DSC291: Machine Learning with Few Labels

Machine Learning Basics
Supervised Learning / Unsupervised Learning

Zhiting Hu

Lecture 3, January 13, 2023

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- Join Piazza https://piazza.com/class/lcliazd0pcx2n5/

Recap: ML Basics

- Bayes' rule
  - prior, posterior

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$

- KL Divergence
  - relation to cross-entropy

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{x} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

- Functional derivatives
  - functional $F(y), y = y(x)$

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)}\eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

  - Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y}\eta(x)dx + O(\epsilon^2)$$

Supervised Learning:
Maximum likelihood vs Maximum Entropy

Supervised Maximum Likelihood

- 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

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

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

Recap: Exponential Family

- A distribution
$$p_\theta(x) = h(x) \exp\{\boldsymbol{\theta} \cdot T(x)\} / Z(\boldsymbol{\theta})$$
  is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter

  - $T(x) \in R^d$: sufficient statistics, features of data x

  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(x) \exp\{\boldsymbol{\theta} \cdot T(x)\}$: normalization factor

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

Maximum Likelihood for Exponential Family

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

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \log p(\boldsymbol{x} \mid \boldsymbol{\theta})$$

$$= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \left(\sum_{i} \theta_i T_i(\boldsymbol{x}) - \log Z(\boldsymbol{\theta}) \right)$$

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

- Take gradient and set to 0

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

$$= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})$$

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

> 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:

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

- 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 <u>variational</u> definition of a distribution!

$$\max_{p} \ \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x})$$

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

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

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:

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

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

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

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

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

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 $\Rightarrow$ 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:

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

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

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

- So feature constraints + MaxEnt $\Rightarrow$ exponential family.

- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

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

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

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

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

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

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

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!

$$\max_{p} \ \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x})$$

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

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

Unsupervised Learning

Unsupervised Learning

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

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
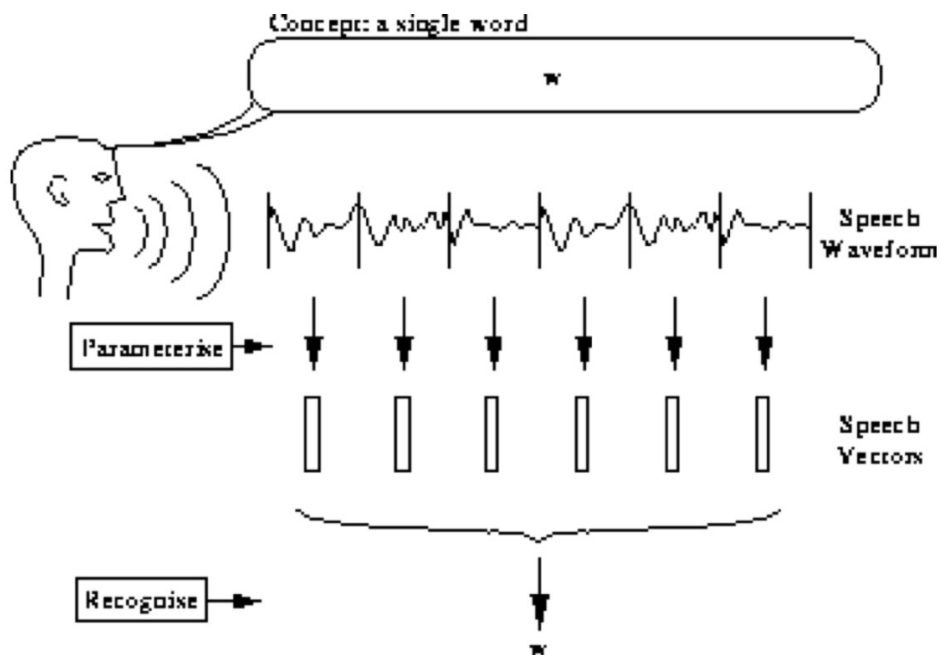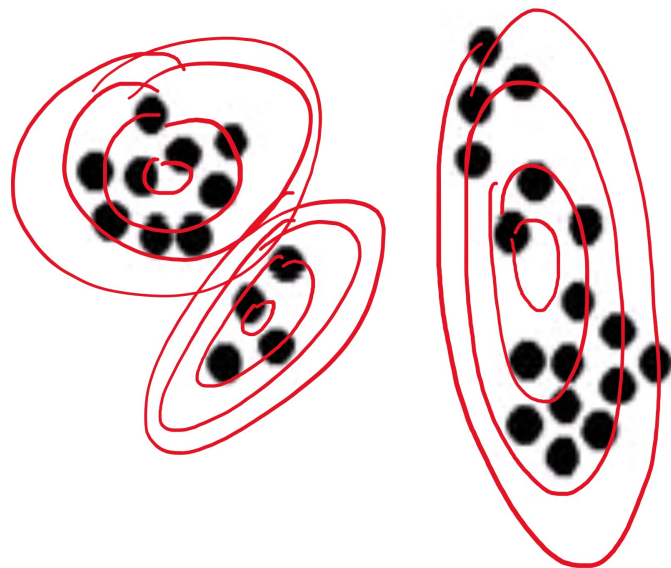   - e.g., speech recognition models, mixture models, ...
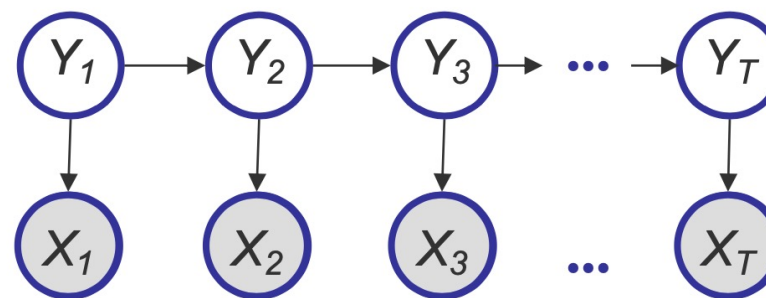
Fig. 1.2 Isolated Word Problem

