Computer Science

CSC580: Principles of Data Science

Parameter Learning and Expectation Maximization

Prof. Jason Pacheco

Parameter Estimation

We have a <u>model</u> in the form of a probability distribution, with unknown **parameters of interest** θ **,**

$$p(X;\theta)$$

Observe data, typically *independent identically distributed (iid),*

$$\{x_i\}_i^N \stackrel{iid}{\sim} p(\cdot;\theta)$$

Compute an **estimator** to approximate parameters of interest,

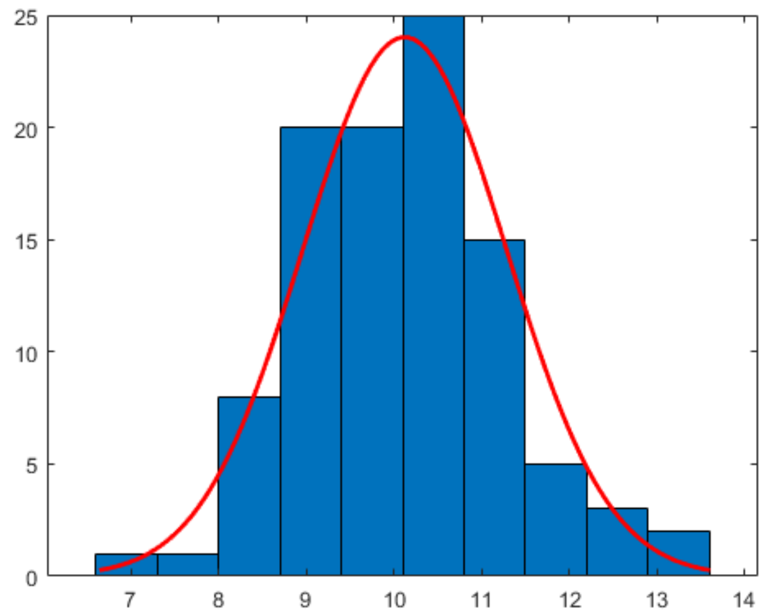
$$\hat{\theta}(\{x_i\}_i^N) \approx \theta$$

Many different types of estimators, each with different properties

Estimating Gaussian Parameters

Suppose we observe the heights of N student at UA, and we model them as Gaussian:

$$\{x_i\}_i^N \sim \mathcal{N}(\mu, \sigma^2)$$



How can we estimate the **mean**?

$$\hat{\mu} = \frac{1}{N} \sum_i x_i \approx \mu$$

Sample mean
$\bar{x}$

How can we estimate the **variance?**

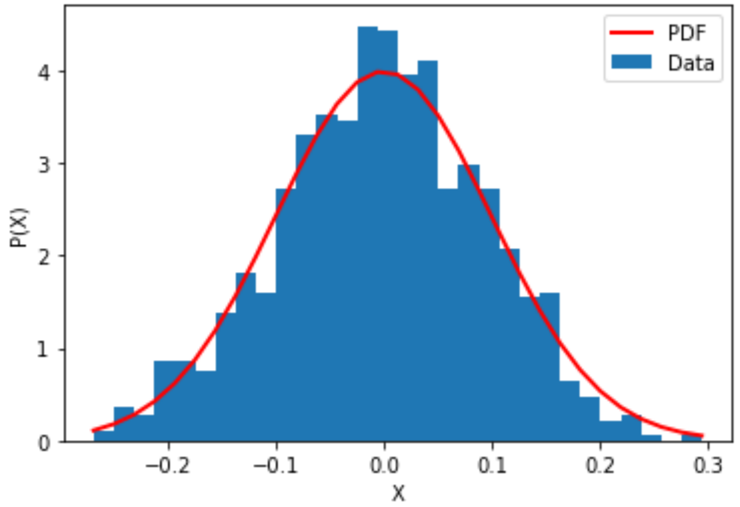
$$\hat{\sigma^2} = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2 \approx \sigma^2$$

Variance estimator uses our previous mean estimate. This is a **plug-in estimator.**

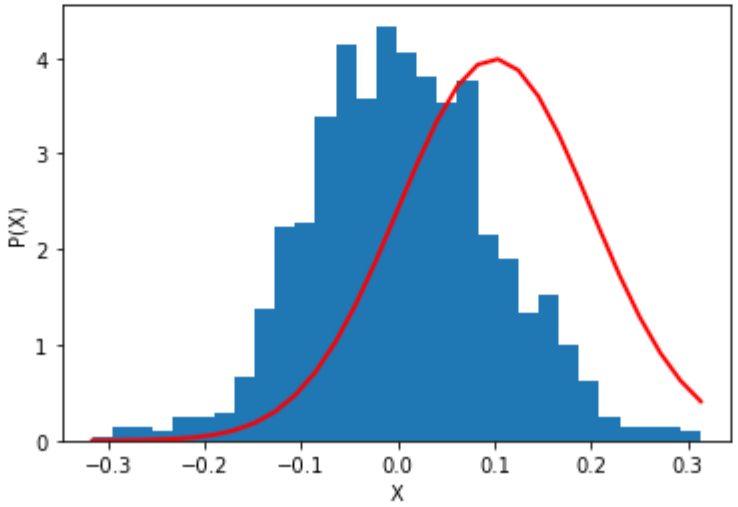
Likelihood (Intuitively)

