Lecture 19: Generative models, maximum likelihood, EM

Introduction to Learning and Analysis of Big Data
Generative models

- Can learning/knowing the distribution help?
  - Yes! Recall Naive Bayes
  - Assume conditional independence

\[
\mathbb{P}[X = x | Y = y] = \prod_{i=1}^{n} P[X_i = x_i | Y = y]
\]

- **symmetric case**: \( p_i = \mathbb{P}[X_i = Y] \) is probability of \( i \)th expert being correct, doesn’t depend on “truth” (true label)
- Bayes-optimal classifier
  - in general,
    \[
    h^*(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(X, Y) \sim \mathcal{D} [Y = y | X = x]
    \]
  - for symmetric Naive Bayes, \( h^*(x) = \text{sign} \left( \sum_{i=0}^{n} \log \frac{p_i}{1-p_i} x_i \right) \)
  - \( 2^n \) vs. \( n + 1 \) parameters — an exponential gap
Generative models

- Generative models: assume the distribution has some known form
- Some parameters of the distributions are unknown
- Goal is to find these parameters.
- Applications:
  - Better classification (e.g. Naive Bayes)
  - Inference: sometimes the parameters tell us something useful about the data generating process
  - Example: mixture of Gaussians — clustering (or soft clustering)
  - later: Hidden Markov Models — part of speech tagging
Generative models, example

- Distribution $\mathcal{D}_\theta$ on $\mathcal{X}$ determined by some parameter(s) $\theta \in \mathbb{R}$
- Example: $\mathcal{D}_\theta = \text{Bernoulli}(\theta)$
- Sample $S \sim \mathcal{D}_\theta^m$ generated by $m$ independent draws from $\mathcal{D}_\theta$
- Goal: recover $\theta$ (parametric density estimation)
- Example: $\mathcal{D}_\theta = \text{Bernoulli}(\theta)$ and $S = \{x_1, \ldots, x_m\}$
- Consider the estimator
  \[
  \hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} x_i
  \]
  - unbiased: $\mathbb{E}_{\mathcal{D}_\theta}[\hat{\theta}] = \theta$
  - Recall Hoeffding: w. prob $\geq 1 - \delta$,
    \[
    |\hat{\theta} - \theta| \leq \sqrt{\frac{\log(2/\delta)}{2m}}
    \]
- Next: Derive this estimator from the Maximum likelihood principle
The Maximum likelihood estimator

- A family of parameters $\Theta \subseteq \mathbb{R}^k$.
- A family of distributions $\{D_\theta \mid \theta \in \Theta\}$.
- Examples:
  - Bernoulli: $\Theta = [0, 1]$, $D_\theta = \text{Bernoulli}(\theta)$
  - Gaussian: $\Theta = \mathbb{R} \times \mathbb{R}_+$, $D_\theta = \mathcal{N}(\theta(1), \theta(2))$.
- Observe a sample $S = (x_1, \ldots, x_m)$ drawn from the distribution $D_\theta^m$.
- Want to estimate $\theta^*$.
- Consider the probability, for $\theta \in \Theta$, that the distribution $D_\theta$ generated the observed sample:

$$\mathbb{P}_{S' \sim D_\theta^m}[S' = S] = \prod_{i=1}^{m} \mathbb{P}_{X \sim D_\theta}[X = x_i].$$

- For convenience, use the log of that: the **Log Likelihood**

$$L(S; \theta) = \log \mathbb{P}_{S' \sim D_\theta^m}[S' = S] = \sum_{i=1}^{m} \log(\mathbb{P}_{X \sim D_\theta}[X = x_i]).$$
The Maximum likelihood estimator

- A family of distributions \( \{ \mathcal{D}_\theta \mid \theta \in \Theta \} \)
- Observe a sample \( S = (x_1, \ldots, x_m) \) drawn from the distribution \( \mathcal{D}_\theta^m \).
- The Log Likelihood of \( \theta \) after observing \( S \):
  \[
  L(S; \theta) := \log P_{S' \sim \mathcal{D}_\theta^m}[S' = S] = \sum_{i=1}^m \log (P_{X \sim \mathcal{D}_\theta}[X = x_i]).
  \]
- The Maximum likelihood estimator: Select the \( \theta \) that maximizes the log likelihood.
  \[
  \hat{\theta} := \arg\max_{\theta} L(S; \theta)
  \]
- Intuition: select the parameter \( \hat{\theta} \) that would make the observed \( S \) the least surprising.
- We hope that with a high probability \( \hat{\theta} \approx \theta^* \).
- \( \hat{\theta} \) is unbiased if for all \( \theta^* \), \( \mathbb{E}_{S \sim \mathcal{D}_{\theta^*}^m}[\hat{\theta}] = \theta^* \).
Maximum likelihood estimator for Bernoulli distributions

