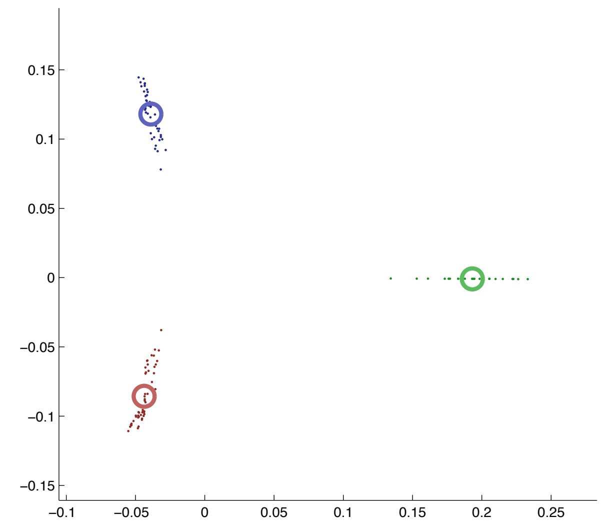
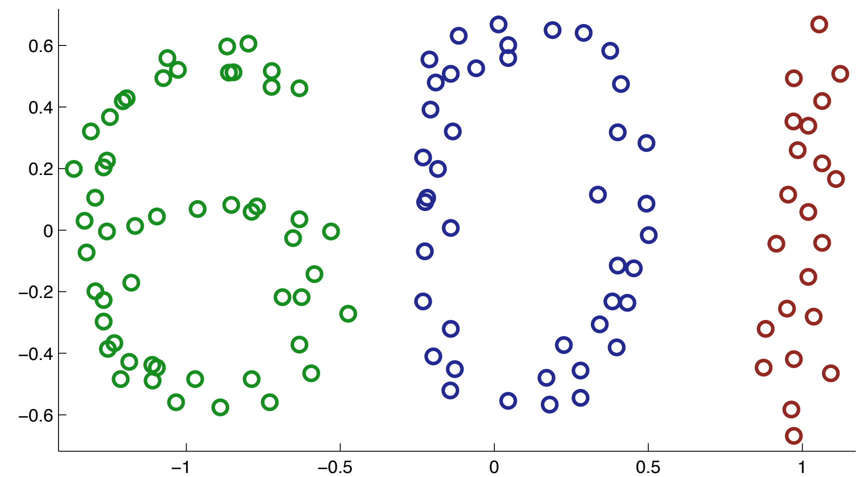
Review

- Supervised v. unsup. v. "other"
- Clustering (for understanding, for compression, or as input to another task)
  - break into "similar" groups
    - what is "similar"?
  - use of spectral embedding
    - mapping back to clusters in original space





Review

- k-means clustering
  - alternating optimization; convergence
  - initialization; multiple restarts; split / merge
- soft k-means
  - mixture of Gaussians model
  - E-step, M-step
  - connection to hard k-means
  - connection to naïve Bayes
  - (un)biasedness

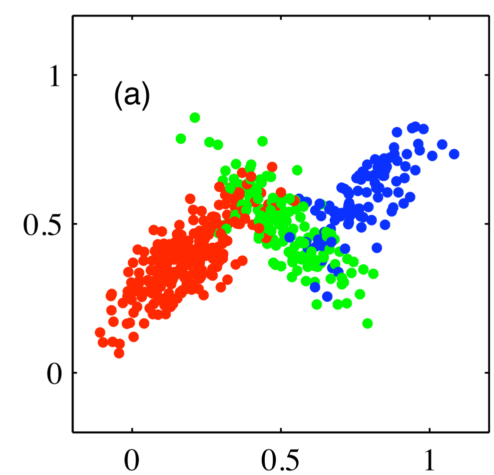


fig 9.5a from Bishop

Review

- EM algorithm
  - general strategy for MLE or MAP with hidden variables (in our case, Z_{ij})
  - we were in the middle of deriving soft k-means as an EM algorithm

Review: soft k-means

- Find soft assignments: "E step"
  - $q_{ij} = P(z_{ij} = 1 \mid x, \theta) \propto \underbrace{P(z_{ij} = 1 \mid \theta)}_{\text{to get normalizer, sum over } j} P(x_{ij} \mid z_{ij} = 1, \theta)$
  - $= p_j N(x_i \mid \mu_j, \Sigma_j)$ for max
- Update means: "M step"
  - $\mu_j = \sum_i q_{ij} x_i \Big/ \sum_i q_{ij}$
- Possibly: update covariances M-step
  - $\Sigma = \sum_i \sum_j q_{ij} (x_i - \mu_j)(x_i - \mu_j)^T \Big/ N$
- Possibly: update cluster weights M-step

$p_j = \sum_i q_{ij} \Big/ N$

Deriving soft k-means

- $P(X_i \mid Z_{ij} = 1, \theta) = \text{Gaussian}(\mu_j, \Sigma_j)$
- $P(Z_{ij} = 1 \mid \theta) = p_j$
- $P(X_i, Z_{i\cdot} \mid \theta) =$ $\prod_j p_j^{z_{ij}} N(x_i \mid \mu_j, \Sigma_j)^{z_{ij}}$
- $L = \ln P(X \mid \theta) =$