Suppose we observe N data points from a Gaussian model and wish to estimate model parameters…

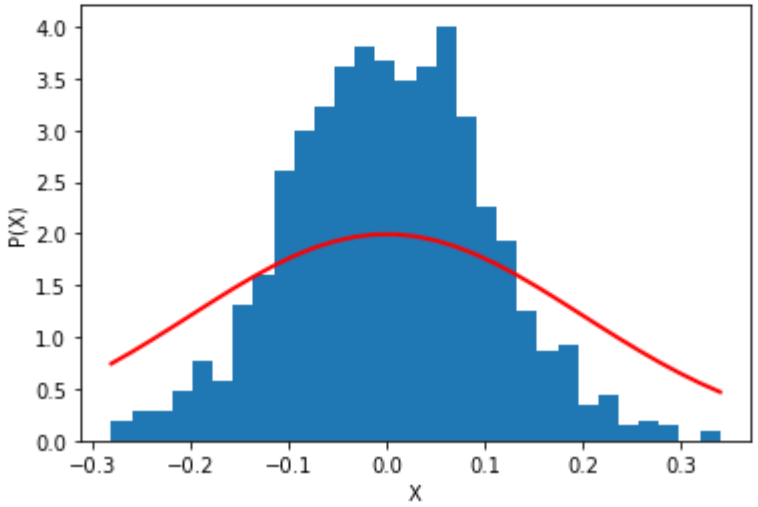
High Likelihood



Low Likelihood (mean)



Low Likelihood (variance)



Likelihood Principle *Given a statistical model, the likelihood function describes all evidence of a parameter that is contained in the data.*

Likelihood Function

Suppose $x_i \sim p(x;\theta)$, then what is the **joint probability** over N *independent identically distributed* (iid) observations $x_1, \ldots, x_N$?

$$p(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

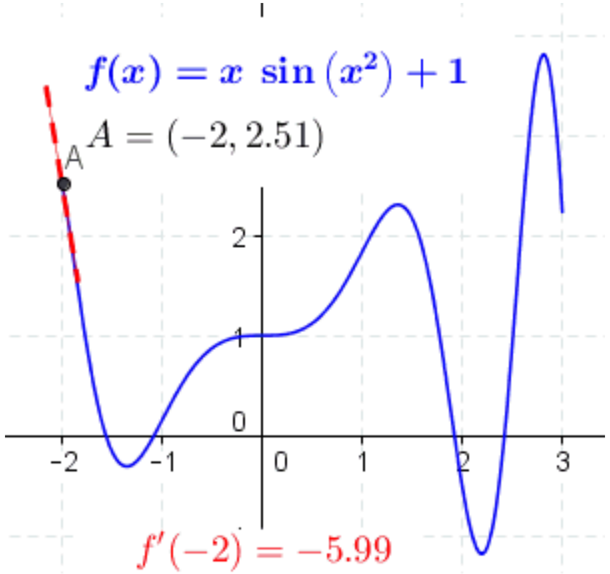
- We call this the **likelihood function**
- It is a function of the parameter θ -- the data are fixed
- Measure of how well parameter θ describes data (*goodness of fit*)

How could we use this to estimate a parameter θ*?*

Maximum Likelihood

Maximum Likelihood Estimator (MLE) as the name suggests, maximizes the likelihood function.

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i; \theta)$$



Question How do we find the MLE?

Answer Remember calculus…

Is $\mathcal{L}_N(\theta)$ concave?

