DSC190: Machine Learning with Few Labels

Unsupervised Learning

Zhiting Hu

Lecture 15, November 1st, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

Unsupervised learning: Variational Inference

Presentations

- **Peiyuan Sun**: Reasoning with Language Model is Planning with World Model
- **Mingyang Yao**: Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions
- **Zhaoxiang Feng**: Learning Equilibria in Matching Markets from Bandit Feedback
- **Bella Wang**: Language Models Are Realistic Tabular Data Generators

Recap: EM Algorithm for GMM

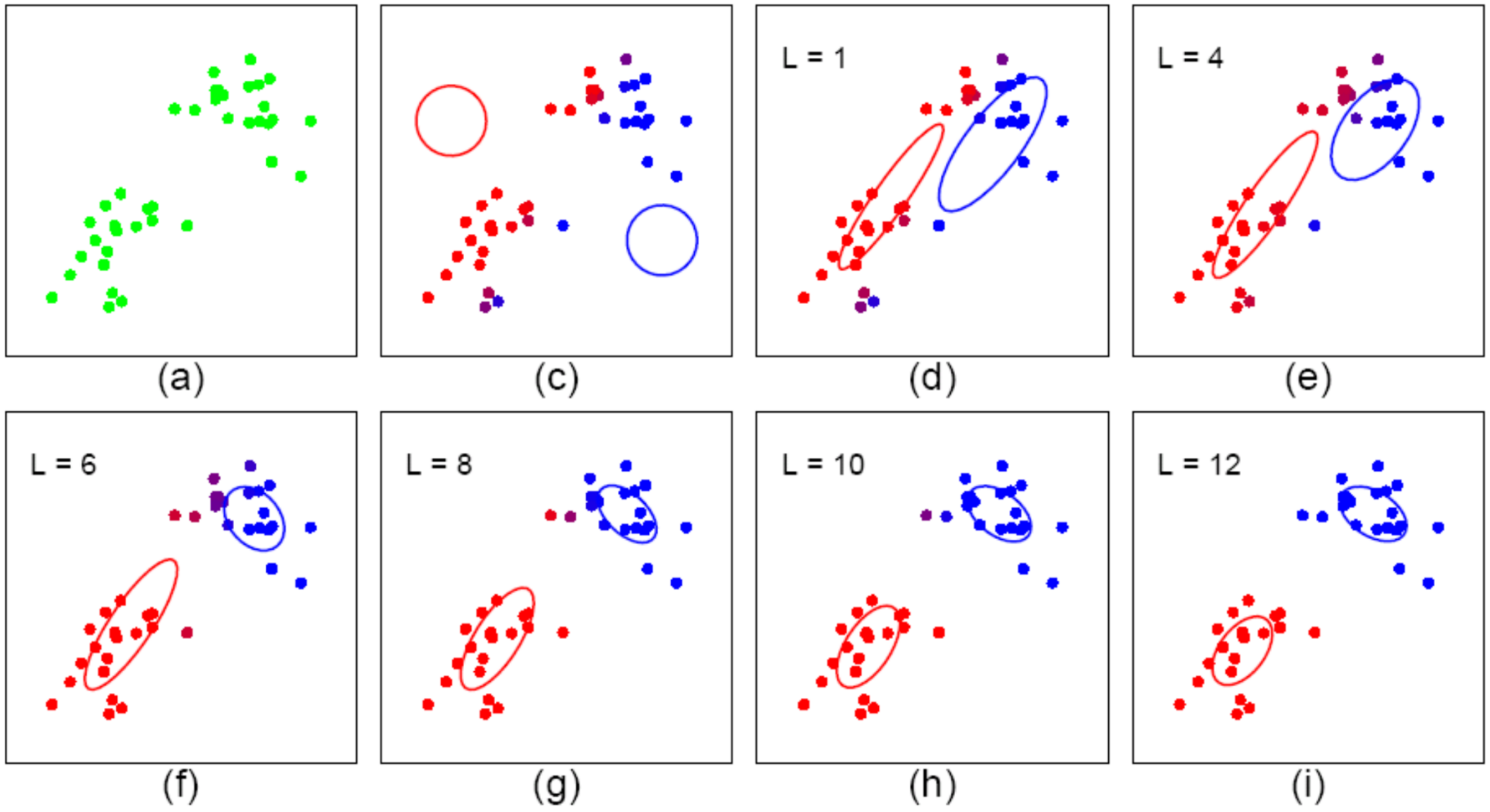
- Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k
- Iterate until convergence:
  - E-step: Evaluate the posterior given current parameters

