DSC190: Machine Learning with Few Labels

Supervised Learning, Unsupervised Learning

Zhiting Hu

Lecture 2, September 28, 2021

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- Office hours
  - Zhiting Hu: Tuesday 3-4pm, SDSC ~~247E~~249E
  - Meng Song: Wednesday 2:30-3:30pm, CSE 4109
- Project
- Presentation

Outline

- Supervised Learning
    - Maximum likelihood estimation (MLE)
    - Duality between MLE and Maximum Entropy Principle

- Unsupervised learning
    - Maximum likelihood estimation (MLE) with latent variables
    - EM algorithm for MLE

Supervised Learning

- Model to be learned $p_\theta(\boldsymbol{x})$
- Observe full data $\mathcal{D} = \{ \boldsymbol{x}^* \}$
    - i.i.d: independent, identically distributed

- Maximum Likelihood Estimation (MLE)
    - The most classical learning algorithm

$$\min_\theta - \mathbb{E}_{x^* \sim \mathcal{D}} \left[\quad \log p_\theta(\boldsymbol{x}^*) \quad \right]$$

- MLE is closely connected to the Maximum Entropy (MaxEnt) principle

Exponential Family

- A distribution

$$p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\} / Z(\boldsymbol{\theta})$$

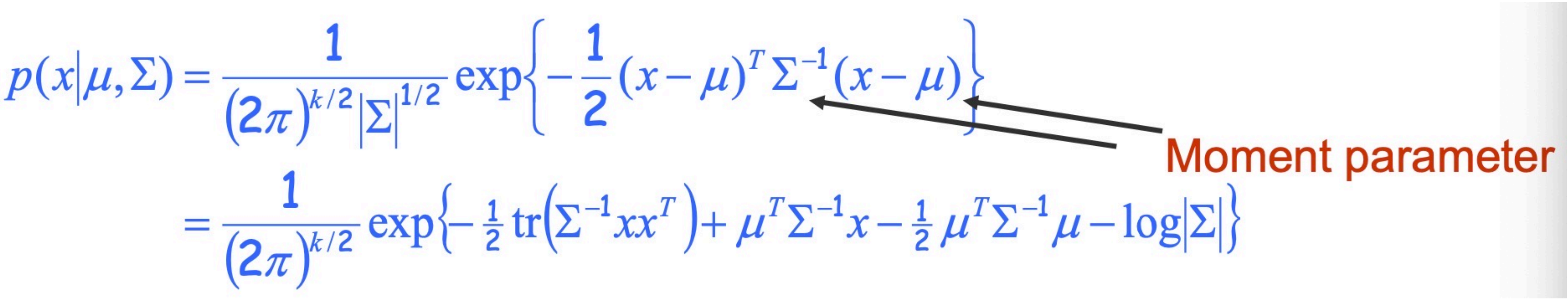
  is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
  - $T(\boldsymbol{x}) \in R^d$: sufficient statistics, features of data $\boldsymbol{x}$
  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\}$: normalization factor

- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Example: Multivariate Gaussian Distribution

- For a continuous vector random variable $\boldsymbol{x} \in R^k$



$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$= \frac{1}{(2\pi)^{k/2}} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\Sigma^{-1}xx^T\right) + \mu^T\Sigma^{-1}x - \tfrac{1}{2}\mu^T\Sigma^{-1}\mu - \log|\Sigma|\right\}$$

Moment parameter

- Exponential family representation

$$\boldsymbol{\theta} = \left[\Sigma^{-1}\mu; -\frac{1}{2}\operatorname{vec}\left(\Sigma^{-1}\right)\right] = \left[\boldsymbol{\theta}_1, \operatorname{vec}\left(\boldsymbol{\theta}_2\right)\right], \ \boldsymbol{\theta}_1 = \Sigma^{-1}\mu \text{ and } \boldsymbol{\theta}_2^{-} = -\frac{1}{2}\Sigma^{-1}$$

$$T(\boldsymbol{x}) = \left[\boldsymbol{x}; \operatorname{vec}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]$$

$$A(\boldsymbol{\theta}) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log|\Sigma| = -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\right) - \frac{1}{2}\log\left(-2\boldsymbol{\theta}_2\right)$$

$$h(x) = (2\pi)^{-k/2}$$

Example: Multinomial Distribution

- For a binary vector random variable $\boldsymbol{x} \in \mathrm{multi}(\boldsymbol{x}|\pi)$