Yes → Unique, closed-form solution →

Approach

- Compute derivative $\frac{d\mathcal{L}_N}{d\theta}$
- Set to zero and solve

$$\frac{d\mathcal{L}_N}{d\theta} = 0 \Rightarrow \hat{\theta}^{\text{MLE}}$$

No → Gradient-based optimization → Still have to compute derivative…

Maximum Likelihood

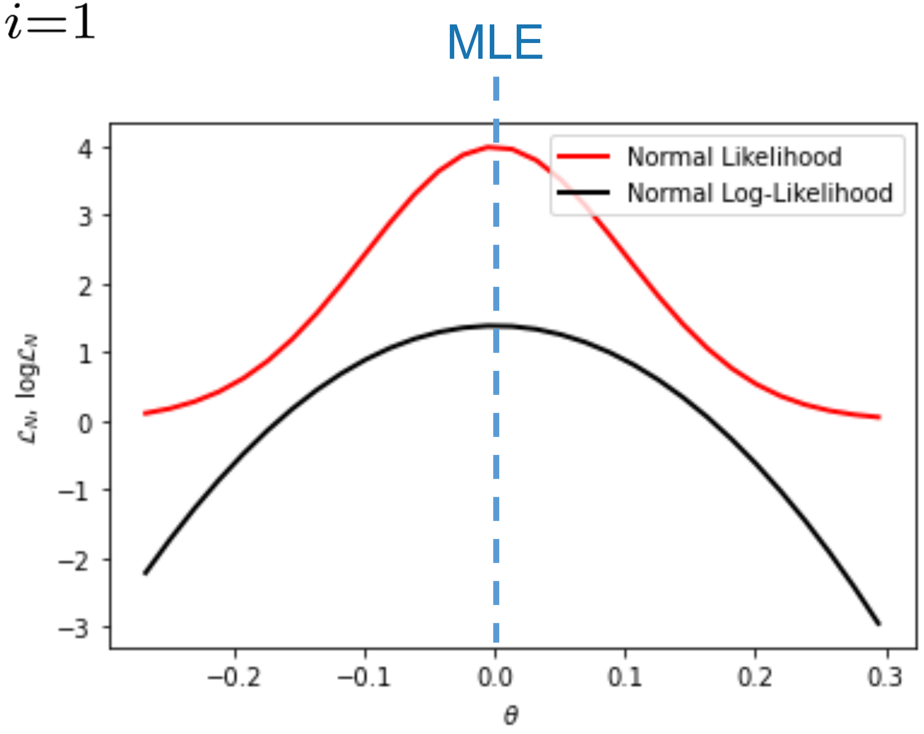
Maximizing log-likelihood makes the math easier (as we will see) and doesn't change the answer (logarithm is an increasing function)

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \; \log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \log p(x_i; \theta)$$

Derivative is a linear operator so,

$$\frac{d}{d\theta} \log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \frac{d}{d\theta} \log p(x_i; \theta)$$

One term per data point
Can be computed in parallel
(big data)



Marginal Likelihood

More often, we have a joint distribution with observations y, unknown variables z, and parameters θ

Marginal likelihood is normalizer of posterior:

$$p(z \mid y) = \frac{p(z)p(y \mid z)}{p(y)}$$

Bayes' Rule

$$p(z, y \mid \theta) = p(z \mid \theta)p(y \mid z, \theta)$$

Prior **Likelihood**

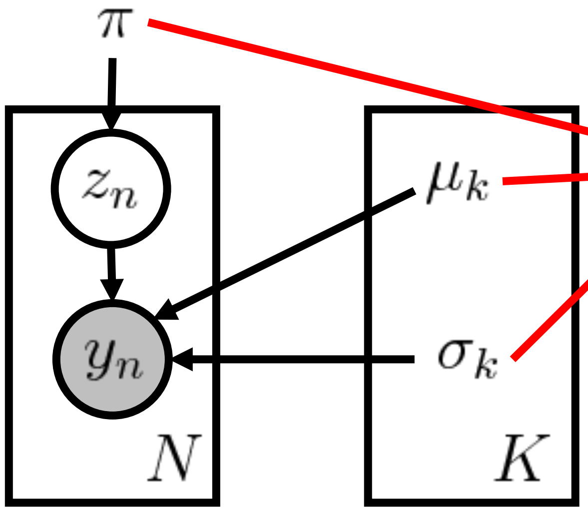
Need to *marginalize* out unknown variables, hence the name *marginal likelihood:*

$$p(y \mid \theta) = \int p(z \mid \theta)p(y \mid z, \theta)\, dz = \mathcal{L}(\theta)$$

Typically, this integral lacks a closed-form solution…so we need to compute *approximate* MLE solutions

Example: Gaussian Mixture Model

Model is often specified in terms of *unknown parameters*



GMM

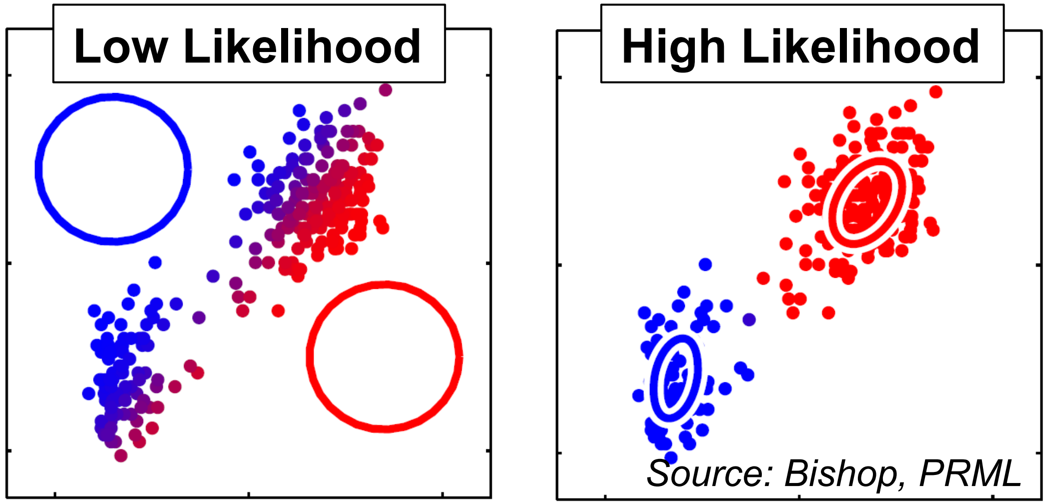
How *likely* are parameters for observed data?