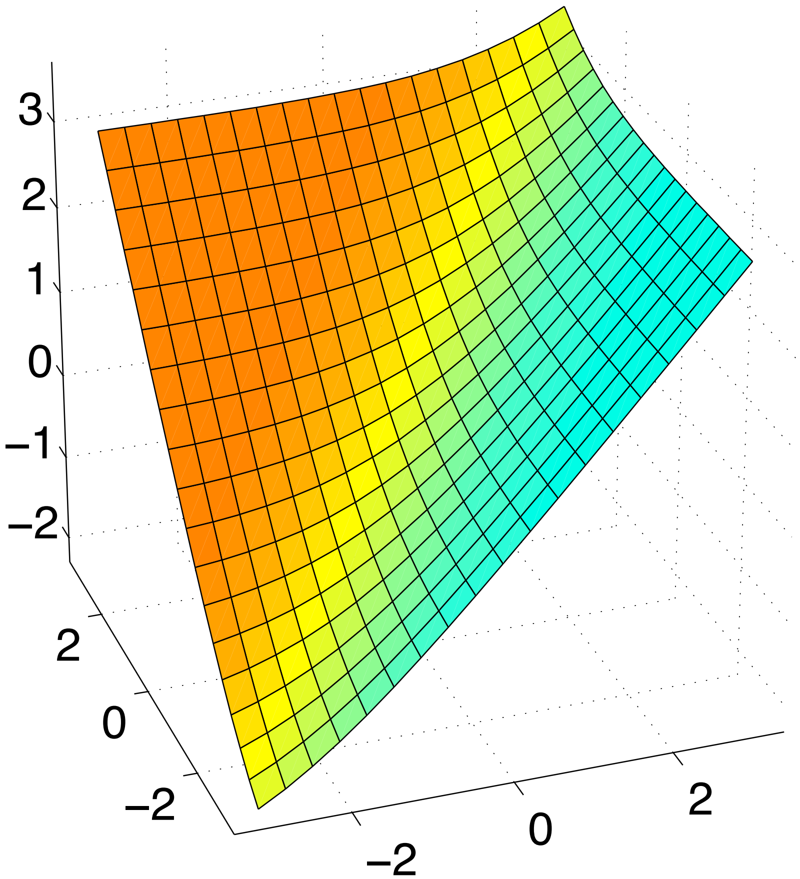
$$\ln \sum_z \exp\left[\sum_i \sum_j z_{ij} \left[\ln p_j + \ln N(x_i \mid \mu_j, \Sigma_j) \right] \right]$$

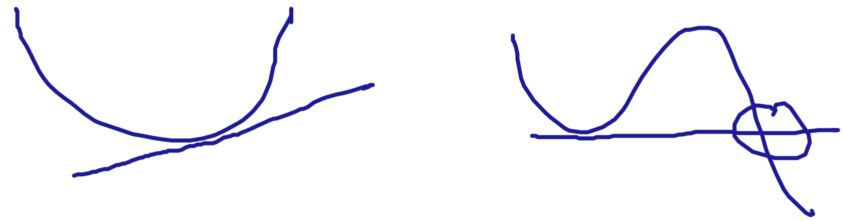
soft max

$\approx \max(x,y)$

$$f(x,y) = \ln(e^x+e^y)$$

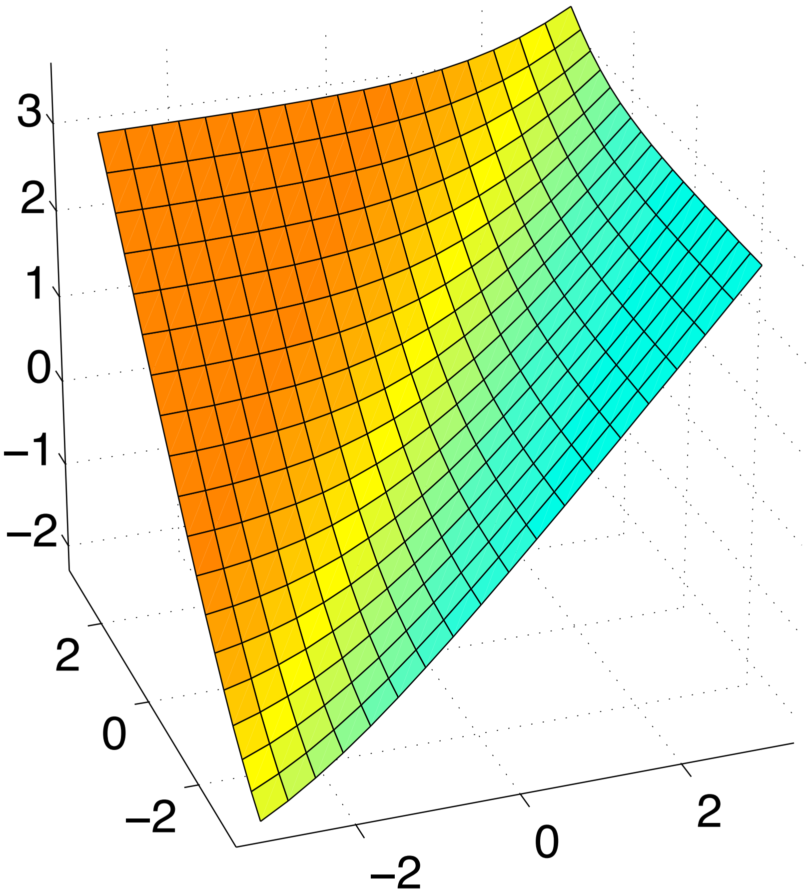
convex





Convex fns are lower-bounded by tangents

$f(x,y) = \ln(e^x+e^y)$



$$f(x,y) \geq f(x_0,y_0) + (x-x_0) f'_x(x_0,y_0) + (y-y_0) f'_y(x_0,y_0)$$

$$f'_x(x_0,y_0) = \frac{1}{e^{x_0}+e^{y_0}} e^{x_0} \equiv p$$

$$f'_y(x_0,y_0) = \frac{1}{e^{x_0}+e^{y_0}} e^{y_0} \equiv q$$

$$p+q = \frac{e^{x_0}+e^{y_0}}{e^{x_0}+e^{y_0}} = 1$$

$$\ln(e^x+e^y) \geq f(x_0,y_0)(p+q) + (x-x_0)p + (y-y_0)q$$

$$= p(x-x_0+f(x_0,y_0)) + q(y-y_0+f(x_0,y_0))$$

$$= p(x-\ln p) + q(y-\ln q)$$

$$\ln p = x_0 - f(x_0,y_0)$$

$$\ln q = y_0 - f(x_0,y_0)$$

In general

- $f(x_1, x_2, \ldots) = \ln(\sum_i \exp(x_i)) \geq$

  $\sum_i q_i (x_i - \ln q_i)$

  for any probability distribution q:

  $q_i \geq 0 \qquad \sum_i q_i = 1$

Optimizing a bound

- $L(\theta) = \ln \sum_z \exp(\ln P(X, Z{=}z \mid \theta)) \geq$

$$L(\theta, q) = \sum_z q_z (\ln P(X, Z=z \mid \theta) - \ln q_z)$$

  - for any distribution $q = \langle q_z \rangle$ $\quad q_z \geq 0 \quad \sum_z q_z = 1$
- Maximizing $L(\theta)$ is hard
- So, maximize $L(\theta, q)$ instead
  - start w/ arbitrary q, max wrt θ
  - then max wrt q to get a tighter bound
  - repeat

The distribution q

- One element q_z for each setting of $z = (z_{11}, z_{12}, \ldots z_{1K}, z_{21}, z_{22} \ldots z_{2K} \ldots\ldots z_{NK})$
- How many?
  - K^N
- Can we work with q efficiently?