$$
\begin{aligned}
p(x|\pi) &= \pi_1^{x_1}\pi_2^{x_1}\cdots\pi_K^{x_K} = \exp\left\{\sum_k x_k \ln \pi_k\right\} \\
&= \exp\left\{\sum_{k=1}^{K-1} x_k \ln \pi_k + \left(1-\sum_{k=1}^{K-1} x_K\right)\ln\left(1-\sum_{k=1}^{K-1}\pi_k\right)\right\} \\
&= \exp\left\{\sum_{k=1}^{K-1} x_k \ln\left(\frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k}\right) + \ln\left(1-\sum_{k=1}^{K-1}\pi_k\right)\right\}
\end{aligned}
$$

- Exponential family representation

$$
\begin{aligned}
\boldsymbol{\theta} &= \left[\ln\left(\pi_k/\pi_K\right);0\right] \\
T(\boldsymbol{x}) &= [\boldsymbol{x}] \\
A(\boldsymbol{\theta}) &= -\ln\left(1-\sum_{k=1}^{K-1}\pi_k\right) = \ln\left(\sum_{k=1}^{K} e^{\theta_k}\right) \\
h(\boldsymbol{x}) &= 1
\end{aligned}
$$

Maximum Likelihood for Exponential Family

$m(\boldsymbol{x})$: the number of times $\boldsymbol{x}$ is observed in *D*

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \left(\sum_i \theta_i T_i(\boldsymbol{x}) - \log Z(\boldsymbol{\theta}) \right) \\
&= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \sum_i \theta_i T_i(\boldsymbol{x}) - N \log Z(\boldsymbol{\theta})
\end{aligned}$$

- Take gradient and set to 0

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \frac{\partial}{\partial \theta_i} \log Z(\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})
\end{aligned}$$

$$\Rightarrow \boxed{\sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})} = \sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) = \boxed{\sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})}$$

At MLE, the expectations of the sufficient statistics under the model must match empirical feature average

Maximum Entropy (MaxEnt)

- Given $\mathcal{D}$, to estimate $p(\boldsymbol{x})$
- We can approach the problem from an entirely different point of view. Begin with some fixed feature expectations:

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) \coloneqq \alpha_i$$

- There may exist many distributions which satisfy them. Which one should we select?
  - MaxEnt principle: the most uncertain or flexible one, i.e., the one with maximum entropy
- This yields a new optimization problem:
  - This is a variational definition of a distribution!

$$\boxed{\begin{aligned} \max_{p} \;\; & \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \\ \text{s.t.} \;\; & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i \\ & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1 \end{aligned}}$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_{i} \theta_i \left(\sum_{x} p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\boldsymbol{x})} = 1 + \log p(\boldsymbol{x}) - \sum_i \theta_i T_i(\boldsymbol{x}) - \mu$$

$$p^*(\boldsymbol{x}) = e^{\mu - 1} \exp \left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu - 1} = \sum_{\boldsymbol{x}} \exp \left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\} \quad \left(\text{since} \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) = 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\boldsymbol{x})} = 1 + \log p(\boldsymbol{x}) - \sum_i \theta_i T_i(\boldsymbol{x}) - \mu$$

$$p^*(\boldsymbol{x}) = e^{\mu - 1} \exp \left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu - 1} = \sum_{\boldsymbol{x}} \exp \left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\} \quad \left(\text{since} \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) = 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

- So feature constraints + MaxEnt ⇒ exponential family.
- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

- So feature constraints + MaxEnt ⇒ exponential family.
- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

plug $p(x|\boldsymbol{\theta})$ back into L, and since $\sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) \coloneqq \alpha_i$:

$$\max_{\theta} L(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \sum_i \theta_i T_i(\boldsymbol{x}) - N \log Z(\boldsymbol{\theta})$$

- Recovers precisely the MLE problem of exponential family

(Homework)

Constraints from Data

- We have seen a case of **convex duality**:
  - In one case, we assume exponential family and show that Maximum Likelihood implies model expectations must match empirical expectations.
  - In the other case, we assume model expectations must match empirical feature counts and show that MaxEnt implies exponential family distribution.

