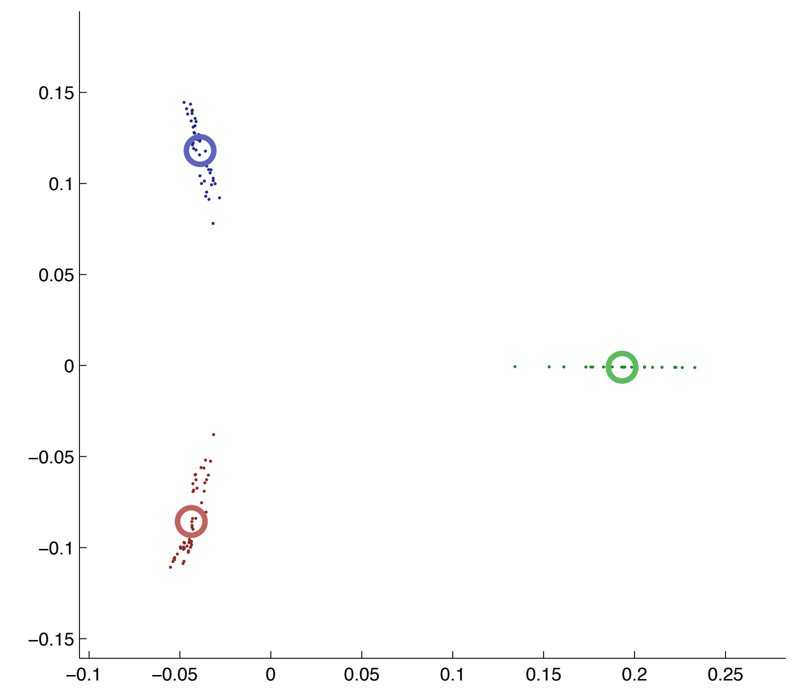
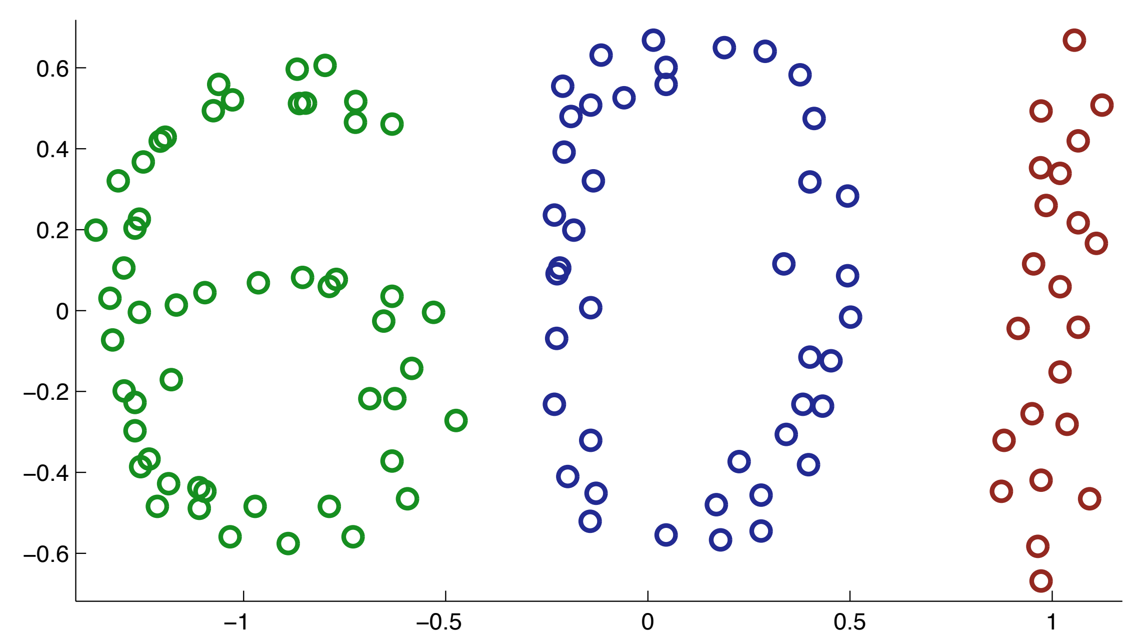
Review

- 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
  - relation to hard k-means
  - connection to naïve Bayes
  - (un)biasedness

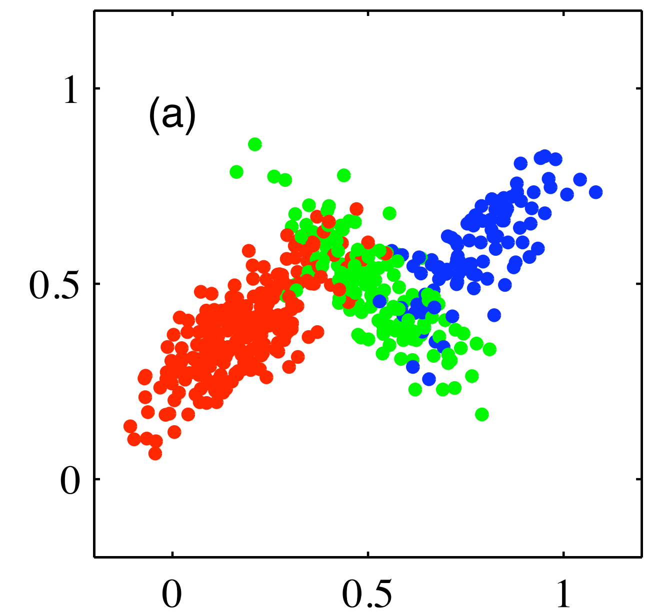


fig 9.5a from Bishop

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"
  - $q_{ij} =$
- Update means: "M step" for max
  - $\mu_j = \sum_i q_{ij} x_i \Big/ \sum_i q_{ij}$
- Possibly: update covariances
  - $\Sigma = \sum_i \sum_j q_{ij} (x_i - \mu_j)(x_i - \mu_j)^T \Big/ N$
- Repeat