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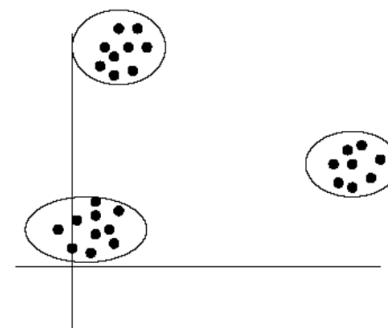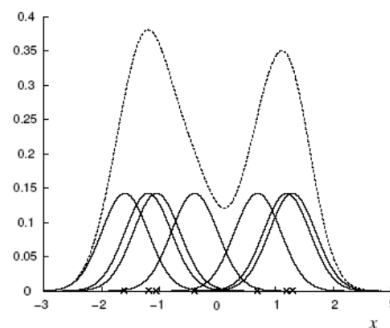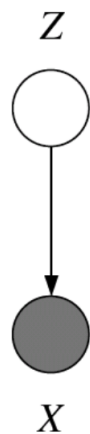
- This model can be used for unsupervised clustering.
  - This model (fit by AutoClass) has been used to discover new kinds of stars in astronomical data, etc.

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

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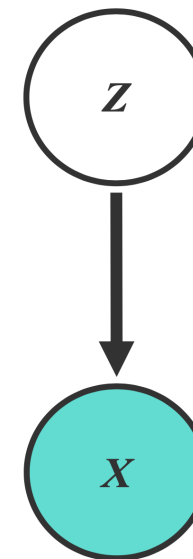
  ❑ The likelihood of a sample:

  Parameters to be learned:

  mixture component

  mixture proportion

  $$p(x_n \mid \mu, \Sigma) = \sum_k p(z^k = 1 \mid \pi) p(x_, \mid z^k = 1, \mu, \Sigma)$$

  $$= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k}\right) = \sum_k \pi_k N(x_, \mid \mu_k, \Sigma_k)$$

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:

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

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

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

  ○ MLE:

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

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

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

- What if we do not know z_n?

Why is Learning Harder?

- **Complete log likelihood:** if both x and z can be observed, then

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

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately

- But given that z is not observed, $\ell_c(\theta; x, z)$ is a random quantity, cannot be maximized directly

- **Incomplete (or marginal) log likelihood:** with z unobserved, our objective becomes the log of a marginal probability:

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

  - All parameters become coupled together
  - In other models when 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:

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

  - 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

$$\ell(\theta; x) = \log p(x \mid \theta)$$

$$= \log \sum_z p(x, z \mid \theta)$$

$$= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)}$$

$$\geq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \qquad \text{Evidence Lower Bound (ELBO)}$$

$$= \sum_z q(z \mid x) \log p(x, z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x)$$

$$= \mathbb{E}_q[\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})] + H(q)$$
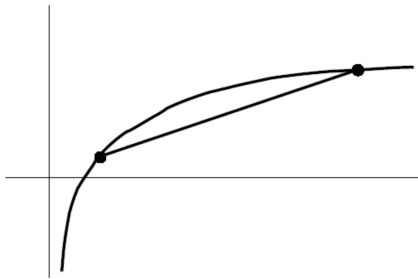
Expectation Maximization (EM)