$$p(z^k = 1 \mid \boldsymbol{x}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x} \mid \mu_j, \Sigma_j)} := \gamma_k$$

  - M-step: Re-estimate the parameters given current posterior

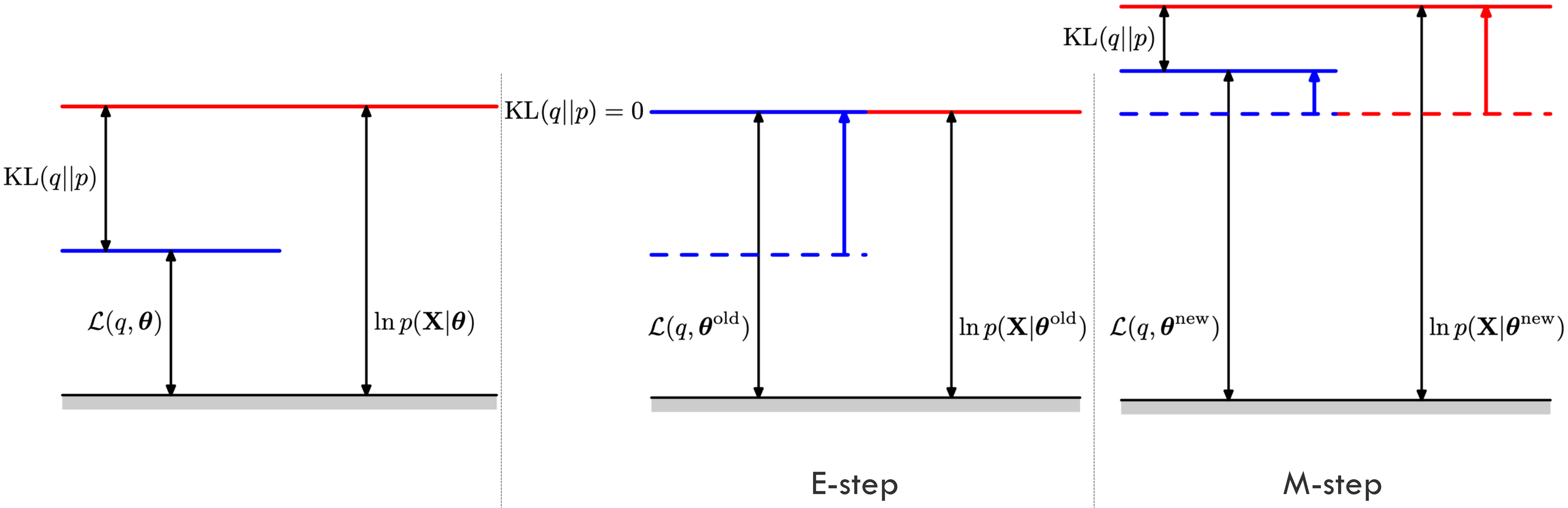
Recap: EM Algorithm for GMM

- Start: "guess" the centroid μ_k and covariance Σ_k of each of the K clusters
- Loop:



Each EM iteration guarantees to improve the likelihood

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$



[PRML, Chap 9.4]

Summary: EM Algorithm

- The EM algorithm is coordinate-decent on $F(q, \theta)$
  - E-step: $q^{t+1} = \arg\min_q F\left(q, \theta^t\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$
  - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) = \text{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

$$= -F(q, \theta) + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

Summary: EM Algorithm

- The EM algorithm is coordinate-decent on $F(q,\theta)$
  - E-step: $q^{t+1} = \arg\min_q F\left(q, \theta^t\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$
  - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) = \text{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

$$= -F(q,\theta) + \text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

- Limitation: need to be able to compute $p(\boldsymbol{z}|\boldsymbol{x}, \theta)$, not possible for more complicated models --- solution: Variational inference