- $\Theta = [0, 1], \mathcal{D}_\theta = \text{Bernoulli}(\theta)$
- Observe $S = (x_1, \ldots, x_m)$.
- $\mathbb{P}_{S' \sim \mathcal{D}_\theta^m}[S' = S] = \prod_{i=1}^m \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_i x_i} (1 - \theta)^{\sum_i (1-x_i)}$
- The log-likelihood:
  $L(S; \theta) = \log \mathbb{P}_{S' \sim \mathcal{D}_\theta^m}[S' = S] = \log(\theta)^{\sum_i x_i} + \log(1 - \theta)^{\sum_i (1-x_i)}$.
- The Maximum likelihood estimator:
  $\hat{\theta} := \arg\max_\theta L(S; \theta)$

Differentiate and set to zero:
\[
\frac{d}{d\theta} L(S; \theta) = \frac{\sum_i x_i}{\theta} - \frac{\sum_i (1-x_i)}{1 - \theta}
\]

Solving $\frac{d}{d\theta} L(S; \theta) = 0$ yields $\hat{\theta} = \frac{1}{m} \sum_{i=1}^m x_i$. 
Maximum Likelihood for continuous densities

- Consider continuous distributions $\mathcal{D}_\theta$. E.g., Gaussian.

- **Density**: A function $f : \mathbb{R}^d \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^d} f(x) \, dx = 1$.

- A density $f$ defines a distribution $\mathcal{D}$:

$$
\mathbb{P}_{X \sim \mathcal{D}}[X \in A] = \int_A f(x) \, dx.
$$

- Parametric family of densities: $\{f_\theta : \theta \in \Theta\}$, $f_\theta$ determines $\mathcal{D}_\theta$.

- In this case $\mathbb{P}_{S' \sim \mathcal{D}_\theta^m}[S' = S] = 0$ always!

- How to do Maximum Likelihood?

- Redefine Log-Likelihood using densities: for $S = (x_1, \ldots, x_m)$,

$$
L(S; \theta) := \sum_{i=1}^{m} \log f_\theta(x_i).
$$

•
Example: Maximum Likelihood for Gaussian distributions

- \( \theta := (\mu, \sigma), \mathcal{D}_\theta = \mathcal{N}(\mu, \sigma^2) \).
- Parametric density:
  \[
  f_\theta(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)
  \]
- Log-likelihood: for \( S \sim \mathcal{N}(\mu, \sigma^2)^m \)
  \[
  L(S; \theta) = \sum_{i=1}^{m} \log f_\theta(x_i) = -m \log(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu)^2
  \]

  To find \( \hat{\theta} \), differentiate by \( \sigma \) and \( \mu \) and set \( \nabla L = 0 \).
  \[
  \frac{\partial}{\partial \mu} L(S; \theta) = \frac{1}{\sigma^2} \sum_{i} (x_i - \mu) = 0
  \]
  \[
  \frac{\partial}{\partial \sigma} L(S; \theta) = \frac{1}{\sigma^3} \sum_{i} (x_i - \mu)^2 - \frac{m}{\sigma} = 0
  \]

- \( \hat{\mu} = \frac{1}{m} \sum_{i} x_i \) (unbiased), \( \hat{\sigma} = \sqrt{\frac{1}{m} \sum_{i} (x_i - \hat{\mu})^2} \) (biased).
Consistency of the Maximum Likelihood Estimator