A more general MaxEnt problem

$$\min_{p} \quad \mathrm{KL}(p(\boldsymbol{x}) \| h(\boldsymbol{x}))$$

$$\overset{\text{def}}{=} \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{h(\boldsymbol{x})} = -\mathrm{H}(p) - \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log h(\boldsymbol{x})$$

$$\text{s.t.} \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i$$

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

$$\Rightarrow \quad p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp \left\{ \sum_{i} \theta_i T_i(\boldsymbol{x}) \right\}$$

Summary

- Maximum entropy is dual to maximum likelihood of exponential family distributions
- This provides an alternative view of the problem of fitting a model into data:
  - The data instances in the training set are treated as constraints, and the learning problem is treated as a constrained optimization problem.
  - We'll revisit this optimization-theoretic view of learning repeatedly in the future!

$$\boxed{\begin{aligned} \max_{p} \;\; & \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \\ \text{s.t.} \;\; & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i \\ & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1 \end{aligned}}$$

Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

[Content adapted from CMU 10-708]

Latent (unobserved) variables

- A variable can be unobserved (latent) because:
  - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
    - e.g., speech recognition models, mixture models, …

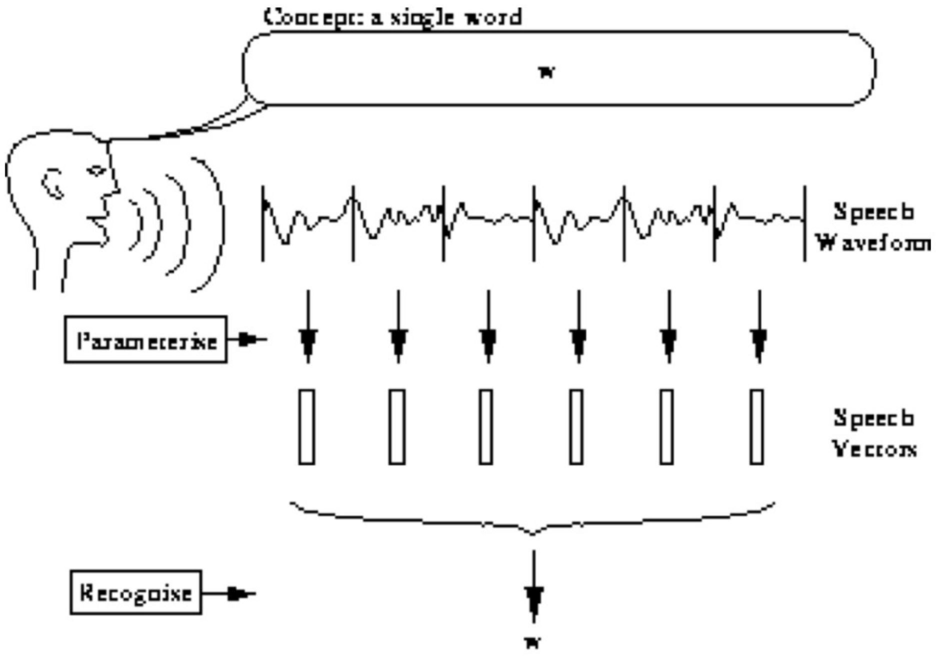
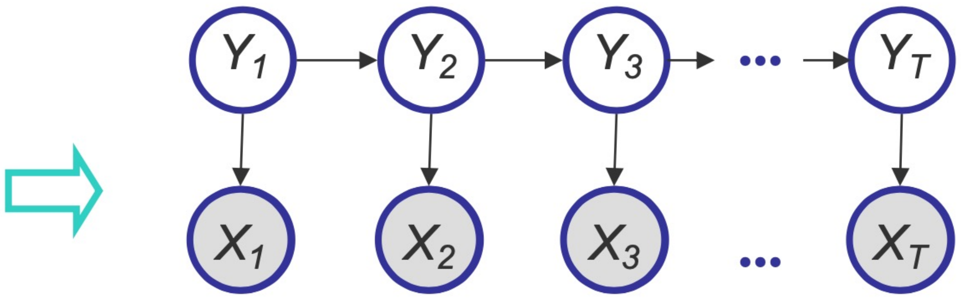
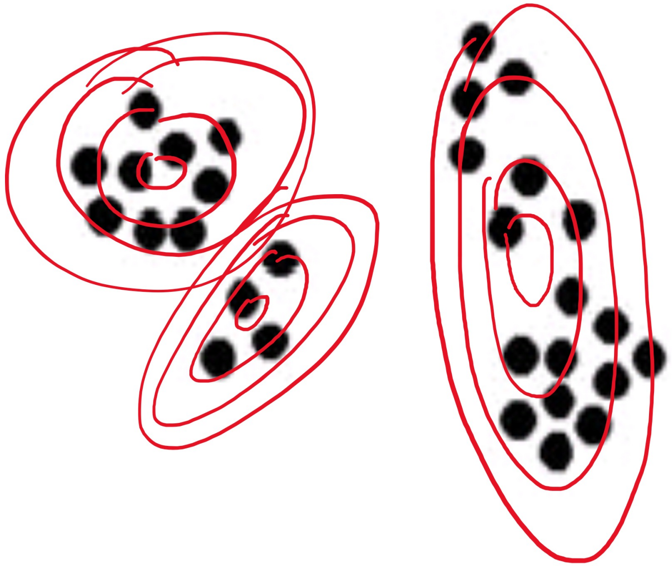