Variational Inference

Content adapted from CMU 10-708 Spring 2017

Inference

- Given a model, the goals of inference can include:
  - Computing the likelihood of observed data $p(\boldsymbol{x}^*)$
  - Computing the marginal distribution over a given subset of variables in the model $p(\boldsymbol{x}_A)$
  - Computing the conditional distribution over a subsets of nodes given a disjoint subset of nodes $p(\boldsymbol{x}_A|\boldsymbol{x}_B)$
  - Computing a mode of the density (for the above distributions) $\mathrm{argmax}_x\, p(\boldsymbol{x})$
  - ....

Variational Inference

- Observed variables $\boldsymbol{x}$, latent variables $\boldsymbol{z}$
- Variational (Bayesian) inference, a.k.a. **variational Bayes**, is most often used to approximately infer the **posterior distribution** over the latent variables

$$p(\boldsymbol{z}|\boldsymbol{x},\theta) = \frac{p(\boldsymbol{z},\boldsymbol{x}|\theta)}{\sum_z p(\boldsymbol{z},\boldsymbol{x}|\theta)}$$

- We cannot directly compute the posterior distribution for many interesting models
  - I.e. the posterior density is in an intractable form (often involving integrals) which cannot be easily analytically solved.

EM and Variational Inference

- The EM algorithm:
  - E-step: $q^{t+1} = \arg\min_{q} F\left(q, \theta^{t}\right)$

    $= p(\boldsymbol{z}|\boldsymbol{x}, \theta^{t}) = \frac{p(\boldsymbol{z}, \boldsymbol{x}|\theta^{t})}{\sum_{z} p(\boldsymbol{z}, \boldsymbol{x}|\theta^{t})}$

    Intractable when model $p(\boldsymbol{z}, \boldsymbol{x}|\theta)$ is complex
  - M-step: $\theta^{t+1} = \arg\min_{\theta} F\left(q^{t+1}, \theta^{t}\right)$

→ Need to approximate $p(\boldsymbol{z}|\boldsymbol{x}, \theta^{t})$ with VI

Example: Bayesian mixture of Gaussians

- The mean μ_k is treated as a (latent) random variable

$$\mu_k \sim \mathcal{N}(0, \tau^2) \text{ for } k = 1, \ldots, K$$

- For each data $i = 1, \ldots, n$

$$z_i \sim \mathrm{Cat}(\pi).$$

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2).$$

- We have
  - observed variables $x_{1:n}$
  - latent variables $\mu_{1:k}$ and $z_{1:n}$
  - parameters $\{\tau^2, \pi, \sigma^2\}$
- $p(x_{1:n}, z_{1:n}, \mu_{1:k} | \tau^2, \pi, \sigma^2) = \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})$

Example: Bayesian mixture of Gaussians

- We can write the posterior distribution as

$$p(\mu_{1:K}, z_{1:n} | x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}$$

- The numerator can be computed for any choice of the latent variables
- The problem is the denominator (the marginal probability of the observations)
  - This integral cannot easily be computed analytically
- We need some approximation..

Variational Inference

Recall that in EM, we assume $q(z|x)$ can be any distribution. E-step shows the optimal $q(z|x)$ is the posterior distribution.

Variational Inference

Recall that in EM, we assume $q(z|x)$ can be any distribution. E-step shows the optimal $q(z|x)$ is the posterior distribution.

The main idea behind variational inference:

- Choose a family of distributions over the latent variables $z_{1:m}$ with its own set of variational parameters ν, i.e.

$$q(z_{1:m}|\nu)$$

- Then, we find the setting of the parameters that makes our approximation q closest to the posterior distribution.
  - This is where optimization algorithms come in.
- Then we can use q with the fitted parameters in place of the posterior.
  - E.g. to form predictions about future data, or to investigate the posterior distribution over the hidden variables, find modes, etc.