- How does the Maximum Likelihood Estimator (MLE) behave as we increase the sample size?
- Denote $S_m \sim D_{\theta^*}^m$
- Log-likelihood (for densities): $L(S_m; \theta) = \sum_{i=1}^{m} \log f_\theta(x_i)$
- $\hat{\theta}_m = \arg\max_{\theta} L(S_m; \theta)$
- Note: $\hat{\theta}_m$ is a random variable!
- Want: $\hat{\theta}_m \to \theta^*$ when $m \to \infty$.
- If this holds, we say that $\hat{\theta}$ is a consistent estimator.
- Under some regularity conditions, $\hat{\theta}_m \to N(\theta^*, \sigma_{\text{MLE}}^2 / m)$ for $m \to \infty$.
- $\sigma_{\text{MLE}}$ measures the sensitivity of $f_\theta(x)$ to small changes in $\theta$.
- This property is called the asymptotic normality of the MLE
- Next goal: compute $\hat{\theta}$ efficiently.
Computing the MLE

- We saw examples where $\hat{\theta}$ had a closed form solution.
- This is not always the case for other families of distributions.
- In general, $\hat{\theta}_m = \arg\max_\theta L(S_m; \theta)$ requires solving a maximization problem.
- Can we do Gradient descent (ascent)?
- $L(S; \cdot)$ might have many local maxima.
- Computing MLE for general parametric distributions is NP-hard.
- Even finding a local maximum is non-trivial in many cases.
- Can do gradient descent (ascent) to a local maximum.
- Better idea: Expectation-Maximization (EM)
- We will see EM through the problem of mixture distributions.
Mixture distributions

- Define $X \sim \mathcal{D}(p, \mu_0, \mu_1)$ as a mixture of two Gaussians:
  $X \sim (1 - p)N(\mu_0, 1) + pN(\mu_1, 1)$.

- To generate $X \sim \mathcal{D}(p, \mu_0, \mu_1)$, flip a $p$-biased coin to decide which of 2 Gaussians to draw $X$ from.

- $X$ has the following density:

  $$f_{p, \mu_0, \mu_1}(x) = (1 - p) \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2}} + p \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2}}$$

- Note: mixture of Gaussians is different from sum of Gaussians! Latter is Gaussian and former is not. [board]

- Log-likelihood of sample $S \sim \mathcal{D}(p, \mu_0, \mu_1)^m$, where $\theta := (p, \mu_0, \mu_1)$:

  $$L(S; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \left( \frac{1 - p}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu_0)^2}{2}} + \frac{p}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu_1)^2}{2}} \right).$$
Mixture distributions and Maximum Likelihood

- Log-likelihood of sample $S \sim \mathcal{D}(p, \mu_0, \mu_1)^m$, where $\theta := (p, \mu_0, \mu_1)$:

$$L(S; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \left( \frac{1-p}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu_0)^2}{2}} + \frac{p}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu_1)^2}{2}} \right).$$

- Problem: log of sum does not decompose nicely!
- “Log of sum is hard; sum of logs is easy”
- Characteristic feature of mixture distributions: do not decompose.
- $\nabla L = 0$ is a transcendental equation, no analytic solution
- Idea: do alternate maximization using **Latent variables**
Latent variables

- Mixture of two Gaussians $D(p, \mu_0, \mu_1) = (1 - p)N(\mu_0, 1) + pN(\mu_1, 1)$
- To generate $X \sim D(p, \mu_0, \mu_1)$:
  - draw $Z \sim \text{Bernoulli}(p)$, then draw $X \sim N(\mu_Z, 1)$.
- Density of $D(p, \mu_0, \mu_1)$ is an expectation over $Z$:
  \[
  f_{p,\mu_0,\mu_1}(x) = (1 - p) \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_0)^2/2} + p \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_1)^2/2}
  \]
  \[
  = \mathbb{E}_{Z \sim \text{Bernoulli}(p)} \left[ \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_Z)^2/2} \right]
  \]
- Log-likelihood of sample $S \sim D(p, \mu_0, \mu_1)^m$:
  \[
  L(S; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \mathbb{E}_{Z_i \sim \text{Bernoulli}(p)} \left[ \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu_{Z_i})^2/2} \right]
  \]
- The $Z = (Z_1, \ldots, Z_m)$ are latent variables.
- Idea: Given $Z$, maximizing for $\theta$ is easy; given $\theta$, expectation over $Z$ is easy.
Latent variables and augmented log-likelihood

- Log-likelihood of sample \( S \sim \mathcal{D}(p, \mu_0, \mu_1)^m \):

\[
L(S; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \mathbb{E}_{Z_i \sim \text{Ber}(p)} \left[ \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu Z_i)^2}{2}} \right]
\]

- Augment \( S \) with the latent variables \( Z = (Z_1, \ldots, Z_m) \)
  - Imagine that we observed \( S = ((x_1, z_1), \ldots, (x_m, z_m)) \)
  - Then we would know for each \( x_i \) which Gaussian it “came from”.