Fig. 1.2 Isolated Word Problem



Latent (unobserved) variables

- A variable can be unobserved (latent) because:
  - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
    - e.g., speech recognition models, mixture models, ...



Latent (unobserved) variables

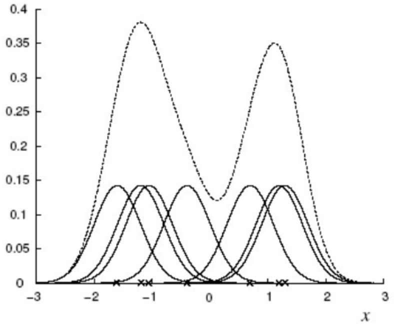
- A variable can be unobserved (latent) because:
    - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
        - e.g., speech recognition models, mixture models, …
    - a real-world object (and/or phenomena), but difficult or impossible to measure
        - e.g., the temperature of a star, causes of a disease, evolutionary ancestors …
    - a real-world object (and/or phenomena), but sometimes wasn't measured, because of faulty sensors, etc.
- Discrete latent variables can be used to partition/cluster data into sub-groups
- Continuous latent variables (factors) can be used for dimensionality reduction (e.g., factor analysis, etc.)

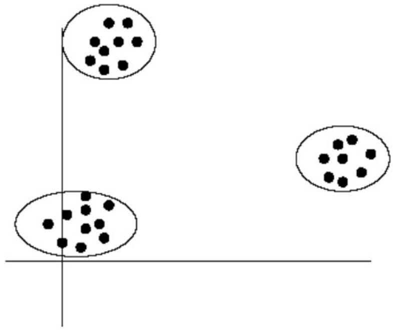
Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components:

$$p(x_n|\mu,\Sigma) = \sum_k \pi_k N(x,|\,\mu_k,\Sigma_k)$$

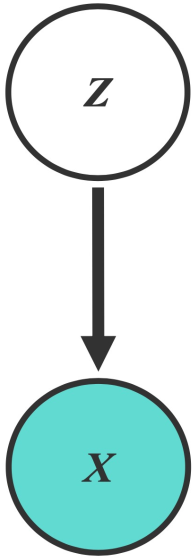
mixture proportion   mixture component





- This model can be used for unsupervised clustering.
  - This model (fit by AutoClass) has been used to discover new kinds of stars in astronomical data, etc.

Example: Gaussian Mixture Models (GMMs)



- Consider a mixture of K Gaussian components:
  - *Z* is a latent class indicator vector:

$$p(z_n) = \text{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

  - *X* is a conditional Gaussian variable with a class-specific mean/covariance

$$p(x_n \mid z_n^k = 1, \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{ -\tfrac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right\}$$

  - The likelihood of a sample:

Parameters to be learned:

mixture component

mixture proportion

$$\begin{aligned} p(x_n | \mu, \Sigma) &= \sum\nolimits_k p(z^k = 1 \mid \pi) p(x, \mid z^k = 1, \mu, \Sigma) \\ &= \sum\nolimits_{z_n} \prod\nolimits_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k} \right) = \sum\nolimits_k \pi_k N(x, \mid \mu_k, \Sigma_k) \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components: $p(x_n|\mu,\Sigma)=\sum_k \pi_k N(x,|\,\mu_k,\Sigma_k)$
- Recall MLE for completely observed data
  - Data log-likelihood:

$$\ell\,(\boldsymbol{\theta}; D) = \log\prod_n p(z_n, x_n) = \log\prod_n p(z_n\,|\,\pi)p(x_n\,|\,z_n,\mu,\sigma)$$

$$= \sum_n \log\prod_k \pi_k^{z_n^k} + \sum_n \log\prod_k N(x_n;\mu_k,\sigma)^{z_n^k}$$

$$= \sum_n\sum_k z_n^k \log\pi_k - \sum_n\sum_k z_n^k \tfrac{1}{2\sigma^2}(x_n - \mu_k)^2 + C$$

  - MLE:

$$\hat{\pi}_{k,MLE} = \arg\max_{\pi} \ell\,(\boldsymbol{\theta}; D),$$

$$\hat{\mu}_{k,MLE} = \arg\max_{\mu} \ell\,(\boldsymbol{\theta}; D) \qquad \Rightarrow \quad \hat{\mu}_{k,MLE} = \frac{\sum_n z_n^k x_n}{\sum_n z_n^k}$$

$$\hat{\sigma}_{k,MLE} = \arg\max_{\sigma} \ell\,(\boldsymbol{\theta}; D)$$

- What if we do not know z_n?

Why is Learning Harder?

- **Complete log likelihood**: if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

$$\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \log p(\boldsymbol{z}|\theta_z) + \log p(\boldsymbol{x}|\boldsymbol{z}, \theta_x)$$

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately
- But given that $\boldsymbol{z}$ is not observed, $\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})$ is a random quantity, cannot be maximized directly
- **Incomplete (or marginal) log likelihood:** with $\boldsymbol{z}$ unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

  - All parameters become coupled together
  - In other models when $\boldsymbol{z}$ is complex (continuous) variables (as we'll see later), marginalization over z is intractable.

Expectation Maximization (EM)

- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})] = \sum_z q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

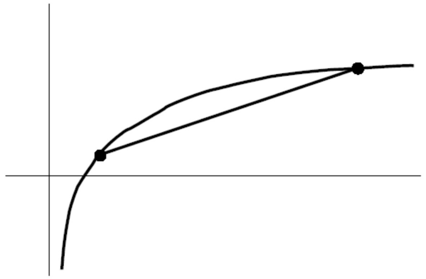
    - A deterministic function of θ
    - Inherit the factorizability of $\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})$
- Use this as the surrogate objective
- Does maximizing this surrogate yield a maximizer of the likelihood?

Expectation Maximization (EM)

- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] = \sum_z q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x},\boldsymbol{z}|\theta)$$

- Jensen's inequality



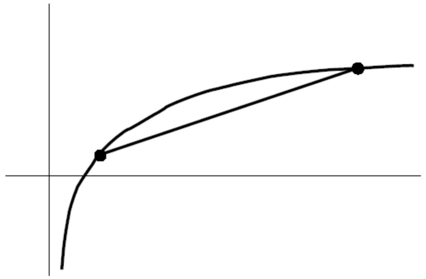
$$\begin{aligned}
\ell(\theta;x) &= \log p(x\mid\theta) \\
&= \log \sum_z p(x,z\mid\theta) \\
&= \log \sum_z q(z\mid x)\frac{p(x,z\mid\theta)}{q(z\mid x)} \\
&\geq \sum_z q(z\mid x)\log\frac{p(x,z\mid\theta)}{q(z\mid x)} \quad \text{Evidence Lower Bound (ELBO)} \\
&= \sum_z q(z\mid x)\log p(x,z\mid\theta) - \sum_z q(z\mid x)\log q(z\mid x) \\
&= \mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] + H(q)
\end{aligned}$$

Expectation Maximization (EM)

- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] = \sum_z q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x},\boldsymbol{z}|\theta)$$

- Jensen's inequality



$$\begin{aligned} \ell(\theta;x) &= \log p(x\mid\theta) \\ &= \log \sum_z p(x,z\mid\theta) \\ &= \log \sum_z q(z\mid x)\frac{p(x,z\mid\theta)}{q(z\mid x)} \\ &\geq \sum_z q(z\mid x)\log\frac{p(x,z\mid\theta)}{q(z\mid x)} \end{aligned}$$

- Indeed we have

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

Lower Bound and Free Energy

- For fixed data $\boldsymbol{x}$, define a functional called the (variational) free energy:

$$F(q,\theta) = -\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] - H(q) \geq \ell(\theta;\boldsymbol{x})$$

- The EM algorithm is coordinate-decent on F
    - At each step t:
        - E-step: $q^{t+1} = \arg\min_q F\left(q,\theta^t\right)$
        - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1},\theta^t\right)$

E-step: minimization of $F(q,\theta)$ w.r.t q

- Claim:

$$q^{t+1} = \mathrm{argmin}_q \, F(q,\theta^t) = p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$$

  - This is the posterior distribution over the latent variables given the data and the current parameters.