$$\theta = \{\pi, \mu_1, \sigma_1, \ldots, \mu_K, \sigma_K\} \qquad \mathcal{Y} = \{y_1, \ldots, y_N\}$$

Marginal Likelihood (likelihood function):

$$p(\mathcal{Y} \mid \theta) = \sum_{z_1} \cdots \sum_{z_N} p(z_1, \ldots, z_N, \mathcal{Y} \mid \theta)$$

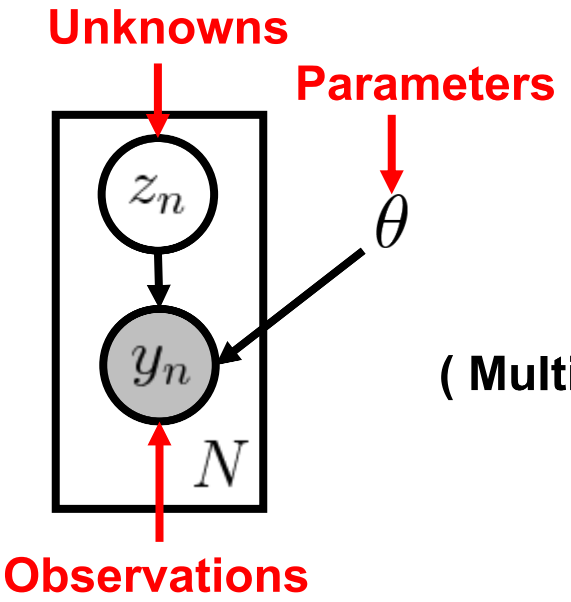
Sum over all possible K^N assignments, which we cannot compute



Intuition Learn / estimate parameters that assign highest probability (under the model) to data we've observed.

Lower Bounding Marginal Likelihood

Conditionally-independent model with partial observations…



$$\log p(\mathcal{Y} \mid \theta) = \log \sum_{z_1} \cdots \sum_{z_N} p(z_1, \ldots, z_N, \mathcal{Y} \mid \theta)$$

(Multiply by q(z)/q(z)=1) $$= \log \sum_{z} p(z, \mathcal{Y} \mid \theta) \left(\frac{q(z)}{q(z)} \right)$$

Shorthand
$z = z_1, \ldots, z_N$

(Definition of Expected Value) $$= \log \mathbf{E}_q \left[\frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right]$$

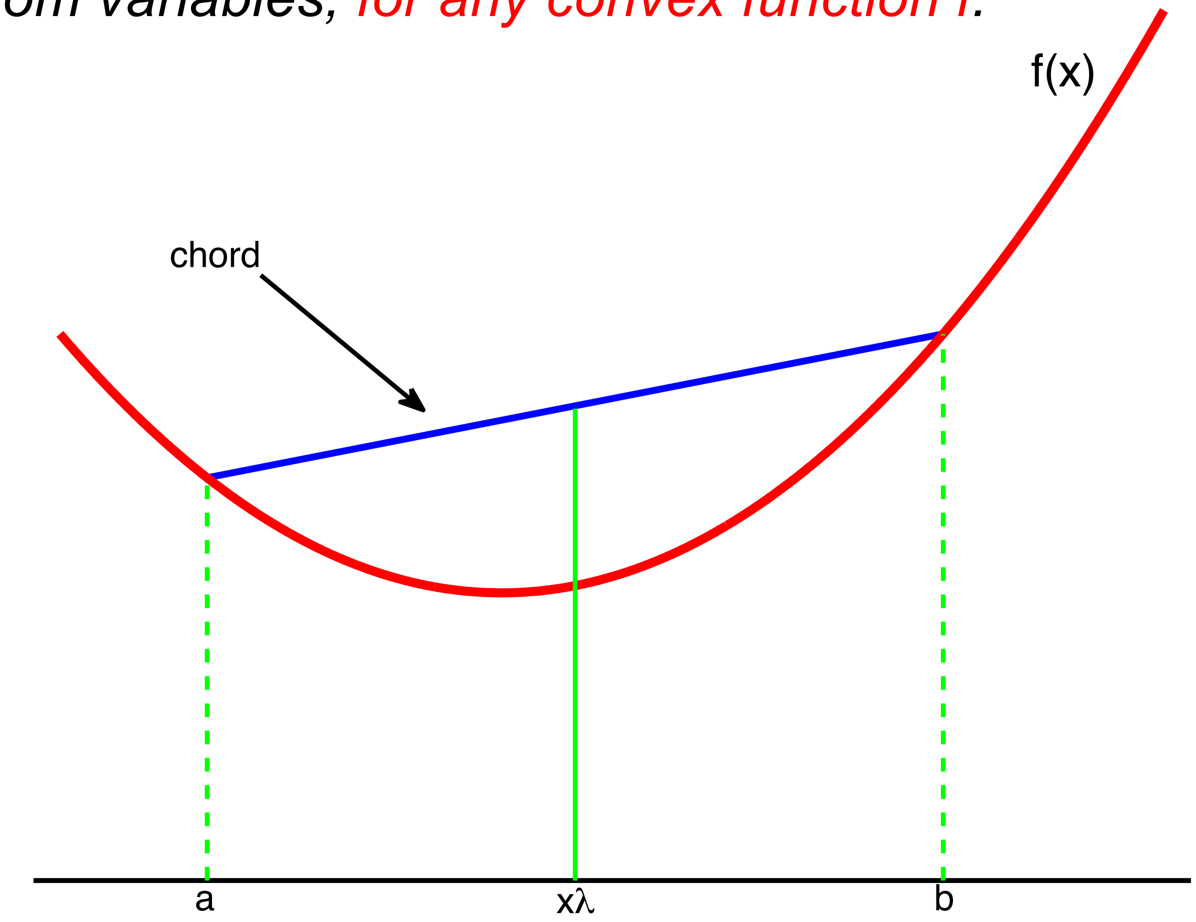
q(z) is *any* distribution with support over Z