Surprisingly, yes.

Example

$k = 2$

opt.

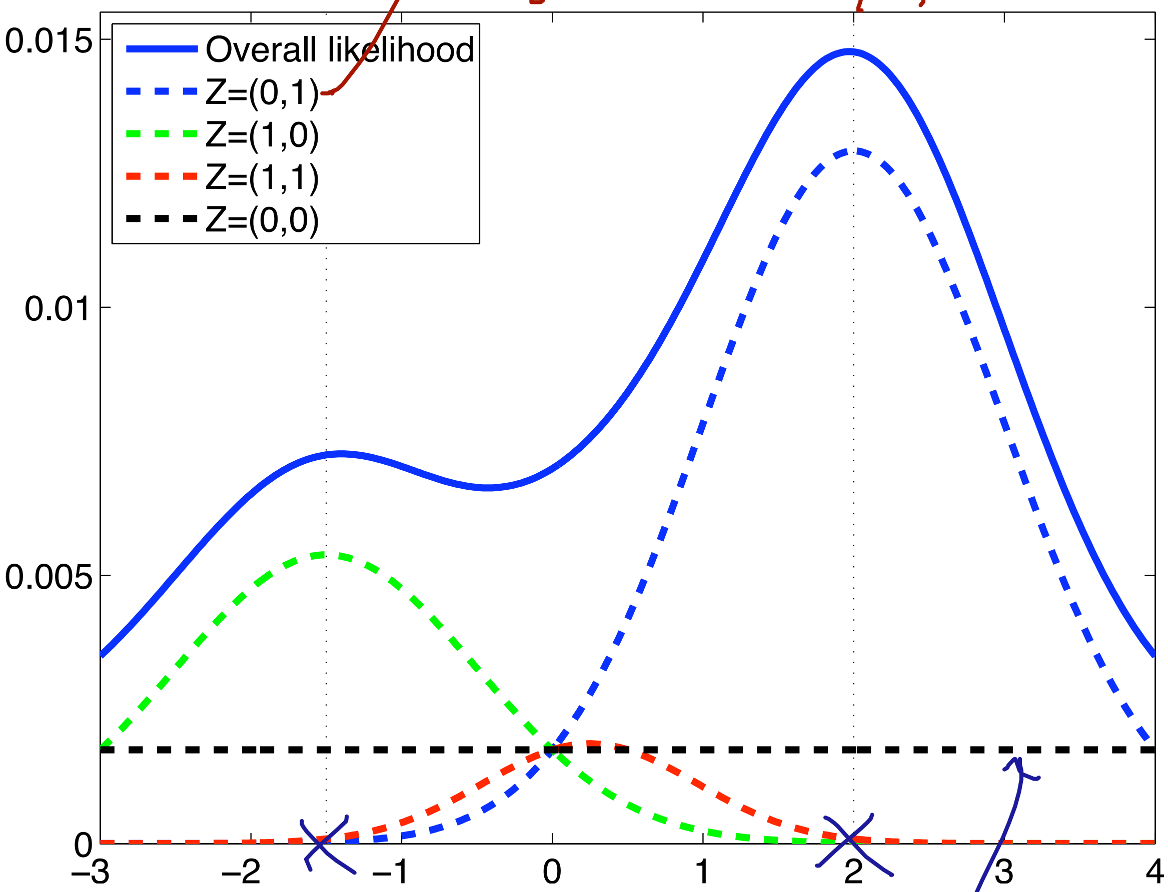
$\mu_0 = 0$ $\mu_1 = \mu$

$\sigma^2 = 1$

$p_0 = .5$ $p_1 = .5$

i.e., $z_1 = 0$ $z_2 = 1$

"x_2 associated w/ μ, x_1 not"



x_1 μ x_2 $Z = (0,0)$

2 data points:

$x_1 = -1.5$

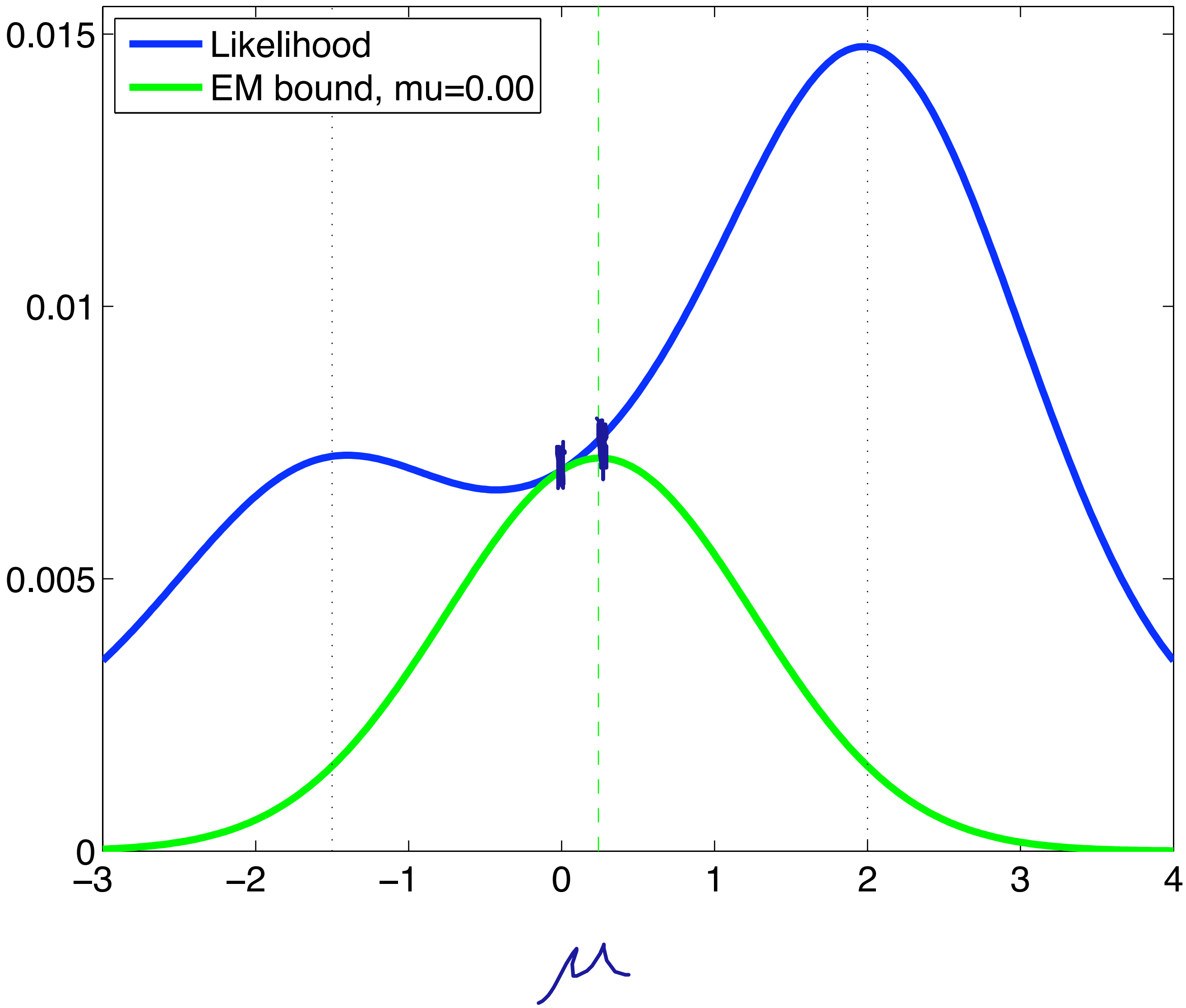
$x_2 = 2.0$

2 hidden vars:

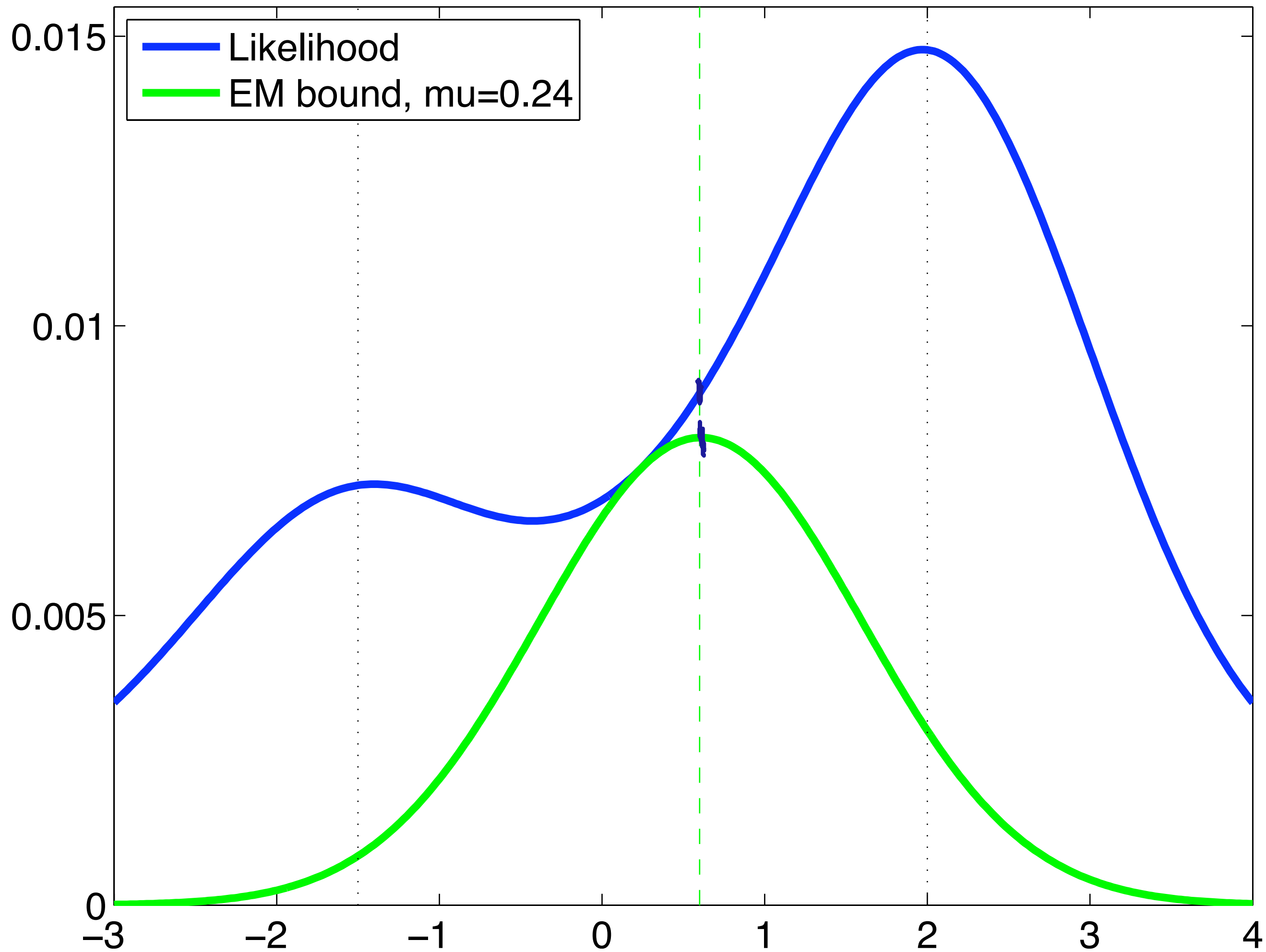
$z_1 \in \{0, 1\}$

$z_2 \in \{0, 1\}$

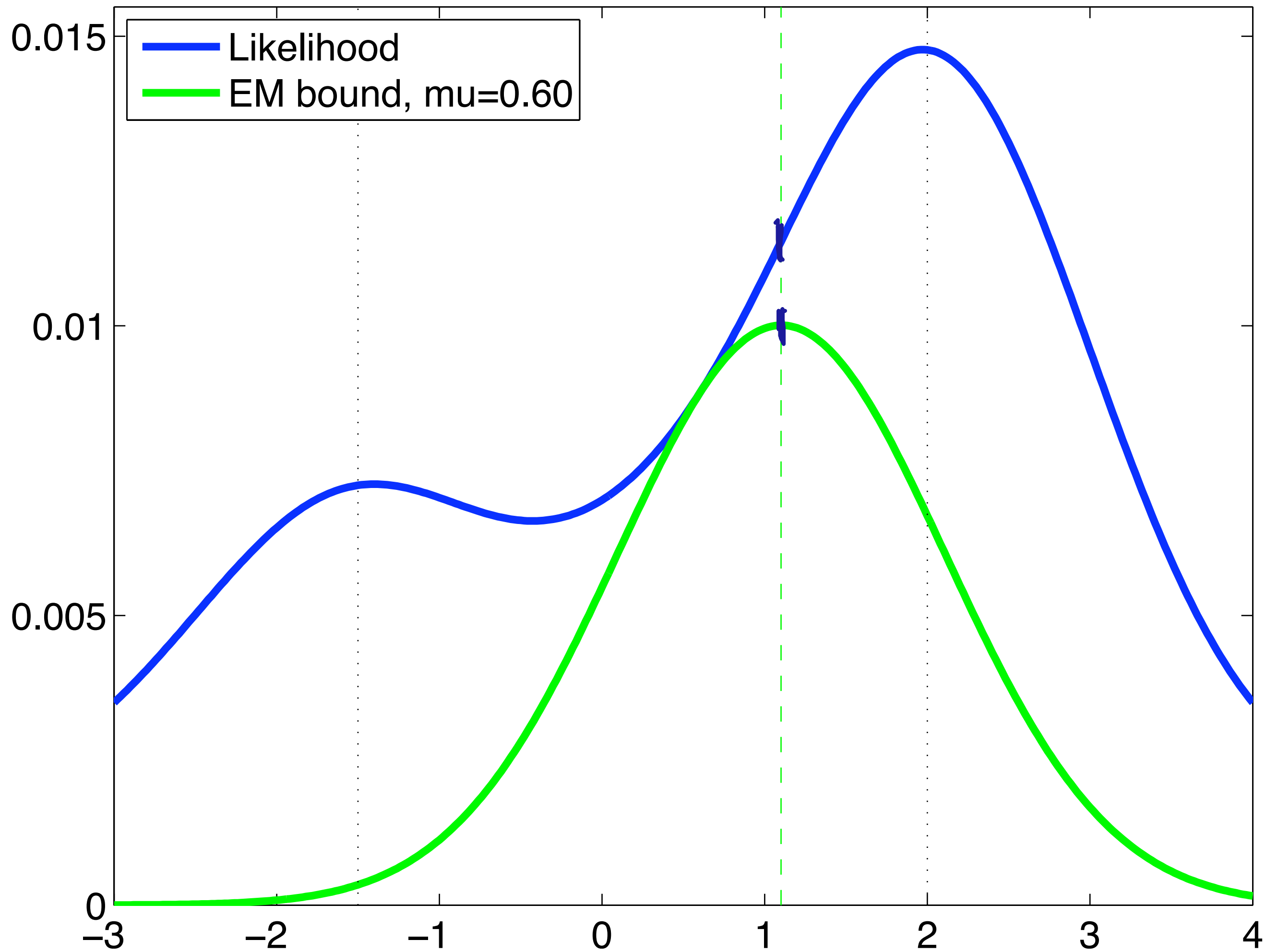
Example



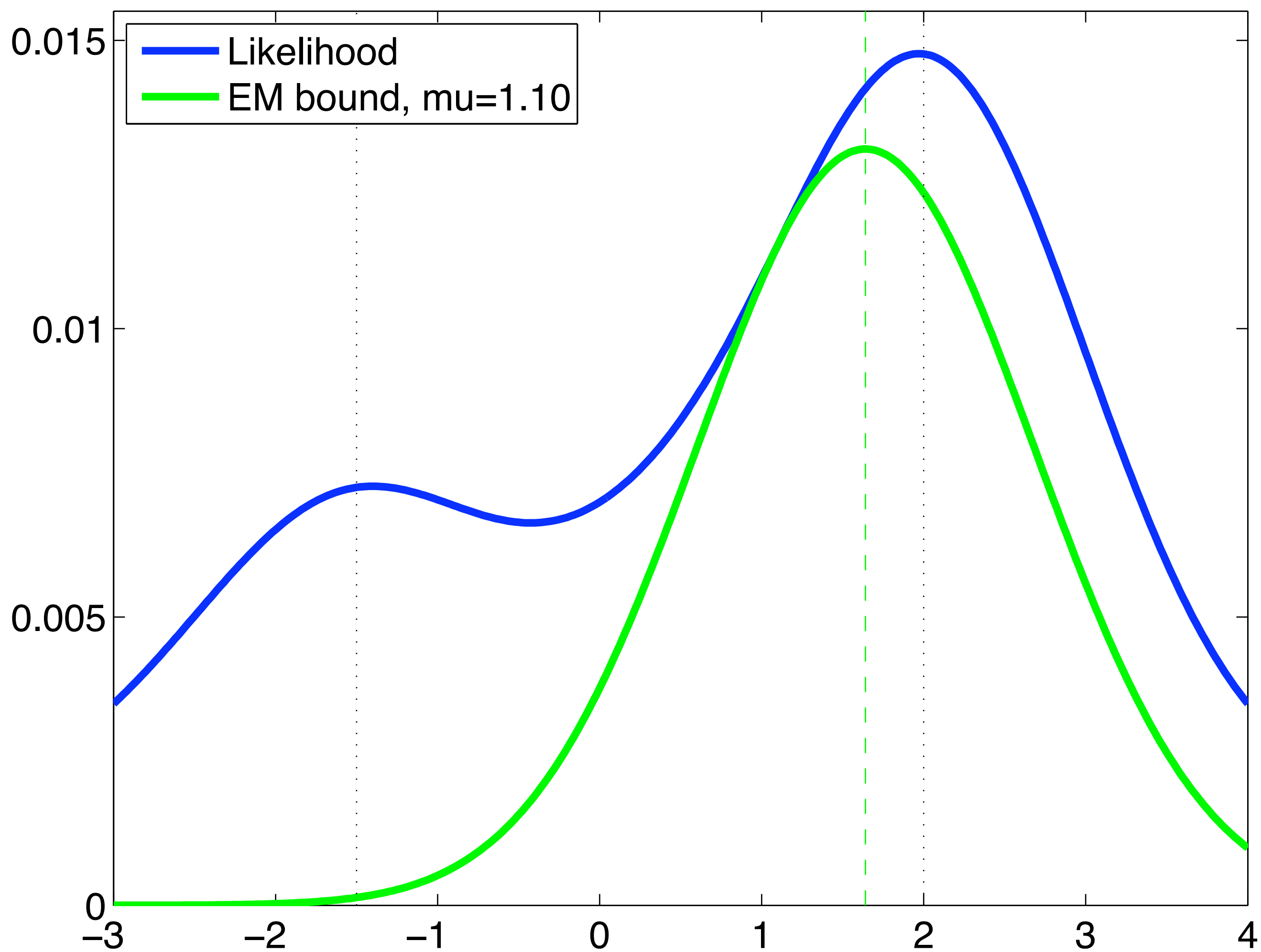
Example



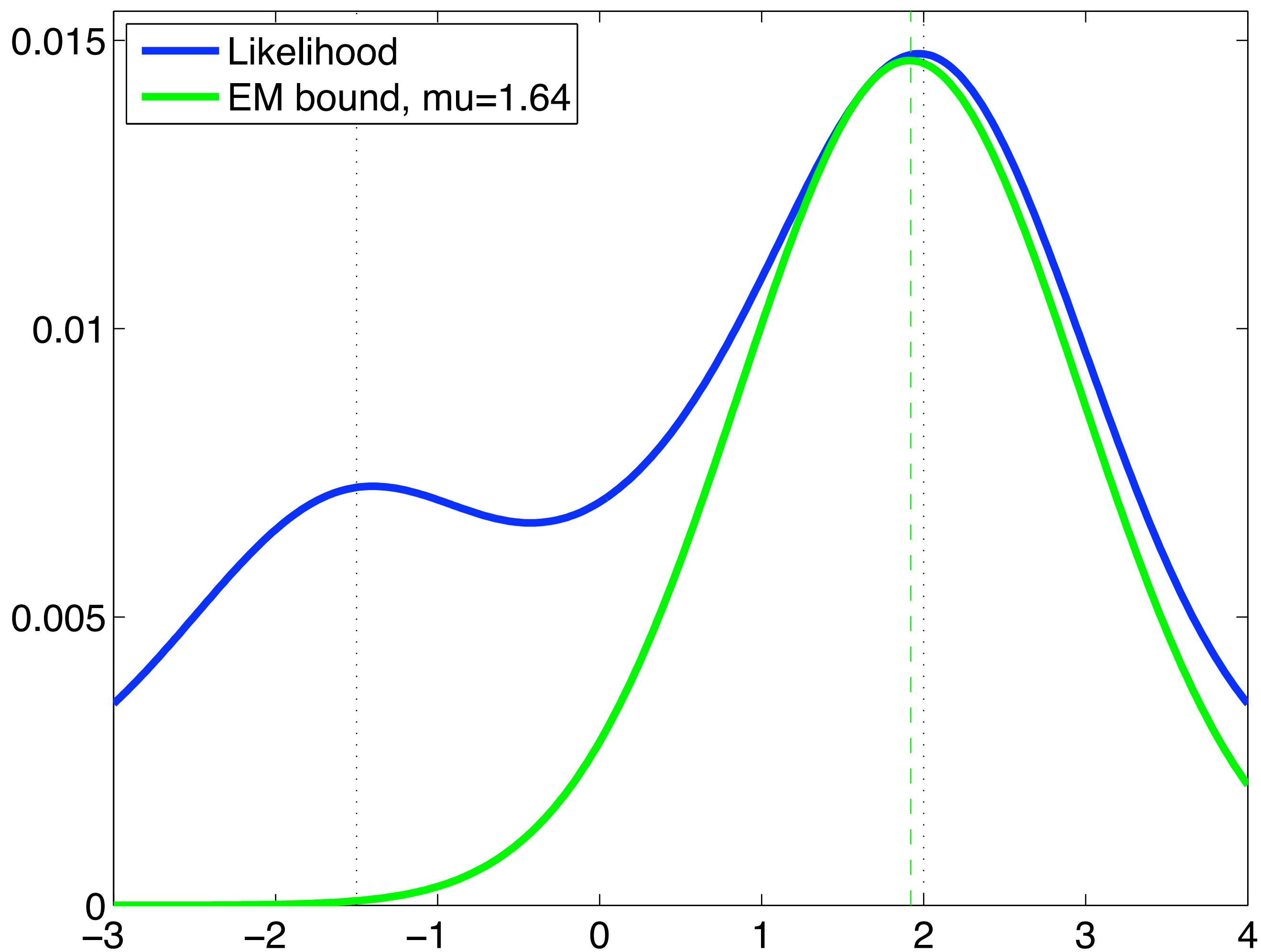
Example



Example

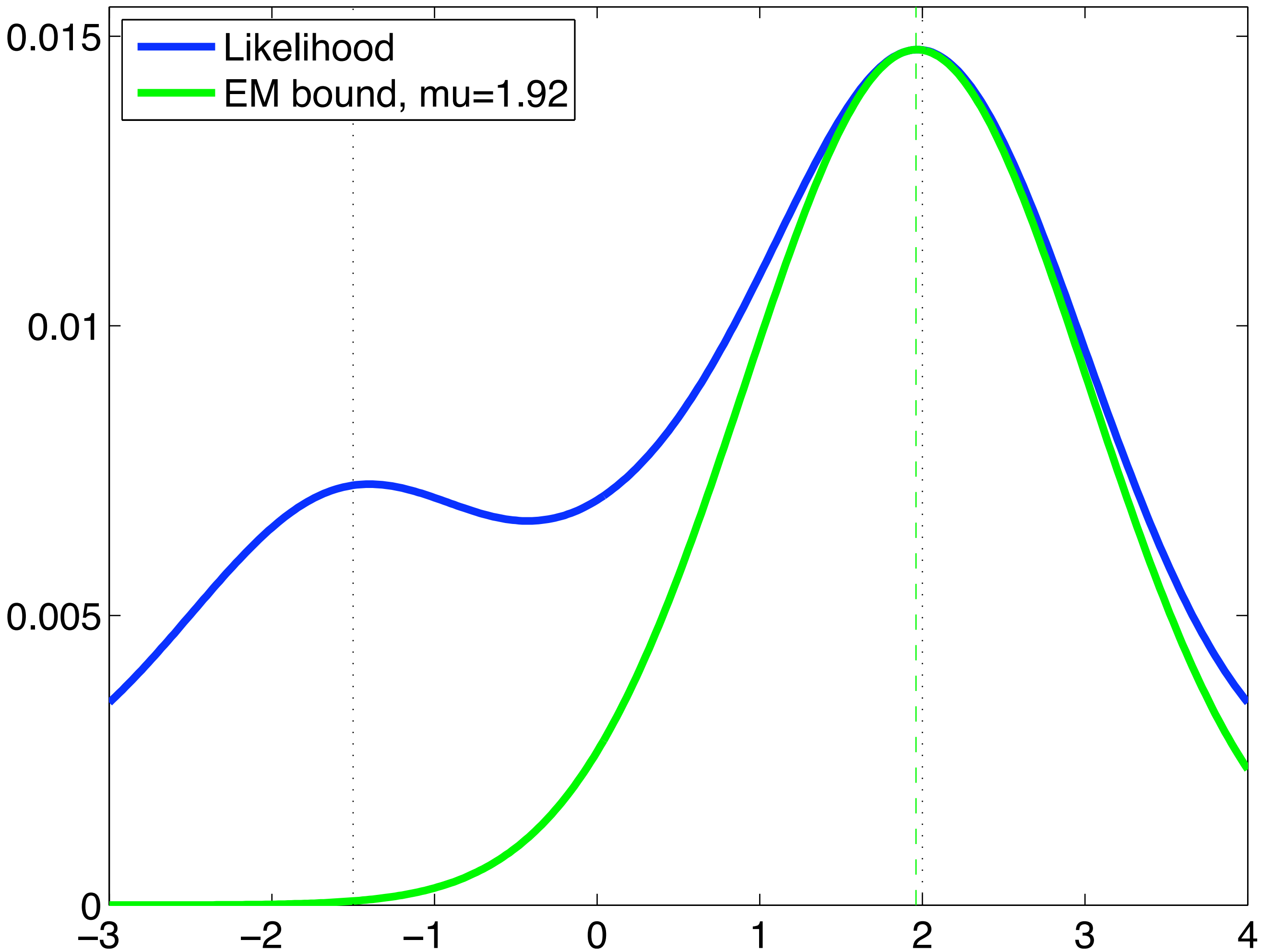


Example



Example

opt $\mu = 1.96$



$\sum_z q_z = 1$

Optimizing: q

$L(\theta, q, \lambda)$

- $L(\theta, q) = \sum_z q_z \overbrace{\ln P(X, Z{=}z \mid \theta)}^{\ell_z} - \sum_z q_z \ln q_z + \lambda\left(\sum_z q_z - 1\right)$

$$\frac{d}{dq}(q \ln q) = q\frac{1}{q} + \ln q = 1 + \ln q$$

$$\frac{d}{dq_z} L(\theta, q, \lambda) = \ell_z - (1 + \ln q_z) + \lambda = 0$$

$$\ln q_z = \ell_z - 1 + \lambda$$

$$q_z = e^{\ell_z} e^{-1+\lambda} \propto e^{\ell_z} = P(X, Z{=}z \mid \theta)$$

$$q_z = P(X, Z{=}z \mid \theta) / P(X \mid \theta) = P(Z{=}z \mid X, \theta)$$