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

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

- Jensen's inequality

$$\ell(\theta; x) = \log p(x | \theta)$$

$$= \log \sum_z p(x, z | \theta)$$

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

$$\geq \sum_z q(z | x) \log \frac{p(x, z | \theta)}{q(z | x)}$$

- Indeed we have

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

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: $\quad q^{t+1} = \arg\min_q F\left(q, \theta^t\right)$

    - M-step: $\quad \theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right)$

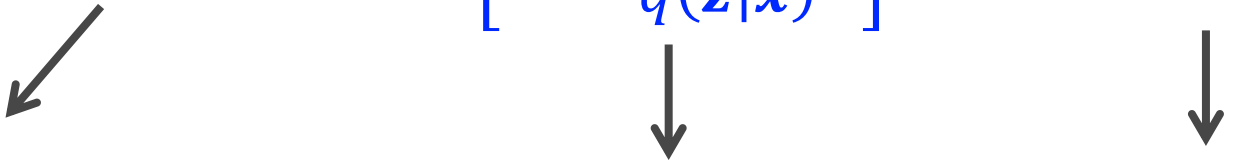# E-step: minimization of $F(q, \theta)$ w.r.t q

- Claim:
$$q^{t+1} = \operatorname{argmin}_q F(q, \theta^t) = p(\mathbf{z}|\mathbf{x}, \theta^t)$$

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

- Proof (easy): recall

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

Independent of q $-F(q, \theta^t)$ ≥ 0

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

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:

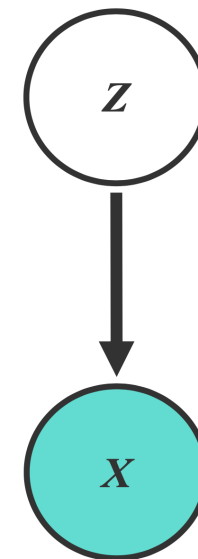$$p(z_n) = \text{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

  - 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:

$$p(x_n \mid \mu, \Sigma) = \sum_k p(z^k = 1 \mid \pi) p(x, \mid z^k = 1, \mu, \Sigma)$$

$$= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k}\right) = \sum_k \pi_k N(x, \mid \mu_k, \Sigma_k)$$

mixture proportion

mixture component

Example: Gaussian Mixture Models (GMMs)

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

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

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

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

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

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

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

Example: Gaussian Mixture Models (GMMs)

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

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

$$\Rightarrow \quad \pi_k^* = \left. \sum_n \langle z_n^k \rangle_{q^{(t)}} \middle/ N \right. = \left. \sum_n \tau_n^{k(t)} \middle/ N \right. = \left. \langle n_k \rangle \middle/ N \right.$$

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

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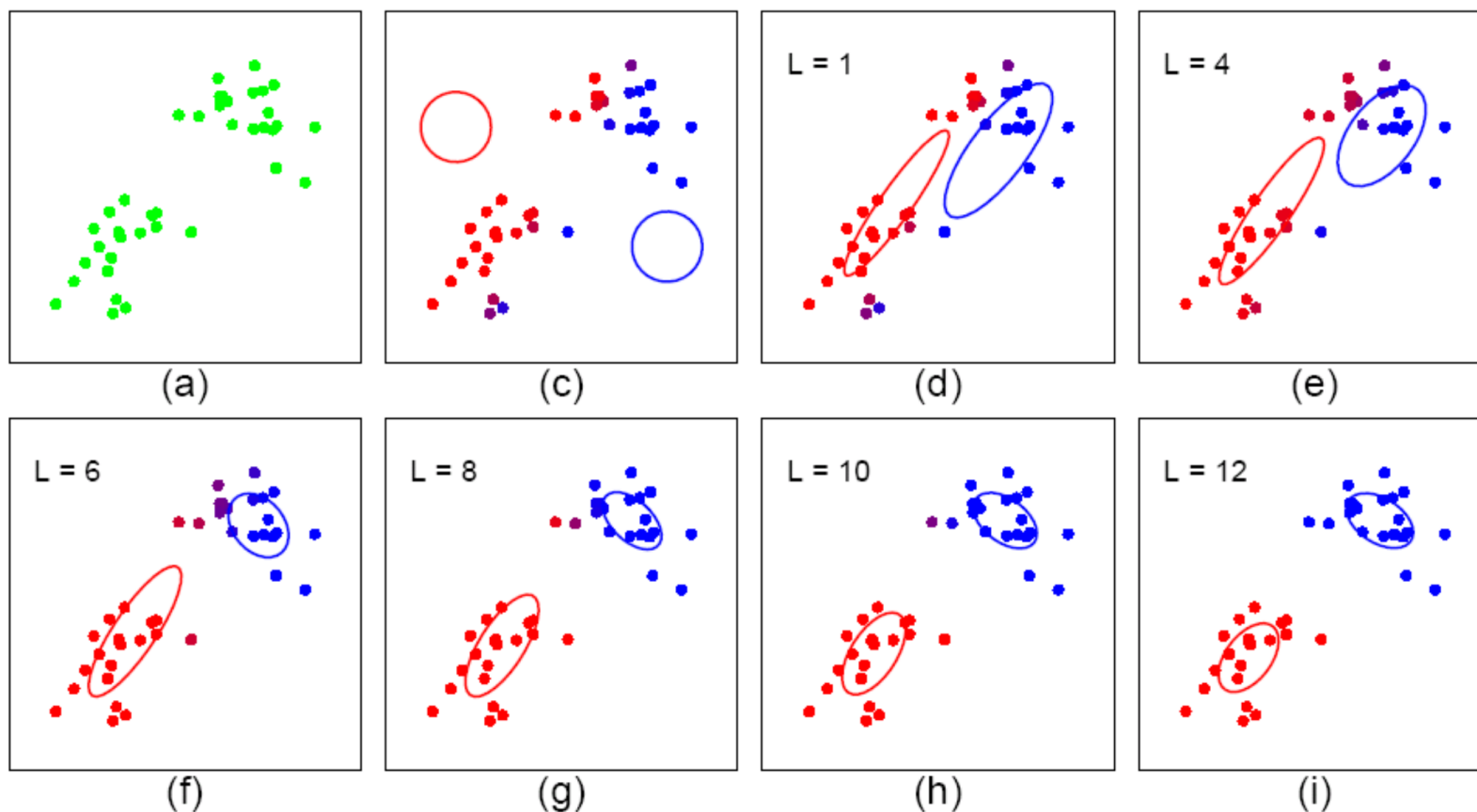
$$\text{Fact}:$$