(Jensen's Inequality) $$\geq \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right]$$

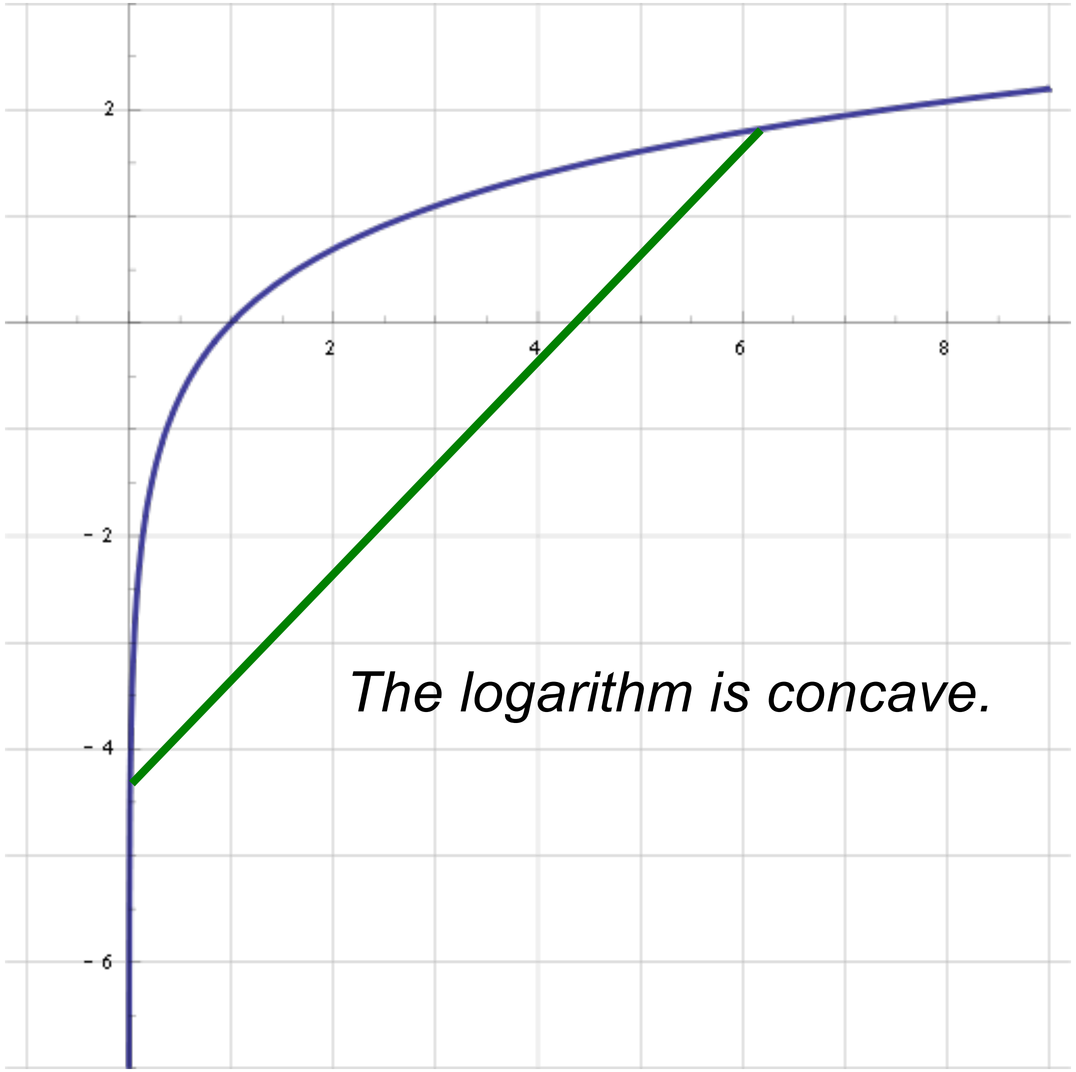
Jensen's Inequality

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

Valid for both discrete (expectations are sums) and continuous (expectations are integrals) random variables, for any convex function f.



$$\ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)]$$



Expectation Maximization

Find tightest lower bound of marginal likelihood,

$$\max_{\theta} \log p(\mathcal{Y} \mid \theta) \geq \max_{q,\theta} \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] \equiv \mathcal{L}(q, \theta)$$

Solve by coordinate ascent…

Initialize Parameters: $\theta^{(0)}$

At iteration t do:

Fix θ

Update q: $q^{(t)} = \arg\max_q \mathcal{L}(q, \theta^{(t-1)})$

Update θ: $\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$

Fix q

Until convergence

Expectation Maximization

Find tightest lower bound of marginal likelihood,

$$\max_{\theta} \log p(\mathcal{Y} \mid \theta) \geq \max_{q,\theta} \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] \equiv \mathcal{L}(q, \theta)$$

Solve by coordinate ascent…

Initialize Parameters: $\theta^{(0)}$

At iteration t do:

E-Step: $q^{(t)} = \arg\max_q \mathcal{L}(q, \theta^{(t-1)})$ (Fix θ)

M-Step: $\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$ (Fix q)

Until convergence

E-Step