- Joint density of \( X, Z \):

\[
g_{p, \mu_0, \mu_1}(x, z) = p^z (1 - p)^{1-z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x - \mu z)^2}{2}}
\]

- Augmented likelihood:

\[
L(S, Z; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \left( p^{z_i} (1 - p)^{1-z_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu z_i)^2}{2}} \right)
\]
Augmented likelihood

- Joint density of observed variable \( X \) and latent variable \( Z \) (the Gaussian that \( X \) “comes from”):

\[
g_{p,\mu_0,\mu_1}(x, z) = p^z (1 - p)^{1-z} \frac{1}{\sqrt{2\pi}} e^{-(x - \mu_z)^2 / 2}
\]

- Simplify augmented likelihood:

\[
L(S, Z; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \left( p^{z_i} (1 - p)^{1-z_i} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_z)^2 / 2} \right)
= \log(p) \sum_i z_i + \log(1 - p) \sum_i (1 - z_i) - \frac{1}{2} \sum_{i=1}^{m} (x_i - \mu_{z_i})^2 - m \log(\sqrt{2\pi})
\]

- Augmented likelihood is “sum of logs” \( \implies \) easy to maximize 😊

\[
\hat{p} = \frac{1}{m} \sum_{i=1}^{m} z_i; \quad \hat{\mu}_1 = \frac{\sum_{i=1}^{m} z_i x_i}{\sum_{i=1}^{m} z_i}, \quad \hat{\mu}_0 = \frac{\sum_{i=1}^{m} (1 - z_i) x_i}{\sum_{i=1}^{m} (1 - z_i)}
\]

- But we need to know the \( z_i \)'s to do that!
The Expectation-Maximizaiton (EM) trick

- Augmented likelihood is easy to maximize.

\[
L(S, Z; p, \mu_0, \mu_1) = \log(p) \sum z_i + \log(1-p) \sum (1 - z_i) - \frac{1}{2} \sum (x_i - \mu z_i)^2 - m \log(\sqrt{2\pi})
\]

- Problem: don’t know the \(z_i\) — they’re latent!

- Solution: estimate the \(z_i\) by the prob. that \(x_i\) is from Gaussian “1”.

- These probabilities are called pseudo-counts.

- The probability that \(X\) is from Gaussian 1:

\[
P[Z = 1|X = x] = \frac{P[Z = 1, X = x]}{P[X = x]}
\]

\[
= \frac{p \frac{1}{\sqrt{2\pi}} e^{-(x - \mu_1)^2/2}}{(1 - p) \frac{1}{\sqrt{2\pi}} e^{-(x - \mu_0)^2/2} + p \frac{1}{\sqrt{2\pi}} e^{-(x - \mu_1)^2/2}}.
\]

- Problem: computing \(P[Z_i = 1|X = x_i]\) requires knowing \((p, \mu_0, \mu_1)\)!

- Solution: alternate maximization; start with a random guess
Expectation-Maximization (EM) for mixture of 2 Gaussians

**input** $(x_1, \ldots, x_m) \in \mathbb{R}^m$

**output** $(p, \mu_0, \mu_1) \in [0, 1] \times \mathbb{R} \times \mathbb{R}$

1: $t \leftarrow 0$, select $(p^{(t)}, \mu_0^{(t)}, \mu_1^{(t)})$ randomly
2: **repeat**
3: E-step: pseudo-counts

$$q_i^{(t)} := \frac{p^{(t)} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_1^{(t)})^2/2}}{(1 - p^{(t)}) \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_0^{(t)})^2/2} + p^{(t)} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_1^{(t)})^2/2}}, \quad i \leq m$$