Variational Inference

- We want to minimize the KL divergence between our approximation $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{v})$ and our posterior $p(\boldsymbol{z}|\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{v}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}))$$

  - But we can't actually minimize this quantity w.r.t q because $p(\boldsymbol{z}|\boldsymbol{x})$ is unknown
- **Question:** how can we minimize the KL divergence?
  - **Hint**: recall the equation that holds for any q:

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

Variational Inference

- We want to minimize the KL divergence between our approximation $q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})$ and our posterior $p(\boldsymbol{z}|\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}))$$

  - But we can't actually minimize this quantity w.r.t q because $p(\boldsymbol{z}|\boldsymbol{x})$ is unknown

- **Question:** how can we minimize the KL divergence?

$$\ell(\theta;\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta)\big)$$

Evidence Lower Bound (ELBO)

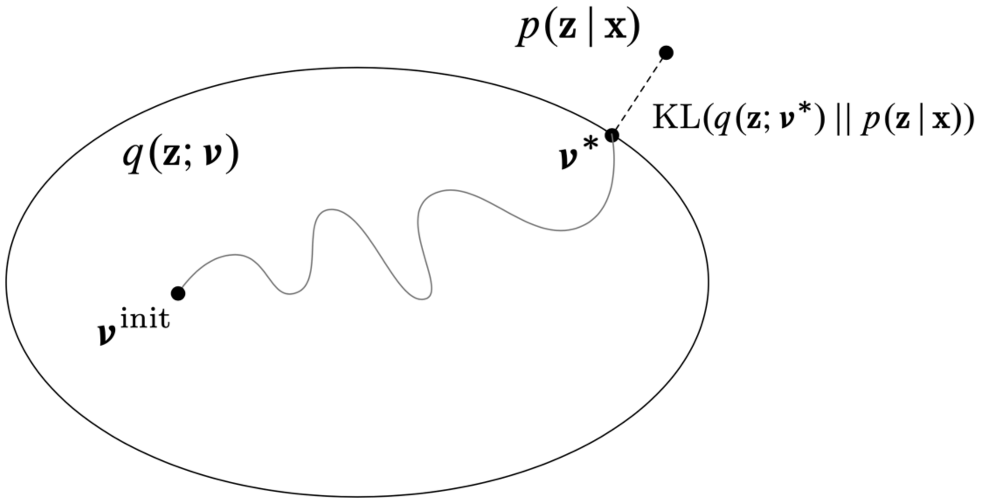
- The ELBO is equal to the negative KL divergence up to a constant $\ell(\theta;\boldsymbol{x})$
- We maximize the ELBO over q to find an "optimal approximation" to $p(\boldsymbol{z}|\boldsymbol{x})$

Variational Inference

- Choose a family of distributions over the latent variables $\boldsymbol{z}$ with its own set of variational parameters ν , i.e. $q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})$
- We maximize the ELBO over q to find an "optimal approximation" to $p(\boldsymbol{z}|\boldsymbol{x})$

$$\operatorname{argmax}_{\nu}\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}\right]$$

$$= \operatorname{argmax}_{\nu}\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}[\log q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})]$$