$$q^{(t)}(z) = \arg\max_q \mathcal{L}(q, \theta^{(t-1)}) \equiv \mathbf{E}_q \left[\log \frac{p(z, y \mid \theta^{(t-1)})}{q(z)} \right]$$

Concave (in $q(z)$) and optimum occurs at,

$$q^{(t)}(z) = p(z \mid y, \theta^{(t-1)})$$

Set q(z) to posterior with current parameters

Initialize Parameters: $\theta^{(0)}$

At iteration t do:

E-Step: $q^{(t)}(z) = p(z \mid y, \theta^{(t-1)})$

M-Step: $\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$

Until convergence

M-Step

$$\theta^{(t)} = \arg\max_{\theta} \mathcal{L}(q^{(t)}, \theta) = \arg\max_{\theta} \mathbf{E}_{q^{(t)}} \left[\log \frac{p(z, y \mid \theta)}{q^{(t)}} \right]$$

Adding / subtracting constants we have,

$$\theta^{(t)} = \arg\max_{\theta} \sum_{z} q^{(t)}(z) \log p(z, y \mid \theta)$$

Intuition We don't know Z, so average log-likelihood over current posterior q(z), then maximize. E.g. weighted MLE.

May lack a closed-form, but suffices to take one or more gradient steps. Don't need to maximize, just improve.

Expectation Maximization

Initialize Parameters: $\theta^{(0)}$

At iteration t do:

E-Step: $q^{(t)}(z) = p(z \mid y, \theta^{(t-1)})$

M-Step: $\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$

Until convergence

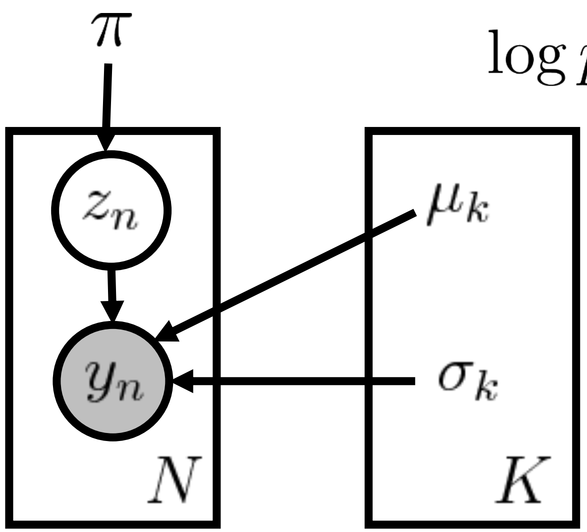
E-Step Compute **expected** log-likelihood under the posterior distribution,

$$q^{(t)}(z) = p(z \mid y, \theta^{(t-1)}) \qquad \mathbf{E}_{q^{(t)}}[\log p(z, y \mid \theta)] = \mathcal{L}(q^{(t)}, \theta)$$