4: M-step: maximize augmented likelihood

$$p^{(t+1)} := \frac{1}{m} \sum_{i=1}^{m} q_i^{(t)}; \quad \mu_1^{(t+1)} := \frac{\sum_{i=1}^{m} q_i^{(t)} x_i}{\sum_{i=1}^{m} q_i^{(t)}}; \quad \mu_0^{(t+1)} := \frac{\sum_{i=1}^{m} (1 - q_i^{(t)}) x_i}{\sum_{i=1}^{m} (1 - q_i^{(t)})}$$

5: $t := t + 1$
6: **until** convergence (no change in $L(S; p^{(t)}, \mu_0^{(t)}, \mu_1^{(t)})$)
EM algorithm in general

- In general: Family of densities $\mathcal{F} = \{f_{\theta}\}$ (not necessarily Gaussian.)
- $X$ is a mixture of distributions from $\mathcal{F}$.
- (Can have mixture of more than 2; here we show 2 for simplicity.)
- The density of $X$: $f_{p,\mu_0,\mu_1}(x) = (1 - p)f_{\mu_0}(x) + pf_{\mu_1}(x) = \mathbb{E}_{Z \sim \text{Ber}(p)}[f_{\mu_Z}(x)]$.
- Augmented likelihood given $z_i$:
  $$L(S, Z; p, \mu_0, \mu_1) = \sum_{i=1}^{m} \log \left[ p^{z_i} (1 - p)^{1 - z_i} f_{\mu_{z_i}}(x_i) \right].$$
- Define expected log-likelihood: given pseudo-counts $Q = (q_1, \ldots, q_m) \in [0, 1]^m$,
  $$F(S, Q; p, \mu_0, \mu_1) := \mathbb{E}_{Z_1 \sim \text{Ber}(q_1), \ldots, Z_m \sim \text{Ber}(q_m)}[L(S, Z; p, \mu_0, \mu_1)].$$
- **ASSUMPTION**: can efficiently compute
  $$(\hat{p}, \hat{\mu}_0, \hat{\mu}_1) = \arg\max_{p, \mu_0, \mu_1} F(S, Q; p, \mu_0, \mu_1)$$
- Note that
  $$F(S, Q; p, \mu_0, \mu_1) = \sum_{i=1}^{m} [(1 - q_i) \log((1 - p)f_{\mu_0}(x_i)) + q_i \log(pf_{\mu_1}(x_i))]$$
  $$= \sum_{i=1}^{m} (1 - q_i) \log(1 - p) + q_i \log(p) + (1 - q_i)L(S; 0, \mu_0, \mu_1) + q_i L(S; 1, \mu_0, \mu_1).$$
EM algorithm in general

- EM algorithm for a general mixture:
  - Initialize random \((p, \mu_0, \mu_1)\)
  - E-step:
    \[
    q_i = \mathbb{P}_{p, \mu_0, \mu_1}[Z_i = 1|X = x_i] = \frac{pf_{\mu_1}(x_i)}{(1 - p)f_{\mu_0}(x_i) + pf_{\mu_1}(x_i)}
    \]
  - M-step:
    \[
    (p', \mu'_0, \mu'_1) = \arg\max_{p, \mu_0, \mu_1} F(S, Q; p, \mu_0, \mu_1)
    \]
  - Repeat E-M steps until convergence.
- Does the EM algorithm converge?
- What does it converge to?
EM convergence analysis — preliminaries

- Discrete distribution $P = (p_1, \ldots, p_k)$, $p_i \geq 0$, $\sum_{i=1}^{k} p_i = 1$
- Entropy: measures amount of “randomness”/”uncertainty”:

$$H(P) = -\sum_{i=1}^{k} p_i \log(p_i) = \mathbb{E}_{i \sim P}[\log(1/p_i)].$$

- For a binary distribution $P = (1 - p, p)$,

$$H(p) = -p \log(p) - (1 - p) \log(1 - p).$$

- The entropy is smallest (= 0) for distributions concentrated on a single value
- The entropy is largest (= $\log k$) for uniform distributions.
- Recall convexity: $f : \mathbb{R} \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}$, $\lambda \in [0, 1]$

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

- Jensen’s inequality: If $X$ is a real random variable and $f : \mathbb{R} \to \mathbb{R}$ is convex, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