- Proof (easy): recall

$$\underbrace{\ell(\theta^t;\boldsymbol{x})}_{\text{Independent of } q} = \underbrace{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}|\theta^t)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]}_{-F(q,\theta^t)} + \underbrace{\mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta^t)\big)}_{\geq 0}$$

  - $F(q,\theta^t)$ is minimized when $\mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta^t)\big) = 0$, which is achieved only when $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$

M-step: minimization of $F(q,\theta)$ w.r.t $\boldsymbol{\theta}$

- Note that the free energy breaks into two terms:

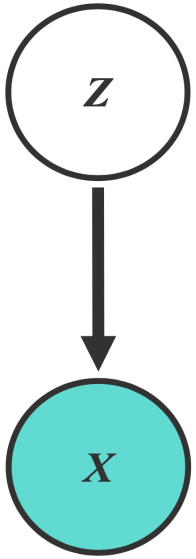
$$F(q,\theta) = -\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] - H(q) \geq \ell(\theta;\boldsymbol{x})$$

  - The first term is the expected complete log likelihood and the second term, which does not depend on q, is the entropy.
- Thus, in the M-step, maximizing with respect to θ for fixed q we only need to consider the first term:

$$\theta^{t+1} = \text{argmax}_\theta\, \mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] = \text{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x},\boldsymbol{z}|\theta)$$

  - Under optimal q^{t+1}, this is equivalent to solving a standard MLE of fully observed model $p(\boldsymbol{x},\boldsymbol{z}|\theta)$, with z replaced by its expectation w.r.t $p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$

Example: Gaussian Mixture Models (GMMs)



- Consider a mixture of K Gaussian components:
  - *Z* is a latent class indicator vector:

$$p(z_n) = \mathrm{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

  - *X* is a conditional Gaussian variable with a class-specific mean/covariance

$$p(x_n \mid z_n^k = 1, \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{- \tfrac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right\}$$

  - The likelihood of a sample:

mixture component

mixture proportion

$$\begin{aligned} p(x_n|\mu,\Sigma) &= \textstyle\sum_k p(z^k = 1 \mid \pi) p(x, \mid z^k = 1, \mu, \Sigma) \\ &= \textstyle\sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k}\right) = \sum_k \pi_k N(x, \mid \mu_k, \Sigma_k) \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components
- The expected complete log likelihood

$$\mathbb{E}_q\left[\ell_c(\boldsymbol{\theta}; x, z)\right] = \sum_n \mathbb{E}_q\left[\log p\left(z_n \mid \pi\right)\right] + \sum_n \mathbb{E}_q\left[\log p\left(x_n \mid z_n, \mu, \Sigma\right)\right]$$

$$= \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \log \pi_k - \frac{1}{2} \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \left(\left(x_n - \mu_k\right)^T \Sigma_k^{-1} \left(x_n - \mu_k\right) + \log\left|\Sigma_k\right| + C\right)$$

- E-step: computing the posterior of z_n given the current estimate of the parameters (i.e., π , μ, Σ)

$$p(z_n^k = 1 \mid x, \mu^{(t)}, \Sigma^{(t)}) = \frac{\pi_k^{(t)} N(x_n, \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_i \pi_i^{(t)} N(x_n, \mid \mu_i^{(t)}, \Sigma_i^{(t)})}$$

$$\nearrow \; p(z_n^k = 1, x, \mu^{(t)}, \Sigma^{(t)})$$

$$\searrow \; p(x, \mu^{(t)}, \Sigma^{(t)})$$

Example: Gaussian Mixture Models (GMMs)