M-Step Maximize expected log-likelihood,

$$\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$$

Example: Gaussian Mixture Model



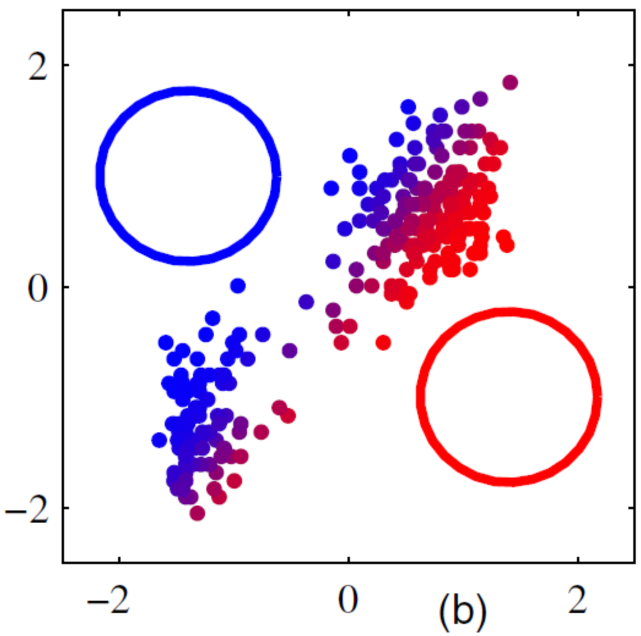
$$\log p(\mathcal{Y} \mid \pi, \mu, \Sigma) \geq \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \log \{\pi_k \mathcal{N}(y_n \mid \mu_k, \Sigma_k)\} = \mathcal{L}(q, \theta)$$

E-Step: $q^{\text{new}} = \arg\max_q \mathcal{L}(q, \theta^{\text{old}})$

$$q^{\text{new}}(z_n = k) = p(z_n = k \mid \mathcal{Y}, \mu^{\text{old}}, \Sigma^{\text{old}}, \pi^{\text{old}})$$

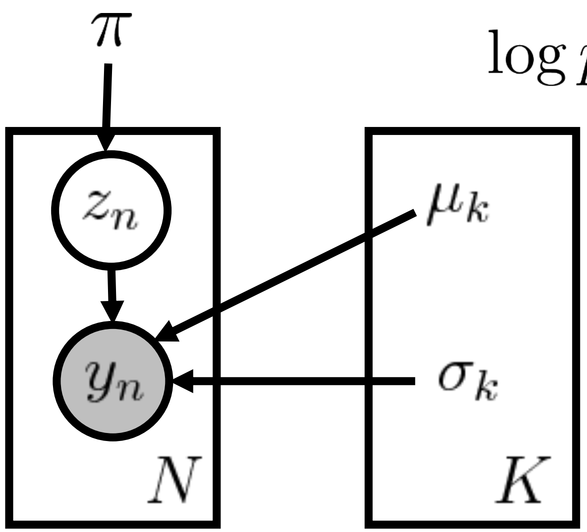
$$= \frac{p(z_n = k, \mathcal{Y} \mid \mu^{\text{old}}, \Sigma^{\text{old}}, \pi^{\text{old}})}{\sum_{j=1}^{K} p(z_n = j, \mathcal{Y} \mid \mu^{\text{old}}, \Sigma^{\text{old}}, \pi^{\text{old}})}$$

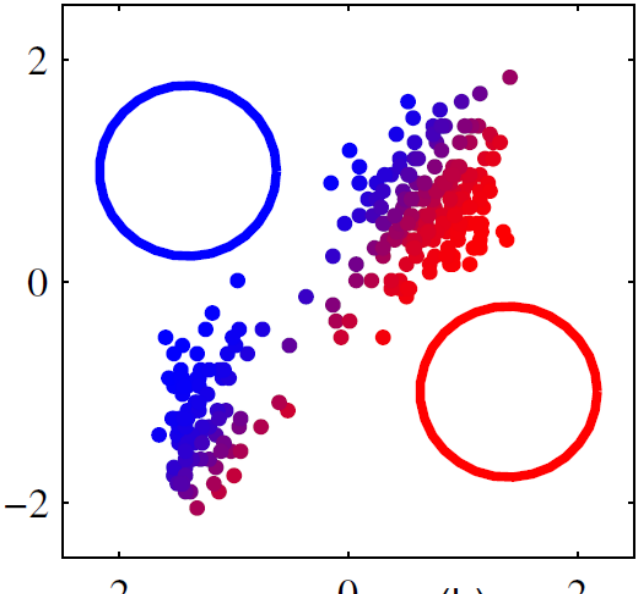
$$= \frac{\pi_k^{\text{old}} \mathcal{N}(y_n \mid \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_n \mid \mu_j^{\text{old}}, \Sigma_j^{\text{old}})}$$