Optimal q is conditional probability

For soft k-means

k^N of these

- $q_z = P(Z=z \mid X, \theta) = \prod_i \prod_j q_{ij}^{z_{ij}}$

$q_{ij} = P(z_{ij}=1 \mid X, \theta) = p_j N(x_i \mid \mu_j, \Sigma_j)$

$N \times k$ of these

$N \times K$

$\boxed{\text{E - step}}$

Simplifying the bound

- $L(\theta, q) = \sum_z q_z \ln P(X, Z{=}z \mid \theta) - \sum_z q_z \ln q_z$

$$= E_{z\sim q}\left(\ln P(X, Z{=}z \mid \theta)\right) + H(q) \leftarrow \text{~~neg~~entropy}$$

$$= E_q\left(\ln \prod_i \prod_j p_j^{z_{ij}} N(x_i \mid \mu_j, \Sigma_j)^{z_{ij}}\right) + H(q)$$

$$= E_q\left(\sum_i \sum_j \left[z_{ij} \ln p_j + z_{ij} \ln N(x_i \mid \mu_j\ \Sigma_j)\right]\right) + H(q)$$

$$= \sum_i \sum_j \underbrace{E_q(z_{ij})}_{q_{ij}} \left[\ln p_j + \ln N(\cdots)\right] + H(q)$$

Optimizing: μ

- $L(\theta, q) = \sum_i \sum_j q_{ij} [\ln p_{ij} + \ln N(X_i \mid \mu_j, \Sigma_j)] + H(q)$