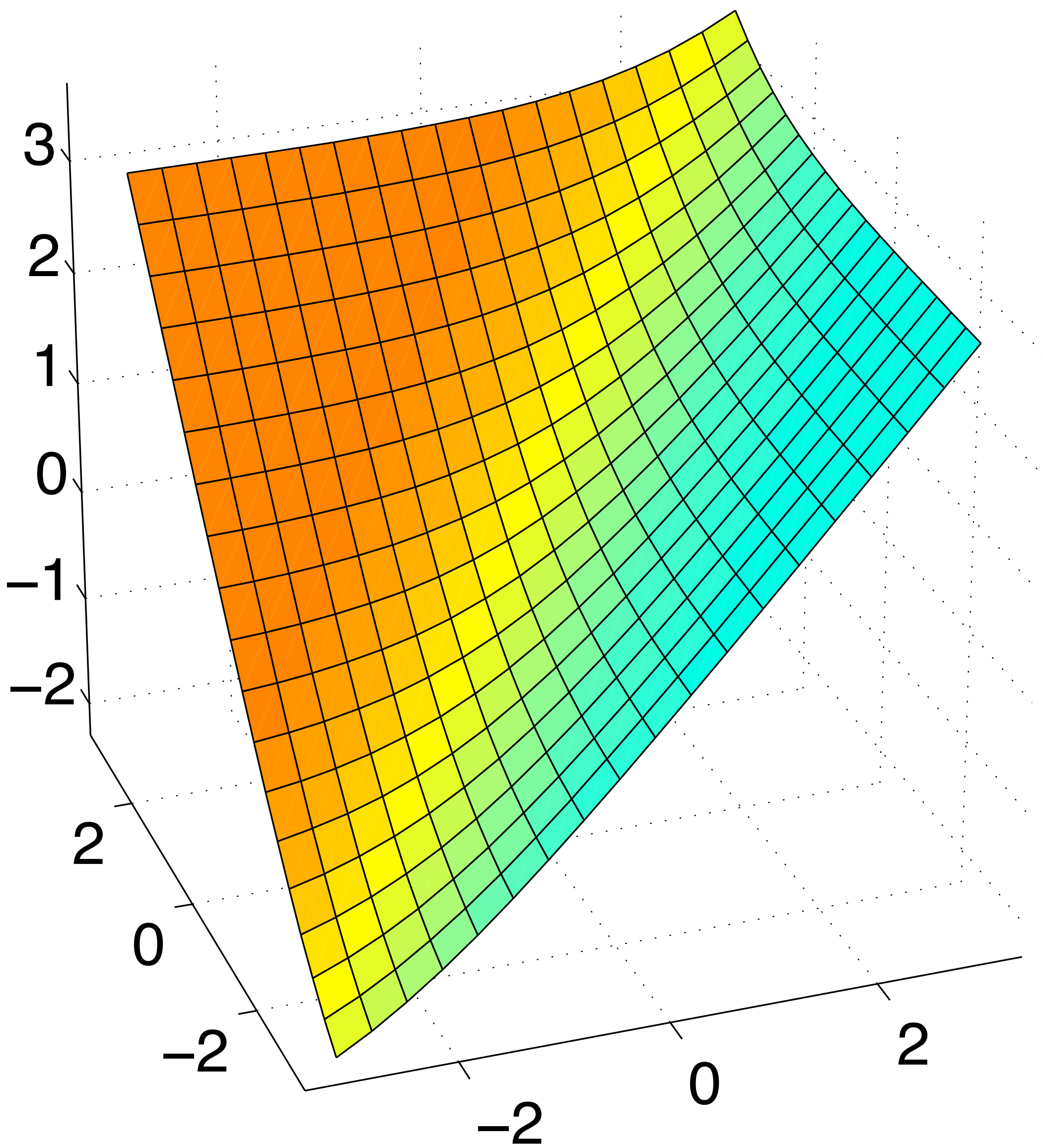
Deriving soft k-means

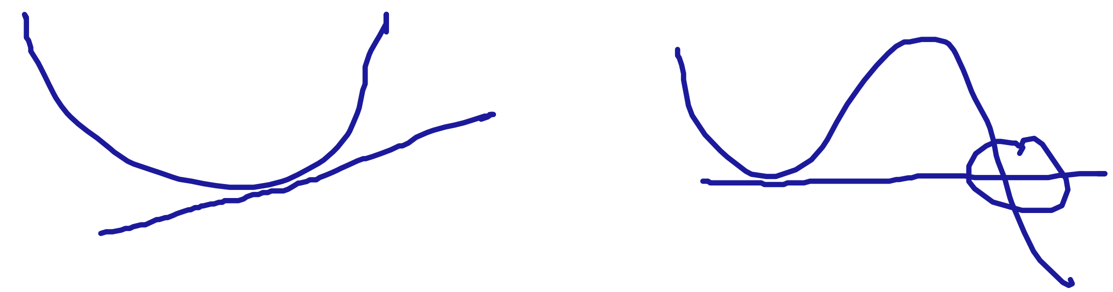
- $P(X_i \mid Z_{ij} = 1, \theta) = \text{Gaussian}(\mu_j, \Sigma_j)$
- $P(Z_{ij} = 1 \mid \theta) = p_j$
- $P(X_i, Z_{i\cdot} \mid \theta) =$ $\prod_j p_j^{z_{ij}} N(x_i \mid \mu_j, \Sigma_j)^{z_{ij}}$
- $L = \ln P(X \mid \theta) =$