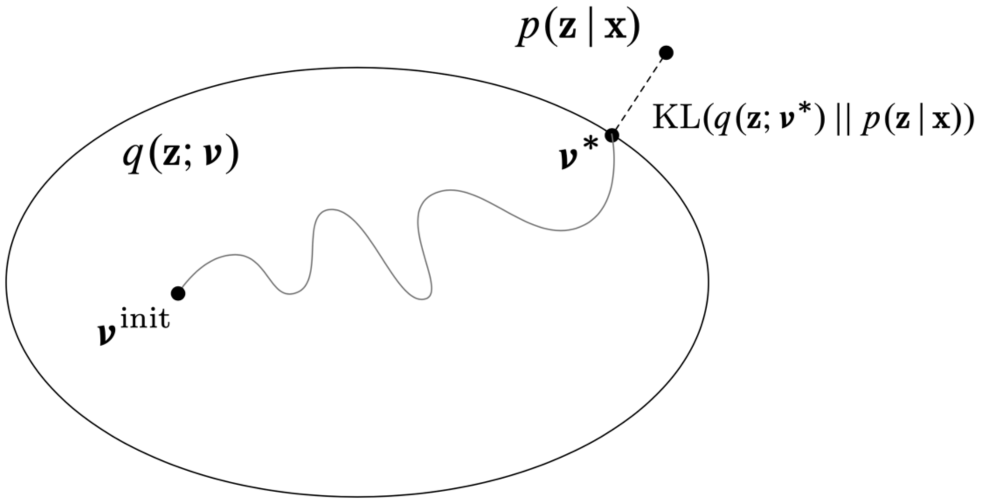


Variational Inference

- Choose a family of distributions over the latent variables $\boldsymbol{z}$ with its own set of variational parameters ν , i.e. $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{v})$
- We maximize the ELBO over q to find an "optimal approximation" to $p(\boldsymbol{z}|\boldsymbol{x})$

$$\operatorname{argmax}_{\nu} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\nu})}\right]$$

$$= \operatorname{argmax}_{\nu} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}[\log q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\nu})]$$



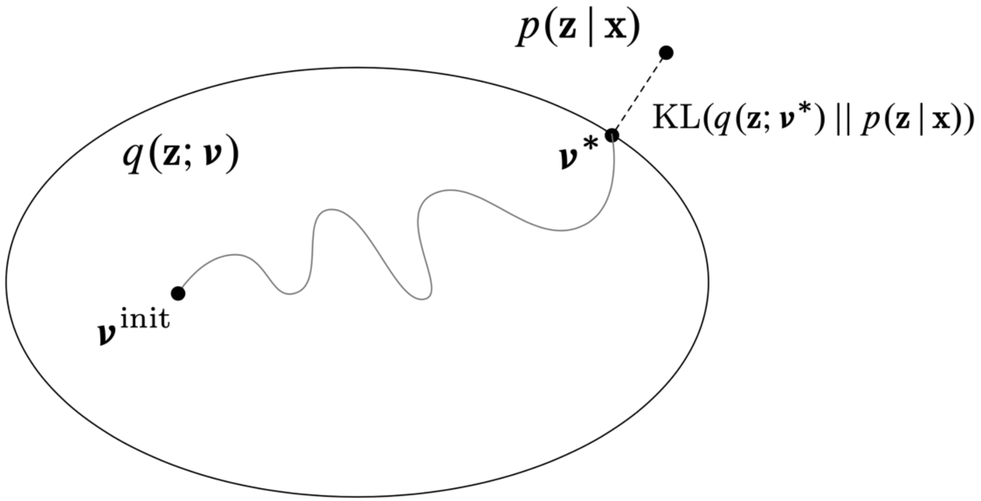
Question: How do we choose the variational family $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{v})$?

Variational Inference

- Choose a family of distributions over the latent variables $\boldsymbol{z}$ with its own set of variational parameters ν , i.e. $q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})$
- We maximize the ELBO over q to find an "optimal approximation" to $p(\boldsymbol{z}|\boldsymbol{x})$

$$\mathrm{argmax}_{\nu}\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}\right]$$

$$= \mathrm{argmax}_{\nu}\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})}[\ \log q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\nu})\ ]$$



Question: How do we choose the variational family $q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})$?

- Factorized distribution -> mean field VI
- Mixture of Gaussian distribution -> black-box VI
- Neural-based distribution -> Variational Autoencoders (VAEs)

Example: Mean Field Variational Inference

- A popular family of variational approximations

- In this type of variational inference, we assume the variational distribution over the latent variables factorizes as

$$q(\boldsymbol{z}) = q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j)$$

  - (where we omit variational parameters for ease of notation)
  - We refer to $q(z_j)$, the variational approximation for a single latent variable, as a "local variational approximation"