$$0 = \sum_i q_{ij} \left(-2 \Sigma_j^{-1} X_i + 2 \Sigma_j^{-1} \mu_j\right)$$

$$\sum_i q_{ij} \cancel{(2 \Sigma_j^{-1})} X_i = \sum_i q_{ij} \cancel{(2 \Sigma_j^{-1})} \mu_j$$

$$\mu_j = \sum_i q_{ij} X_i \Big/ \sum_i q_{ij}$$

$\boxed{\text{M - step}}$

Optimizing: p_j, Σ_j

- Won't derive here, but:
  - max wrt p and Σ are just like MLE
  - count each X_i w/ weight $E_q(Z_{ij})$

The EM algorithm

- Want to maximize $L(\theta) = \log P(X \mid \theta)$
- Hidden variables Z, so that
  - $L(\theta) = \log \sum_z P(X, Z = z \mid \theta)$
- Use bound: for any distribution q, $\log(\sum_z \exp(\ln P(X, Z = z \mid \theta))) \geq$

$q_z \geq 0$

$\sum_z q_z = 1$

$$\sum_z q_z (\ln P(X, Z = z \mid \theta) - \ln q_z) + H(q)$$

The EM algorithm

- Alternating optimization
  - of $L(\theta,q) = E_{Z\sim q}[\ln P(X, Z \mid \theta)] - H(q)$
  - E-step:

    $q \leftarrow P(Z \mid X, \theta)$ conditional dist'n

    simplify $L(\theta, q)$
  - M-step: max wrt θ

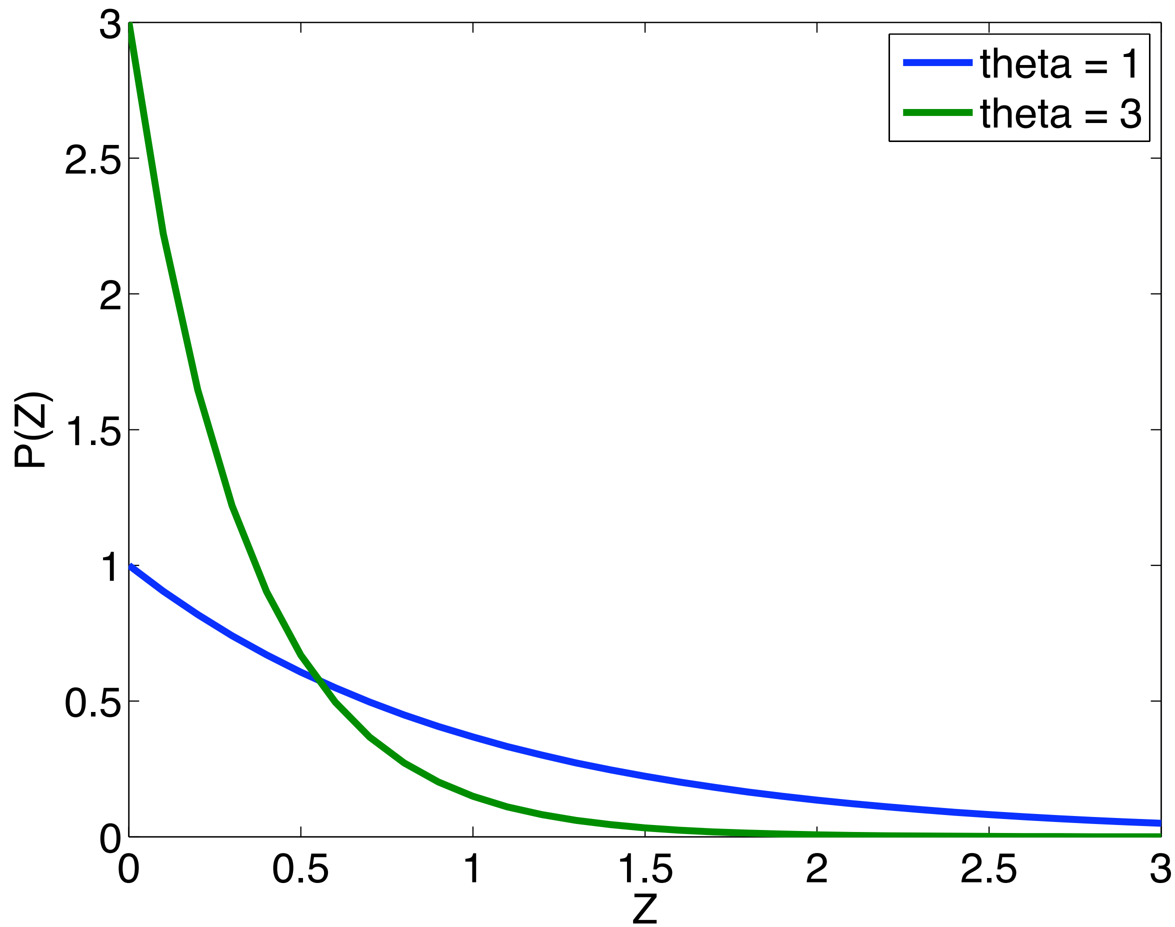
Example: failure times

- You're GE, testing light bulbs to estimate failure rate / lifetime
  - run torture test on 1000 bulbs for 1000 hrs
  - data: 503 bulbs fail at times $X_1, X_2, \ldots, X_{503}$
  - 497 bulbs are still going after 1000 hrs
- Or, you're an MD running a 5-year study, estimating mortality rate due to Emacsitis
  - of 1000 patients, 214 die at times $X_1, \ldots, X_{214}$
  - remaining 786 are alive at end of study

EM for survival analysis

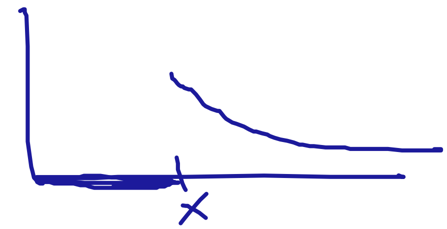
- Hidden data: when would remaining samples have failed, if we had been patient enough to watch that long?
  - $Z_i = X_i$ for failed samples
  - $Z_i \geq X_i$ for remaining samples
- $P(X_i = x \mid \theta) = \theta e^{-\theta x}$ for $x \geq 0$

Exponential distribution



- $P(X = x \mid \theta) = \theta e^{-\theta x}$ (for $x \geq 0$)

Properties of exponential distribution



- $E(X \mid \theta) =$ $1/\theta$
- $P(Z = z \mid \theta, Z \geq X) =$ $\theta e^{\theta(z-X)}$ $z \geq X$
- $E(Z \mid \theta, Z \geq X) =$ $X + 1/\theta$

EM algorithm for survival analysis

- E-step: for each ~~censored~~ point, compute
  - $E(Z_i \mid X_i) =$ $\left\{ \begin{array}{ll} x_i & \text{if } i \text{ obs} \\ x_i + 1/\theta & \text{if } i \text{ cens} \end{array} \right\} = \hat{x}_i$
- M-step: compute MLE
  - with fully-observed data, MLE is:

    $\frac{1}{\theta} = \sum_i x_i / N$
  - with censored data:

    $1/\theta = \sum_i \hat{x}_i / N$

Fixed point

- If there are K censored observations, EM converges to:

$$\frac{1}{\theta} = \left[\frac{1}{N} \sum_i x_i\right] \frac{N}{N-K}$$

$\uparrow$ # censored $= K$

- Note: it's unusual to have closed-form expression for fixed point

More examples of EM

- Regression / classification with missing input values
- Learning parameters of Kalman filters ← tracking a robot position
- Learning params of hidden Markov models ← speech recognition
  - "forward-backward", "Baum-Welsh"
- Learning parameters of NL parsers
  - "inside-outside"