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

soft max

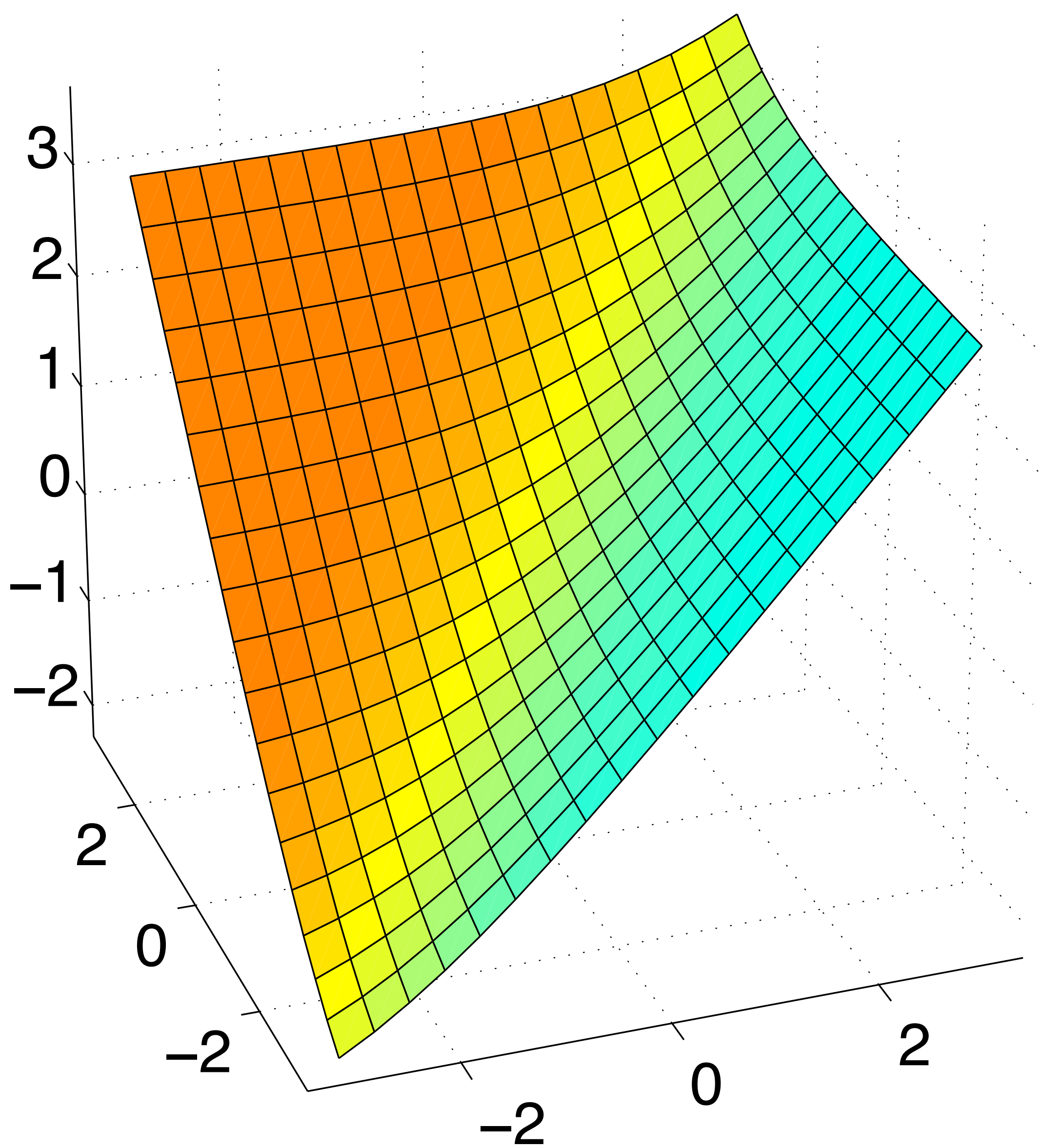
$\approx \max(x,y)$ $$f(x,y) = \ln(e^x+e^y)$$ ← convex





Convex fns are lower-bounded by tangents

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



In general

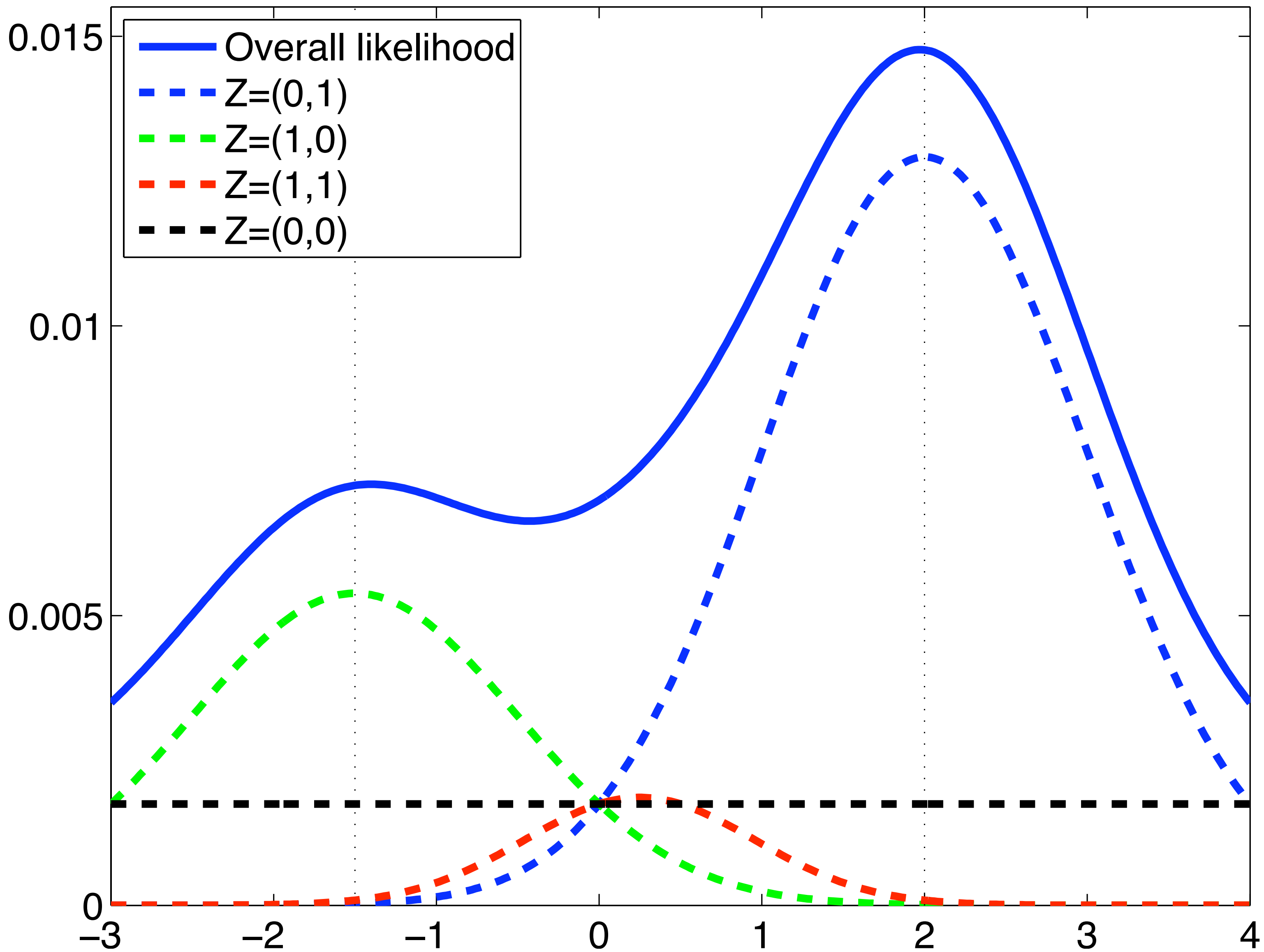
- $f(x_1, x_2, \ldots) = \ln(\sum_i \exp(x_i)) \geq$

  for any probability distribution q:

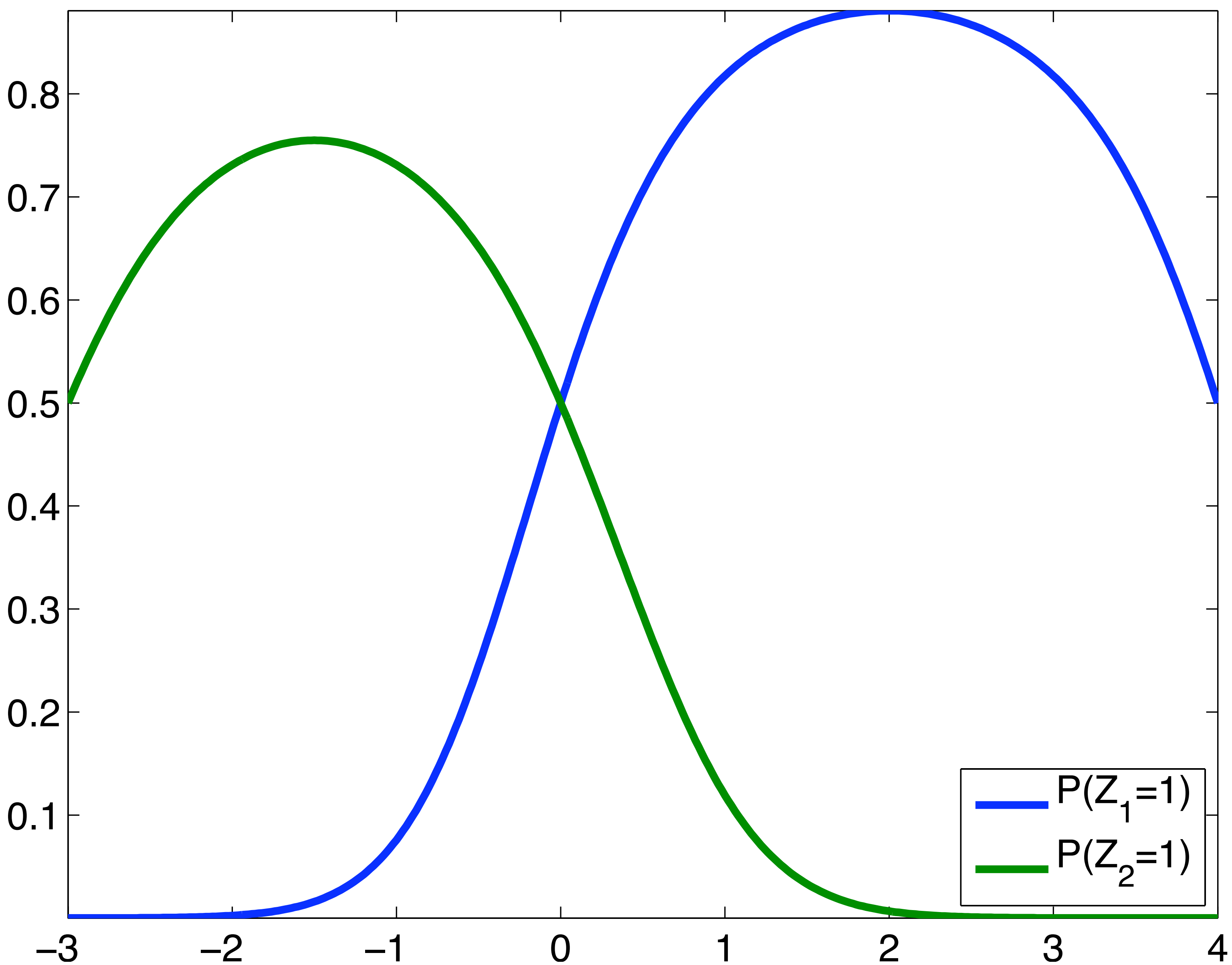
Optimizing a bound

- $L(\theta) = \ln \sum_{z} \exp(\ln P(X, Z{=}z \mid \theta)) \geq$
    - for any distribution $q = \langle q_z \rangle$
- 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