- For concave $f$ (i.e., $-f$ is convex, such as $\log(\cdot)$), the inequality is reversed.
EM as alternate maximization

- Mixture distribution with parameters $\theta = (p, \mu_0, \mu_1)$
- Observe $X \sim f_\theta(x) = (1 - p)f_0(x) + pf_1(x)$, latent $Z \sim \text{Ber}(p)$
- Augmented log-likelihood:

$$L(S, Z; \theta) = \sum_{i=1}^{m} \log P_\theta[X = x_i, Z = z_i] = \sum_{i=1}^{m} \log \left[ p^{z_i}(1 - p)^{1 - z_i} f_{z_i}(x_i) \right]$$

- Expected log-likelihood

$$F(S, Q; \theta) = \mathbb{E}_{Z \sim \text{Ber}(q_1) \times \ldots \times \text{Ber}(q_m)}[L(S, Z; \theta)]$$

- One step of EM (start with $Q, \theta$):
  - E-step: $q'_i \leftarrow P_\theta[Z = 1|X = x_i] = \frac{pf_1(x_i)}{(1 - p)f_0(x_i) + pf_1(x_i)}$; Set $Q' = \{q'_i\}$
  - M-step: $\theta' \leftarrow \text{argmax}_\theta F(S, Q'; \theta)$

- Define

$$G(Q, \theta) = F(S, Q; \theta) + \sum_{i=1}^{m} H(q_i)$$

- Claim: EM is performing alternate maximization on $G$.
  - E-step performs $Q' \leftarrow \text{argmax}_{Q \in [0,1]^m} G(Q, \theta)$ [to prove]
  - M-step performs $\theta' \leftarrow \text{argmax}_\theta G(Q', \theta)$ [obvious]
Proof of claim

Claim: E-step, \( q_i' = \mathbb{P}_\theta[Z_i = 1|X = x_i] \) for \( i \leq m \), is equivalent to

\( Q' \leftarrow \text{argmax}_Q G(Q, \theta) \), where (denote \( q_{i1} := q_i, q_{i0} = 1 - q_i \))

\[
G(Q, \theta) = F(S, Q; \theta) + \sum_{i=1}^m H(q_i) = \sum_{i=1}^m \sum_{j=0}^1 q_{ij} \log \mathbb{P}_\theta[Z = z_i, X = x_i] - \sum_{i=1}^m \sum_{j=0}^1 q_{ij} \log q_{ij}
\]

Proof:

We will show that for any \( Q \), \( G(Q, \theta) \leq L(S; \theta) \).

Then we will show that \( G(Q', \theta) = L(S; \theta) \).

\[
G(Q, \theta) = \sum_{i=1}^m \left( \sum_{j=0}^1 q_{ij} \log \frac{\mathbb{P}_\theta[X = x_i, Z = j]}{q_{ij}} \right)
\]

\([\text{Jensen}]\]

\[
\leq \sum_{i=1}^m \log \left( \sum_{j=0}^1 q_{ij} \frac{\mathbb{P}_\theta[X = x_i, Z = j]}{q_{ij}} \right)
\]

\[
= \sum_{i=1}^m \log \left( \sum_{j=0}^1 \mathbb{P}_\theta[X = x_i, Z = j] \right)
\]

\[
= \sum_{i=1}^m \log \mathbb{P}_\theta[X = x_i] = L(S; \theta).
\]

Continue proof next slide...
Proof of claim, cont

- Claim: E-step, \( q_{ij} = \mathbb{P}_\theta[Z_i = j|X = x_i] \) for \( i \leq m \), is equivalent to \( Q' = \text{argmax}_Q G(Q, \theta) \), where

\[
G(Q, \theta) = F(S, Q; \theta) + \sum_{i=1}^{m} H(q_i) = \sum_{i=1}^{m} \sum_{j=0}^{1} q_{ij} \log \mathbb{P}_\theta[Z = z_i, X = x_i] - \sum_{i=1}^{m} \sum_{j=0}^{1} q_{ij} \log q_{ij}.
\]

- Proof:
  
  ▶ Prev. slide: showed \( G(Q, \theta) \leq L(S; \theta) \)
  
  ▶ now put \( q'_{ij} = \mathbb{P}_\theta[Z = j|X = x_i] \).