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

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

Example: Gaussian Mixture Models (GMMs)

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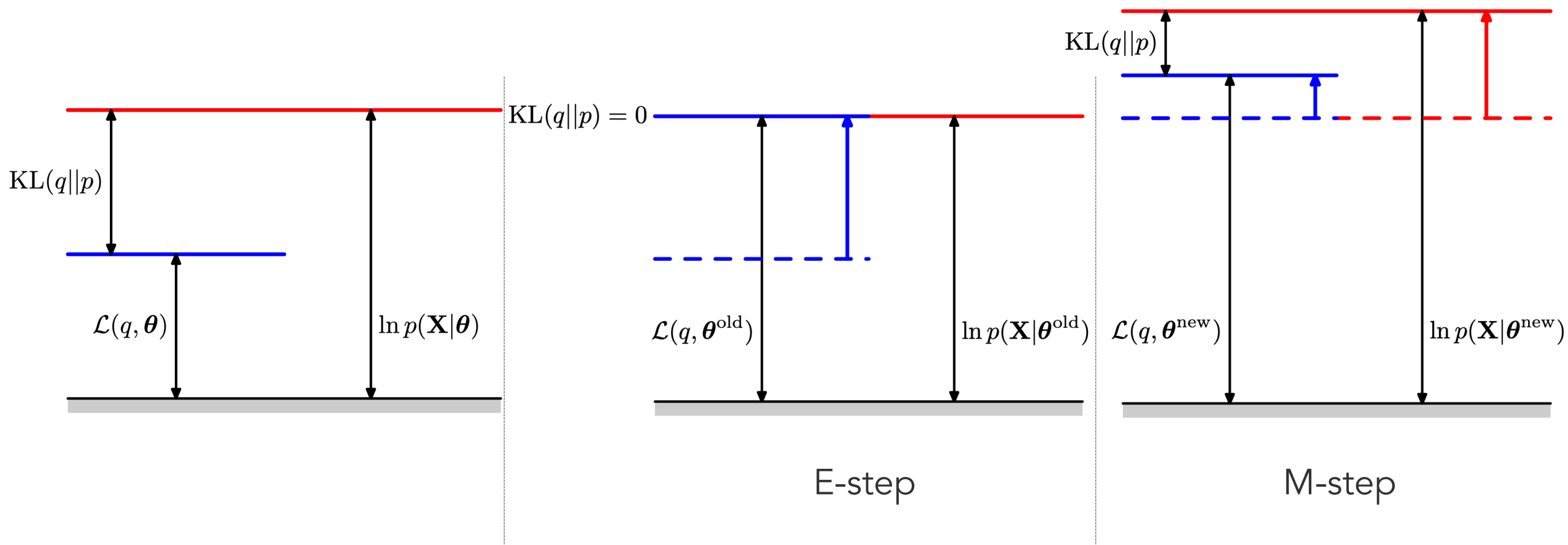- Loop:

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

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?