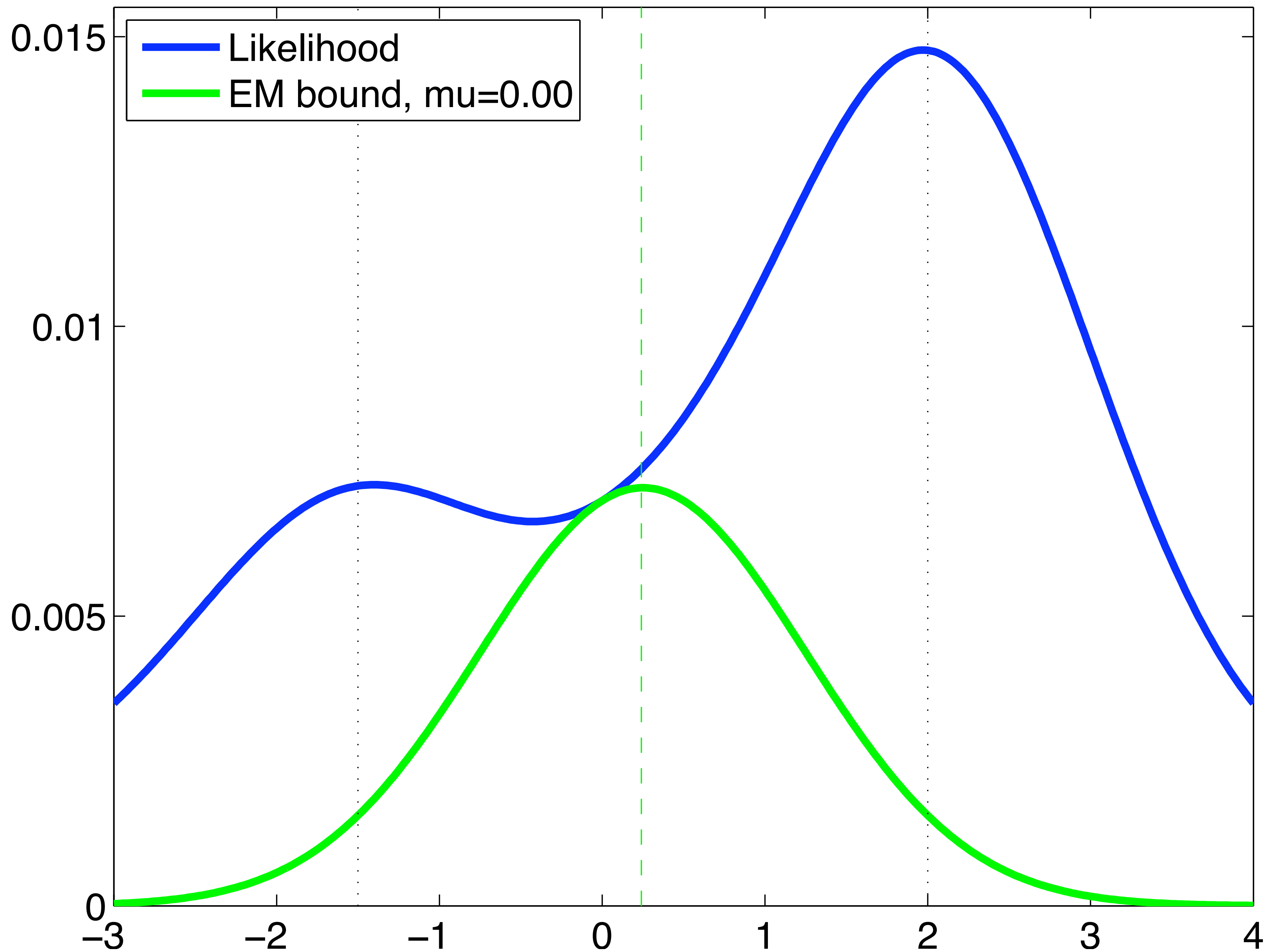
Example



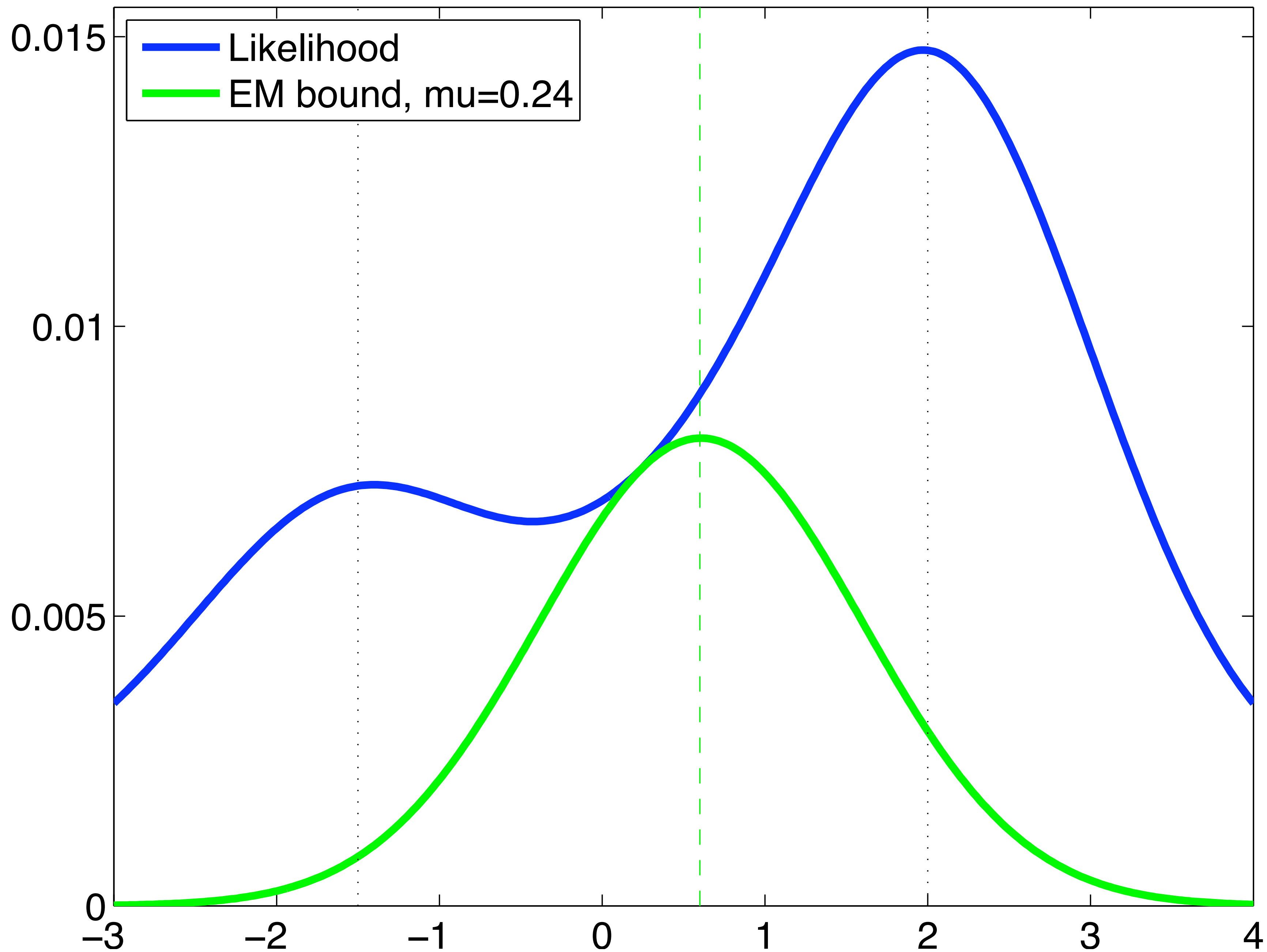
Example



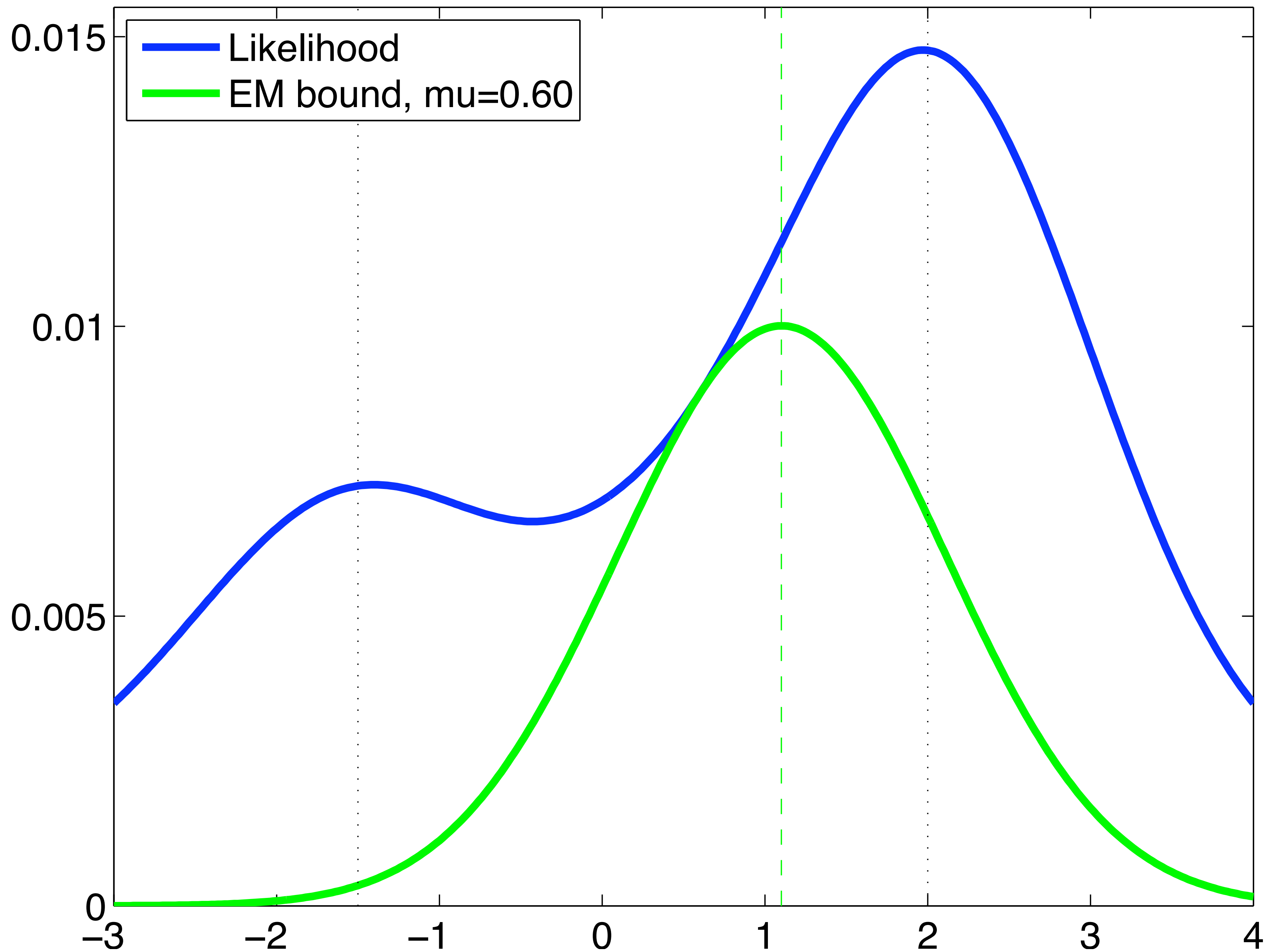
Example



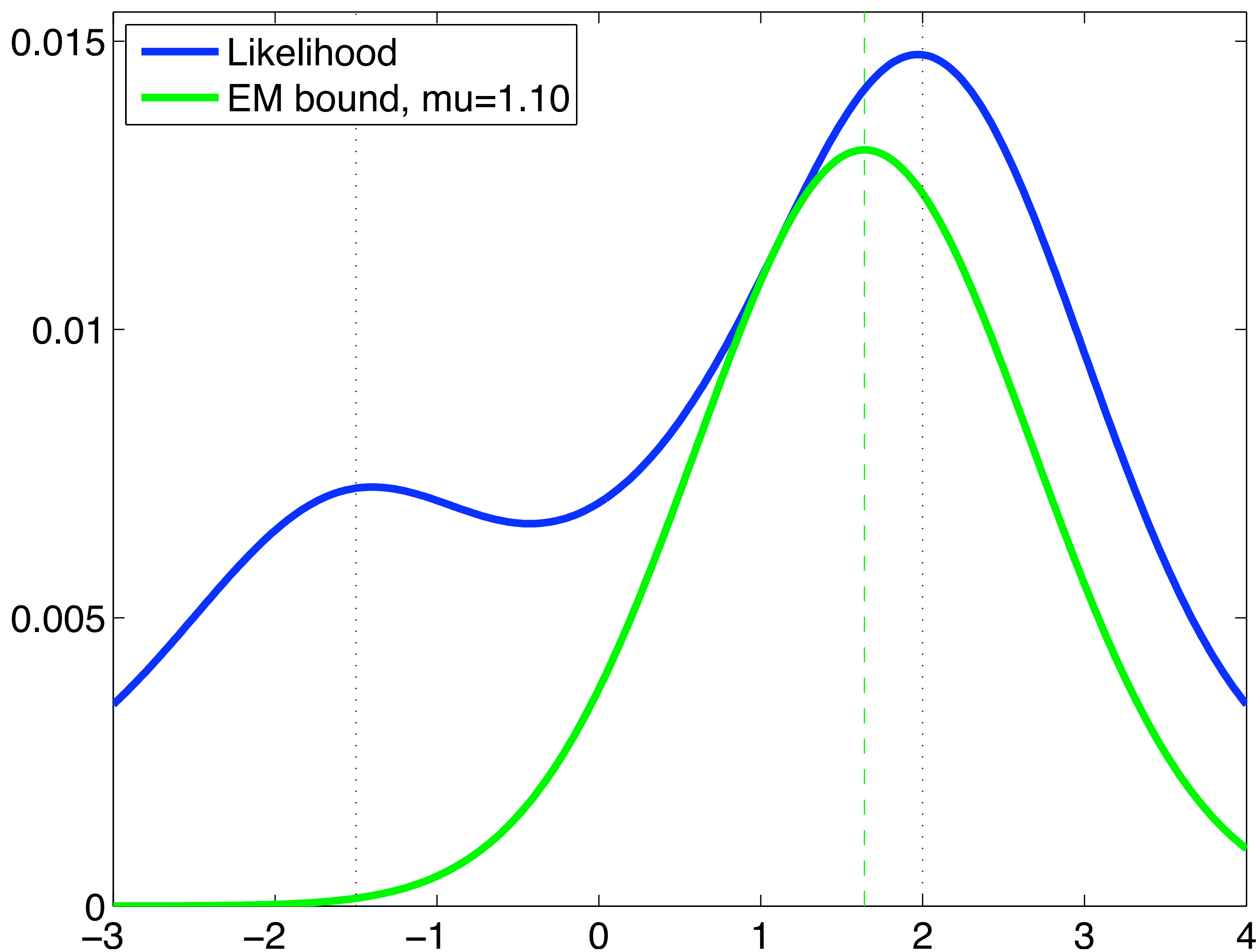
Example



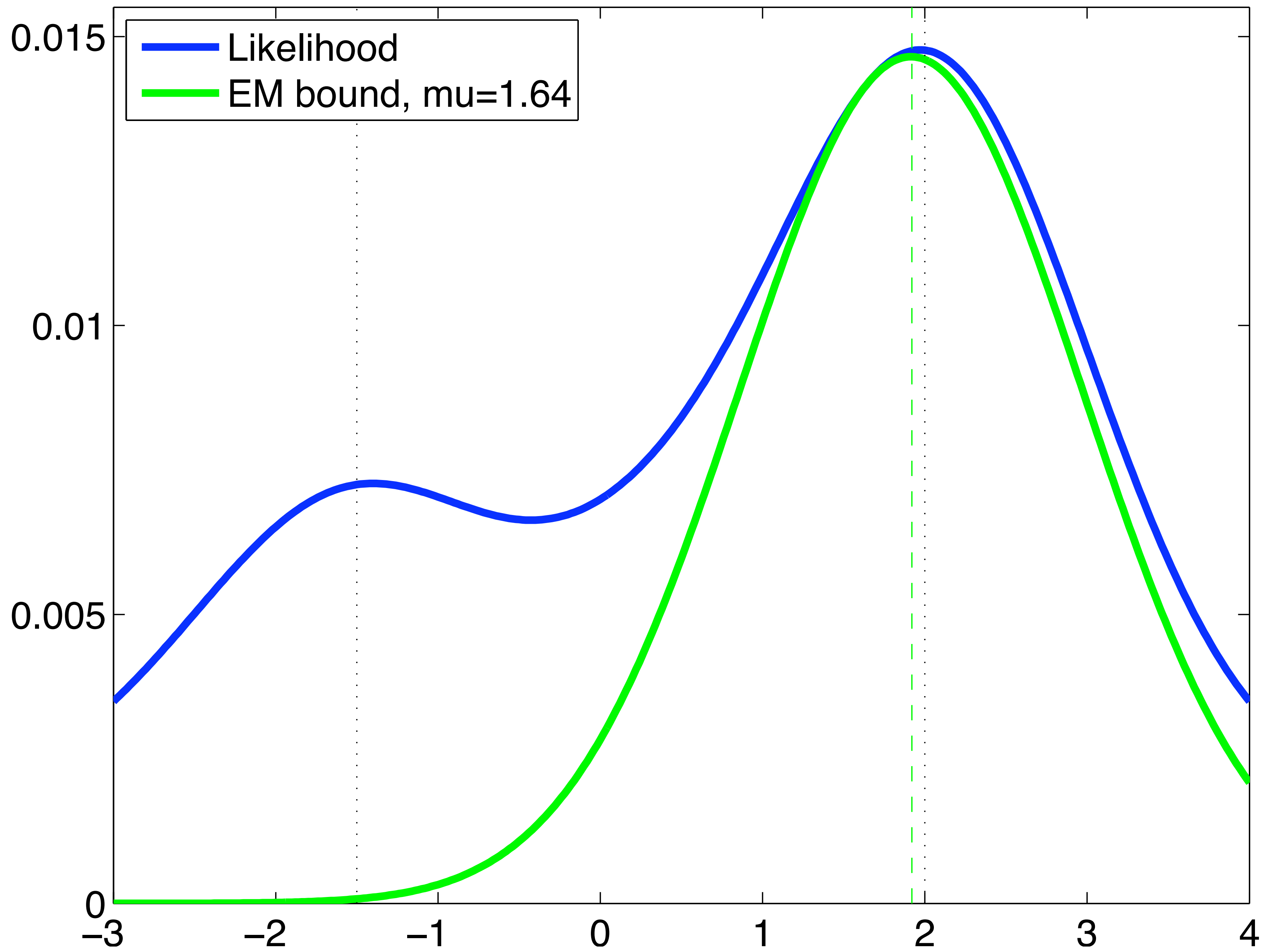
Example



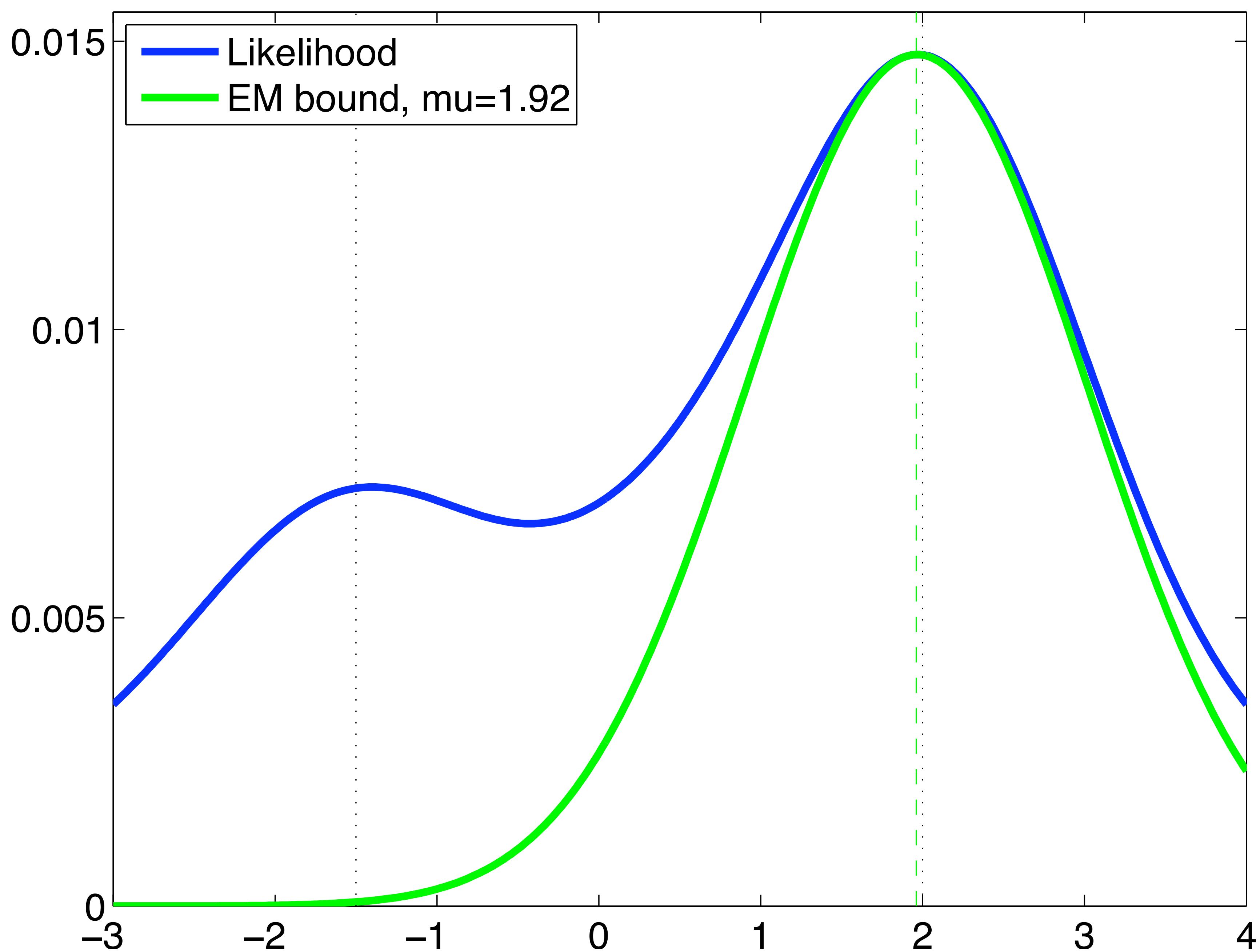
Example



Example



Example



Optimizing: q