\[
G(Q', \theta) = \sum_{i=1}^{m} \left( \sum_{j=0}^{1} \mathbb{P}_\theta[Z = j|X = x_i] \log \frac{\mathbb{P}_\theta[X = x_i, Z = j]}{\mathbb{P}_\theta[Z = j|X = x_i]} \right)
\]

\[
= \sum_{i=1}^{m} \sum_{j=0}^{1} \mathbb{P}_\theta[Z = j|X = x_i] \log \mathbb{P}_\theta[X = x_i]
\]

\[
= \sum_{i=1}^{m} \log \mathbb{P}_\theta[X = x_i] \sum_{j=0}^{1} \mathbb{P}_\theta[Z = j|X = x_i]
\]

\[
= \sum_{i=1}^{m} \log \mathbb{P}_\theta[X = x_i] = L(S; \theta).
\]

▶ conclude: \( G(Q', \theta) = L(S; \theta) \) for \( Q' \) chosen in E-step, so \( Q' \) maximizes \( G(Q, \theta) \).
EM convergence: conclusion

- \( G(Q, \theta) = \sum_{i=1}^{m} \sum_{j=0}^{1} [q_{ij} \log \mathbb{P}_\theta[Z = z_i, X = x_i] - q_{ij} \log q_{ij}] \)
- Just proved: EM is performing alternate maximization
  - E-step: \( Q^{(t+1)} \leftarrow \arg\max_{Q \in [0,1]^m} G(Q^{(t)}, \theta^{(t)}) \)
  - M-step: \( \theta^{(t+1)} \leftarrow \arg\max_{\theta} G(Q^{(t+1)}, \theta) \)
  - also, we saw that for every round \( t \), \( G(Q^{(t)}, \theta^{(t)}) = L(S; \theta^{(t)}) \).

**Theorem:** \( L(S; \theta^{(t+1)}) \geq L(S; \theta^{(t)}) \)

**Proof:**
\[
L(S; \theta^{(t)}) = G(Q^{(t)}, \theta^{(t)}) \leq G(Q^{(t+1)}, \theta^{(t)}) \leq G(Q^{(t+1)}, \theta^{(t+1)}) = L(S; \theta^{(t+1)})
\]

Conclude: the EM update never decreases the likelihood, converges to local maximum (note: can take infinite number of steps; also assumed likelihood function is **bounded**)

What about convergence of the **value of the parameter** \( \theta^{(t)} \)?

Pathological examples exist where \( \theta^{(t)} \) “cycles” (does not converge)

Sufficient for convergence of \( \theta^{(t)} \): unique maximizer at M-step.
EM vs gradient ascent

- EM convergence can be slow 😞
- Unlike gradient ascent, doesn’t require choosing step size! 😊
- Also, constraints on parameters (such as \( p \in [0, 1] \)) automatically enforced in EM 😊
- For mixture problems, EM almost universally preferred to gradient methods 😊
Choosing the number of mixture components $k$

- Generative model: $X \sim \sum_{i=1}^{k} p_i N(\mu_i, \sigma^2)$
- How to choose $k$?
- Maximum likelihood $L(S; \theta)$ is monotonically increasing in $k$ (why?)
- does that mean that $k = m$ is optimal?
- No! It will badly overfit.
- How to avoid overfitting?
- Cross-validation!
- Optimal $k^*$ will maximize likelihood on held-out data.
Summary

- When data $X \sim D_\theta$ is modeled by a **parametric family** of distributions
- **Maximum likelihood** technique used to learn parameters $\theta$
- Log-likelihood: $L(S; \theta) = \sum_{i=1}^{m} \log P_\theta[X = x_i]$
- Easy cases: can find $\arg\max_\theta L(S; \theta)$ analytically
- Hard cases: use EM
  - Start with random guess for $\theta$
  - E-step: posterior probabilities $Q$ of mixture latent variables
  - M-step: maximize expected (w.r.t. $Q$) augmented log-likelihood
  - Repeat until convergence
- EM guarantees monotonic convergence to local maximum (assuming bounded likelihood function).
- In general, maximizing likelihood is NP-hard.
- EM is preferred to gradient-based methods for learning density mixtures.