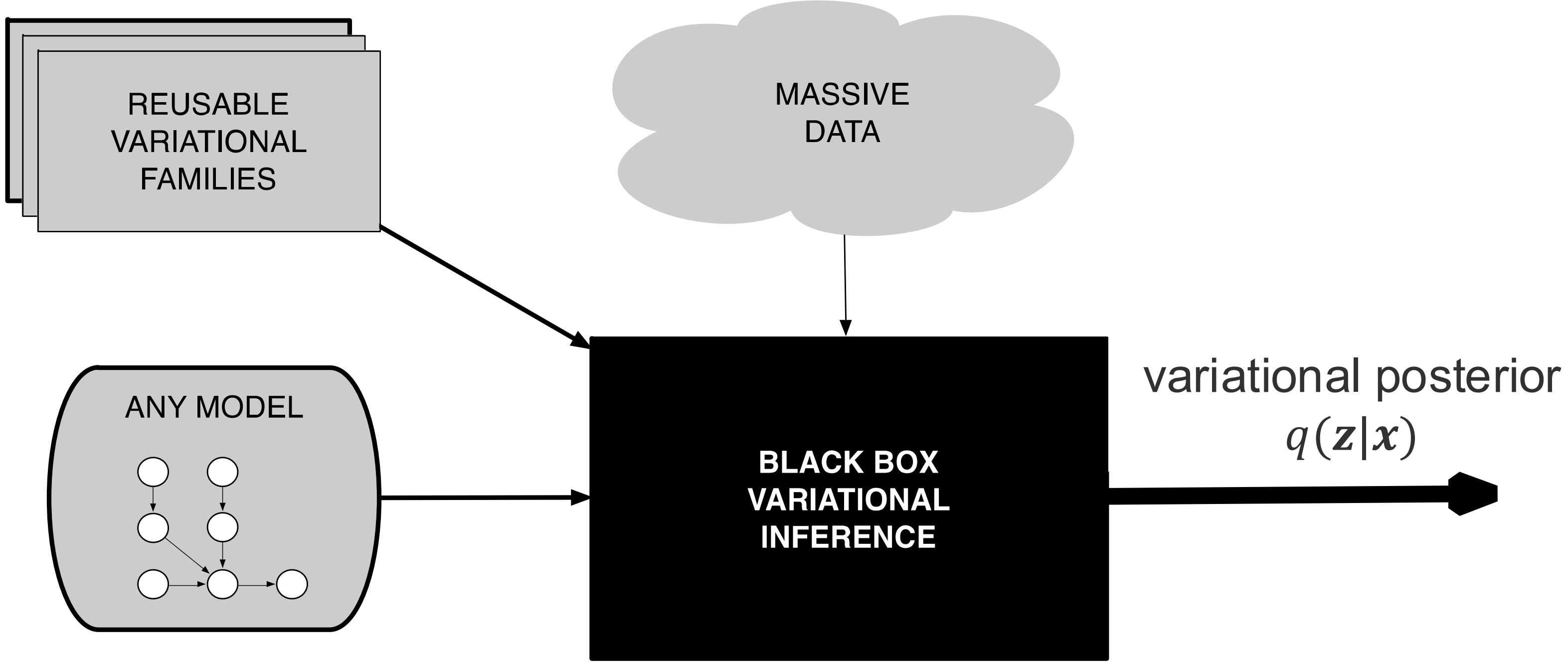
- In the above expression, the variational approximation $q(z_j)$ over each latent variable z_j is independent

Black-box Variational Inference

Black-box Variational Inference (BBVI)

- We have derived variational inference specific for Bayesian Gaussian (mixture) models
- There are innumerable models
- Can we have a solution that does not entail model-specific work?