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

For soft k-means

- $q_z = P(Z=z \mid X, \theta)$

Simplifying the bound

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

Optimizing: μ

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

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, $\log(\sum_z \exp(\ln P(X, Z = z \mid \theta))) \geq$

The EM algorithm

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

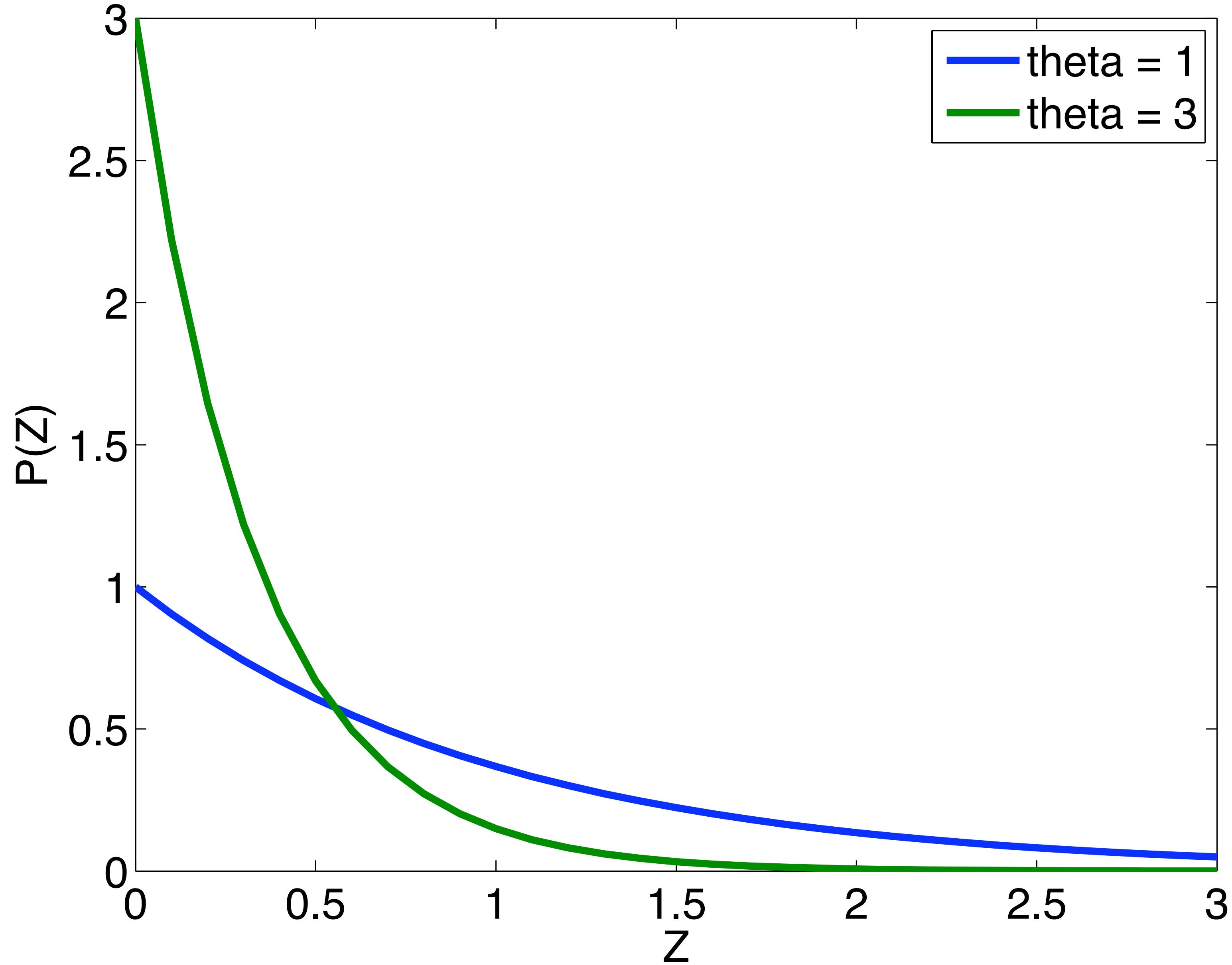
Example: failure times

- 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

- $E(X \mid \theta) =$
- $P(Z = z \mid \theta, Z \geq X) =$
- $E(Z \mid \theta, Z \geq X) =$

EM algorithm for survival analysis

- E-step: for each censored point, compute
  - $E(Z_i \mid X_i) =$
- M-step: compute MLE
  - with fully-observed data, MLE is:
  - with censored data:

Fixed point

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

- 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
- Learning params of hidden Markov models
    - “forward-backward”, “Baum-Welsh”
- Learning parameters of NL parsers
    - “inside-outside”