mathbf{z}|\mathbf{x})} [\log p_{\boldsymbol{\theta}}(\mathsf{X},\mathsf{Z})]}$
  - MLE on the joint density $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$, instead of the marginal $p_{\boldsymbol{\theta}}(\mathbf{x})$

A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1 (1977), pp. 1–38, I. Csiszár and G. Tusnády. "Information geometry and alternating minimization procedures". *Statistics & Decisions*, vol. Supplement, no. 1 (1984), pp. 205–237.


Iteration 3
Iteration 5
Iteration 16

Recap: Expectation-Maximization (EM)

- 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 $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

Auto-Encoder

- High dimensional input $\mathbf{x} \in \mathbb{R}^d$
- Linearly project (encode) to a low dimensional code $\mathbf{h} = V\mathbf{x} \in \mathbb{R}^k$
- Linearly reconstruct (decode) so that $\mathbf{x} \approx U\mathbf{h}$:

$$\min_{U \in \mathbb{R}^{d \times k}} \min_{V \in \mathbb{R}^{k \times d}} \mathbb{E}\|UV\mathbf{x} - \mathbf{x}\|_2^2$$

- Apply change-of-variable to arrive at the equivalent problem:

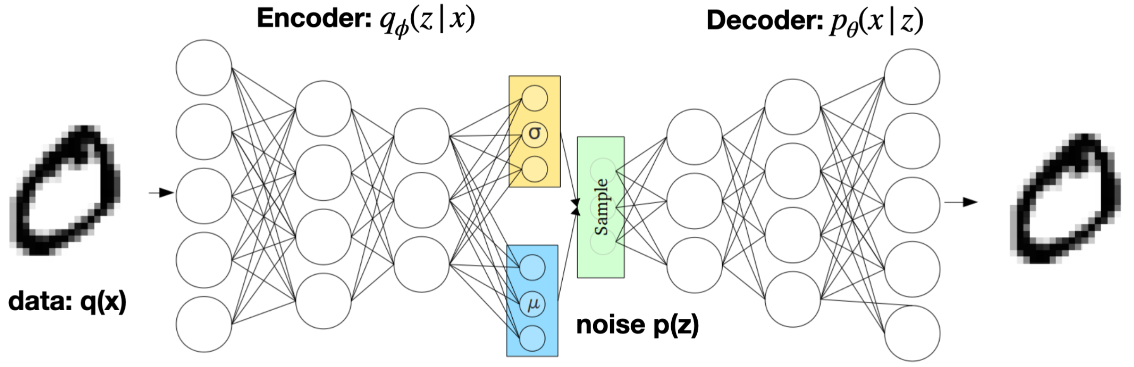
$$\min_{W \in \mathbb{R}^{d \times d}} \mathbb{E}\|W\mathbf{x} - \mathbf{x}\|_2^2, \quad \text{s.t.} \quad \operatorname{rank}(W) \leq k$$

- Optimal W given by principle component analysis (PCA)

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
- Decoder: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, from latent $\mathbf{z}$ to observation $\mathbf{x}$
- Encoder: $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, 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 $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$



D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representation*. 2014, D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning*. 2014.

A Closer Look

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

- Stochastic gradient w.r.t. $\boldsymbol{\theta}$ is standard as long as we have $p_{\boldsymbol{\theta}}$ explicitly
- Stochastic gradient w.r.t. ϕ can be computed via the log-trick:

$$\nabla_\phi \mathbb{E}_{q_\phi} f_\phi(\mathsf{X}, \mathsf{Z}) = \mathbb{E}_{q_\phi}[f_\phi \nabla_\phi \log(f_\phi q_\phi)]$$

- Can choose prior $p(\mathbf{z})$ and posterior $q_\phi(\mathbf{z}|\mathbf{x})$ so that the regularization term is available in closed-form
  - e.g., $\mathsf{KL}(\mathcal{N}(\mathbf{m}, S) \parallel \mathcal{N}(\mathbf{0}, \mathbb{I}_d)) = \frac{1}{2}\left[\operatorname{tr}(S) + \|\mathbf{m}\|_2^2 - 1 - \log\det S\right]$

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 $p = \mathbf{T}_{\#} r$.

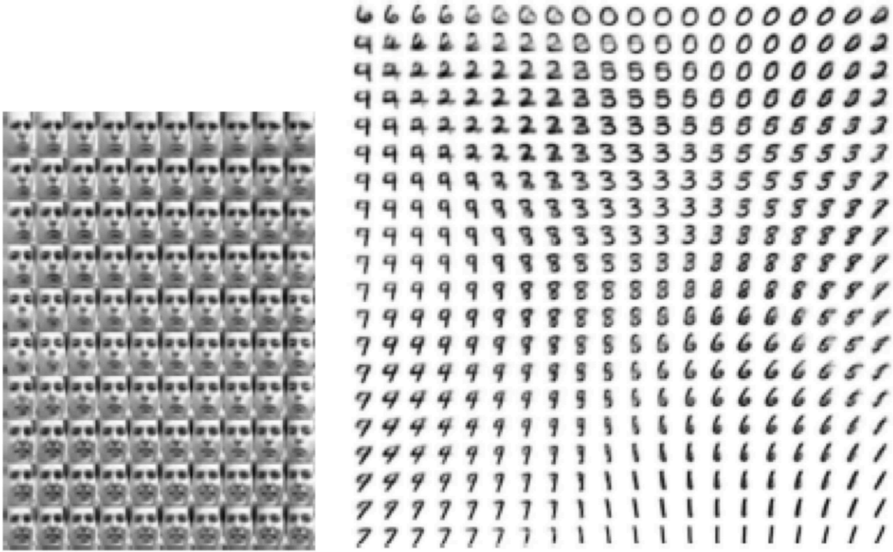
- It follows that $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = (\mathbf{T}_{\boldsymbol{\theta}} \times \mathrm{Id})_{\#} r$, where $\mathbf{T}_{\boldsymbol{\theta}} : \mathbb{R}^{z+x} \to \mathbb{R}^x$
- Similarly, $q_{\boldsymbol{\phi}}(\mathbf{x},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

A Trivial Look

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

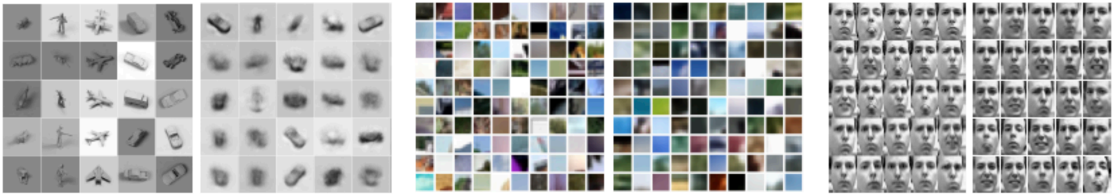
- Can apply change-of-variable to compute density of $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = (\mathbf{T}_{\boldsymbol{\theta}} \times \mathrm{Id})_\# r$
- Can sample from $q_\phi(\mathbf{x}, \mathbf{z}) = (\mathrm{Id} \times \mathbf{S}_\phi)_\# s$; recall $s(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})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

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



(a) Learned Frey Face manifold

(b) Learned MNIST manifold



(a) NORB (b) CIFAR (c) Frey