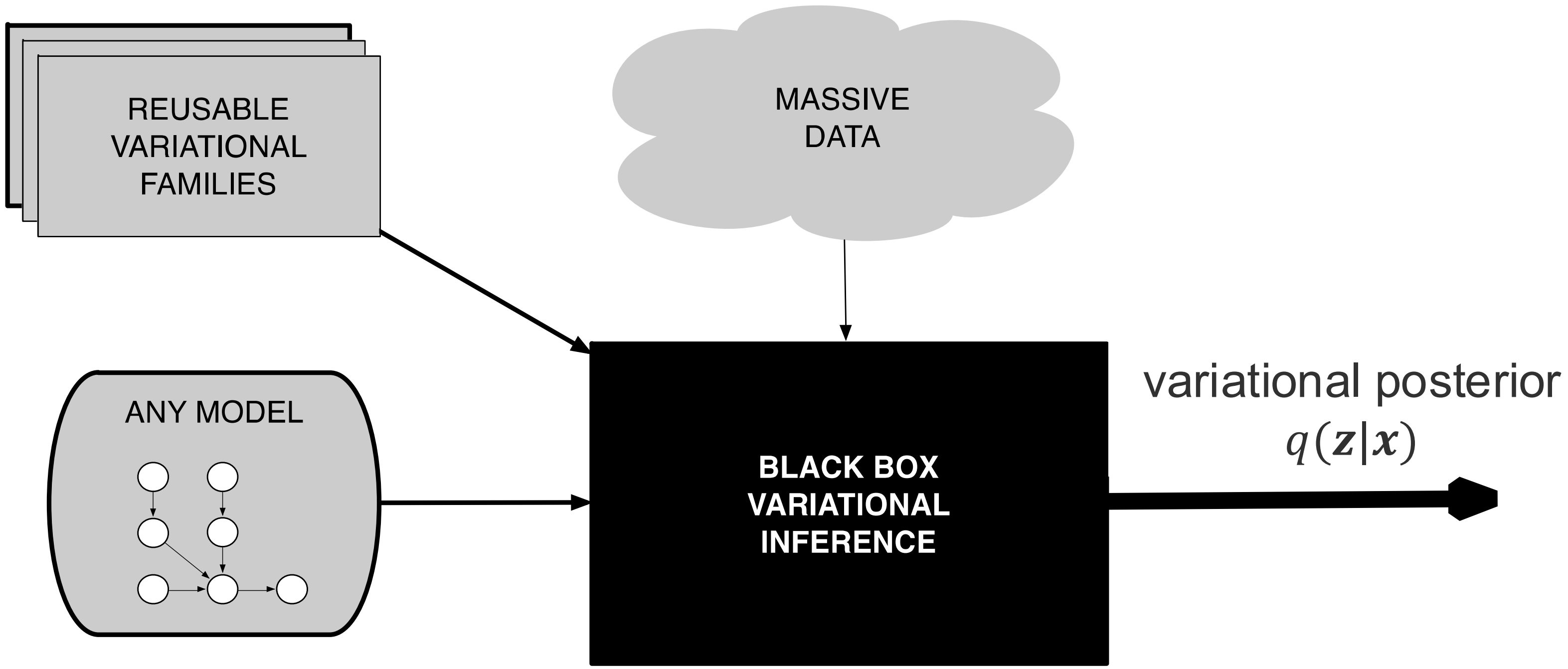
Black-box Variational Inference (BBVI)



- Easily use variational inference with **any model**
- Perform inference with **massive data**
- **No mathematical work** beyond specifying the model

(Courtesy: Blei et al., 2018)

Black-box Variational Inference (BBVI)



- Sample from $q(.)$
- Form noisy gradients (without model-specific computation)
- Use stochastic optimization

(Courtesy: Blei et al., 2018)

Black-box Variational Inference (BBVI)

- Probabilistic model: $\boldsymbol{x}$ -- observed variables, $\boldsymbol{z}$ -- latent variables
- Variational distribution $q_\lambda(\boldsymbol{z}|\boldsymbol{x})$ with parameters λ, e.g.,
  - Gaussian mixture distribution:
    - "A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific nonzero amount of error by a Gaussian mixture model with enough components." (Deep Learning book, pp.65)
  - Deep neural networks

- ELBO:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\lambda})}[\log p(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\lambda})}[\log q(\boldsymbol{z}|\lambda)]$$

- Want to compute the gradient w.r.t variational parameters λ

[Ranganath et al.,14]

Questions?