- M-step: computing the parameters given the current estimate of z_n

$$\pi_k^* = \arg\max \langle l_c(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \frac{\partial}{\partial \pi_k} \langle l_c(\boldsymbol{\theta}) \rangle = 0, \forall k, \quad \text{s.t.} \sum_{\mathrm{k}} \pi_k = 1$$

$$\Rightarrow \quad \pi_k^* = \sum\nolimits_n \langle z_n^k \rangle_{q^{(t)}} \Big/ N = \sum\nolimits_n \tau_n^{k(t)} \Big/ N = \langle n_k \rangle \Big/ N$$

$$\mu_k^* = \arg\max \langle l(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \quad \mu_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} x_n}{\sum_n \tau_n^{k(t)}}$$

$$\Sigma_k^* = \arg\max \langle l(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \quad \Sigma_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \tau_n^{k(t)}}$$

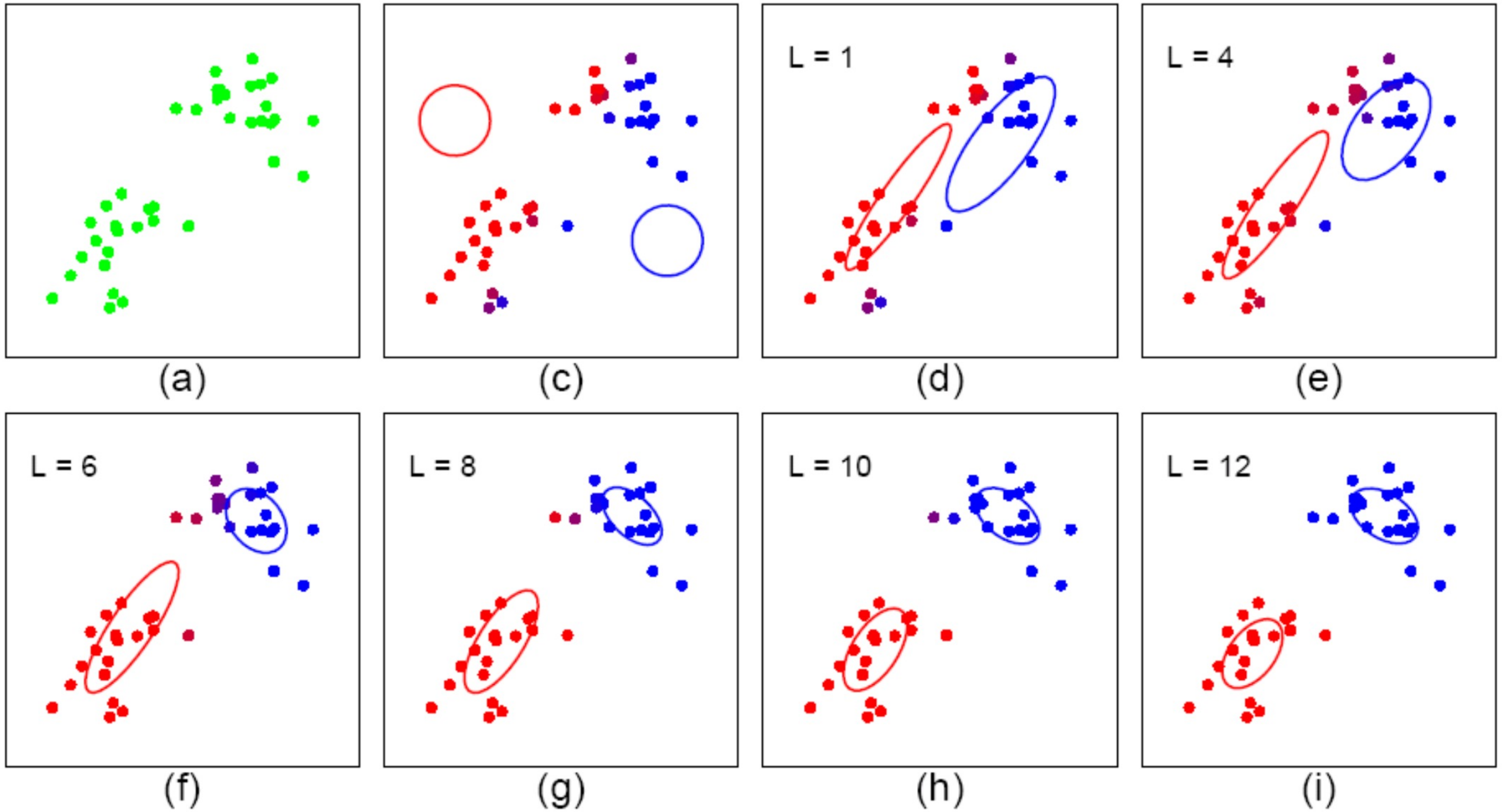
Fact:

$$\frac{\partial \log |\mathrm{A}^{-1}|}{\partial \mathrm{A}^{-1}} = \mathrm{A}^T$$

$$\frac{\partial \mathbf{x}^T \mathrm{A} \mathbf{x}}{\partial \mathrm{A}} = \mathbf{x}\mathbf{x}^T$$

Example: Gaussian Mixture Models (GMMs)

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



(a) (c) (d) (e) (f) (g) (h) (i)

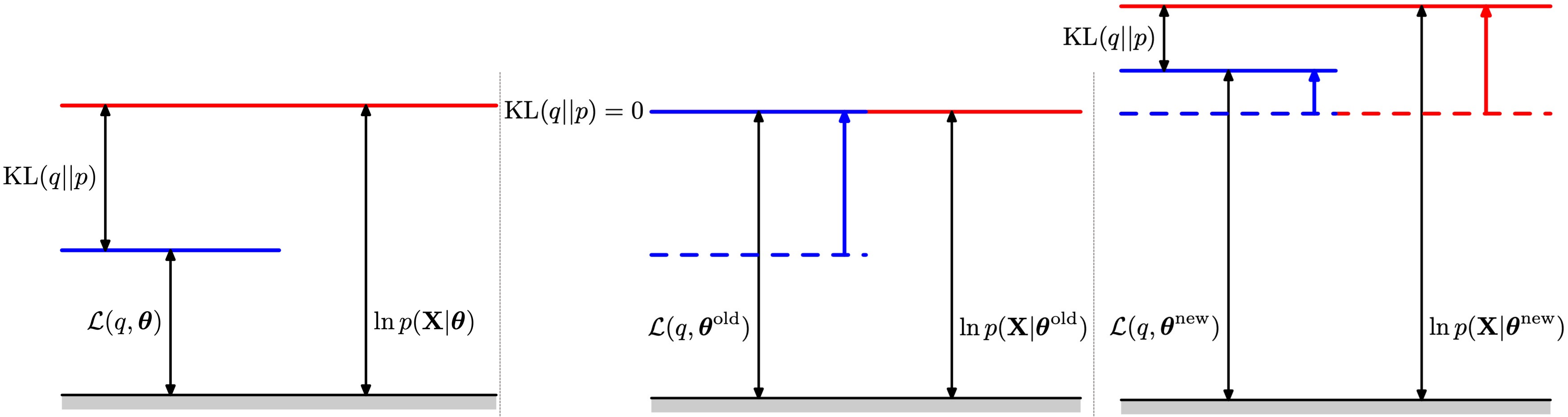
Summary: EM Algorithm

- A way of maximizing likelihood function for latent variable models. Finds MLE of parameters when the original (hard) problem can be broken up into two (easy) pieces
    - Estimate some "missing" or "unobserved" data from observed data and current parameters.
    - Using this "complete" data, find the maximum likelihood parameter estimates.

- Alternate between filling in the latent variables using the best guess (posterior) and updating the parameters based on this guess:
    - E-step: $q^{t+1} = \arg\min_{q} F\left(q, \theta^{t}\right)$
    - M-step: $\theta^{t+1} = \arg\min_{\theta} F\left(q^{t+1}, \theta^{t}\right)$

Each EM iteration guarantees to improve the likelihood

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



E-step

M-step

[PRML, Chap 9.4]

EM Variants

- Sparse EM
    - Do not re-compute exactly the posterior probability on each data point under all models, because it is almost zero.
    - Instead keep an "active list" which you update every once in a while.

- Generalized (Incomplete) EM:
    - It might be hard to find the ML parameters in the M-step, even given the completed data. We can still make progress by doing an M-step that improves the likelihood a bit (e.g. gradient step).

Summary

- Supervised Learning
  - Maximum likelihood estimation (MLE)
  - Duality between MLE and Maximum Entropy Principle

- Unsupervised learning
  - Maximum likelihood estimation (MLE) with latent variables
  - EM algorithm for MLE

Questions?