Commonly refer to $q(z_n)$ as *responsibility*

Example: Gaussian Mixture Model





(b)

$$\log p(\mathcal{Y} \mid \pi, \mu, \Sigma) \geq \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \log \{\pi_k \mathcal{N}(y_n \mid \mu_k, \Sigma_k)\} = \mathcal{L}(q, \theta)$$

M-Step: $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q^{\text{new}}, \theta)$

Start with mean parameter μ_k,

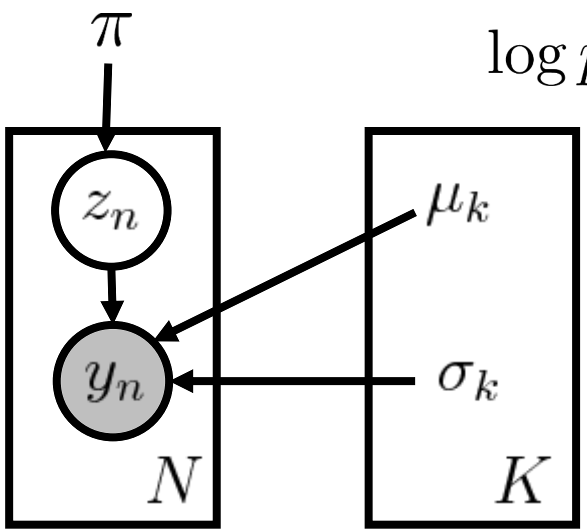
$$0 = \nabla_{\mu_k} \mathcal{L}(q^{\text{new}}, \theta)$$

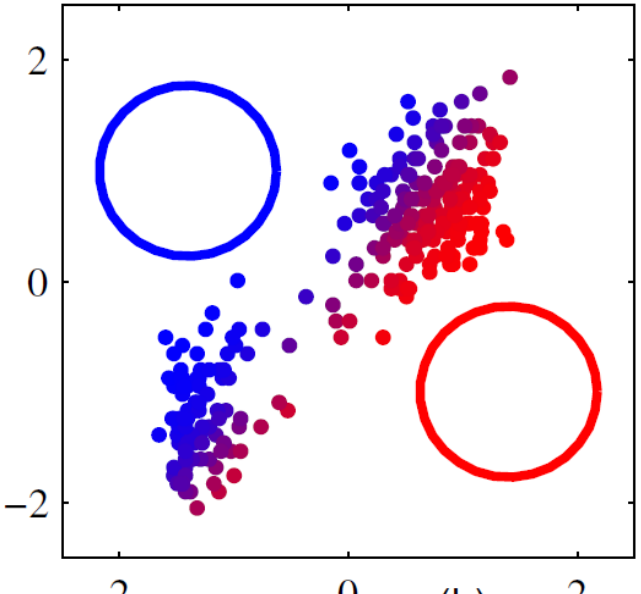
$$0 = \sum_{n=1}^{N} \nabla_{\mu_k} \mathbf{E}_{z_n \sim q^{\text{new}}} [\log \mathcal{N}(y_n \mid \mu_{z_n}, \Sigma_{z_n})]$$

$$0 = -\sum_{n=1}^{N} q^{\text{new}}(z_n = k) \Sigma_k (y_n - \mu_k)$$

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} q^{\text{new}}(z_n = k) y_n \quad \text{where} \quad N_k = \sum_{n=1}^{N} q(z_n = k)$$

Example: Gaussian Mixture Model





(b)

$$\log p(\mathcal{Y} \mid \pi, \mu, \Sigma) \geq \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \log \{\pi_k \mathcal{N}(y_n \mid \mu_k, \Sigma_k)\} = \mathcal{L}(q, \theta)$$

M-Step: $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q^{\text{new}}, \theta)$

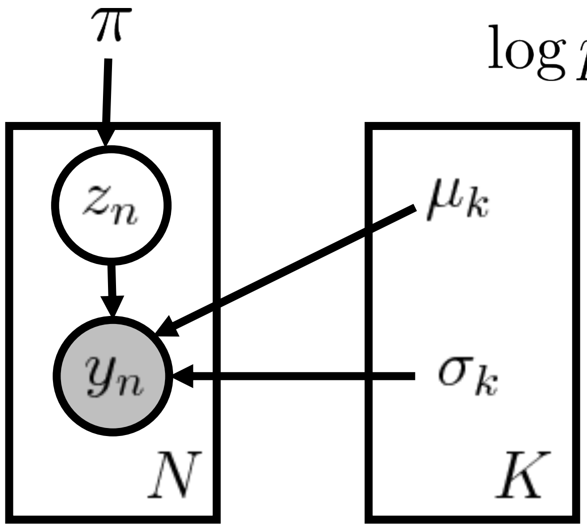
Repeat for remaining parameters,

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} q(z_n = k)(y_n - \mu_k^{\text{new}})(y_n - \mu_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

- Solving for mixture weights requires a bit more work
- Need constraint $\sum_k \pi_k = 1$
- Use Lagrange multiplier approach

Example: Gaussian Mixture Model



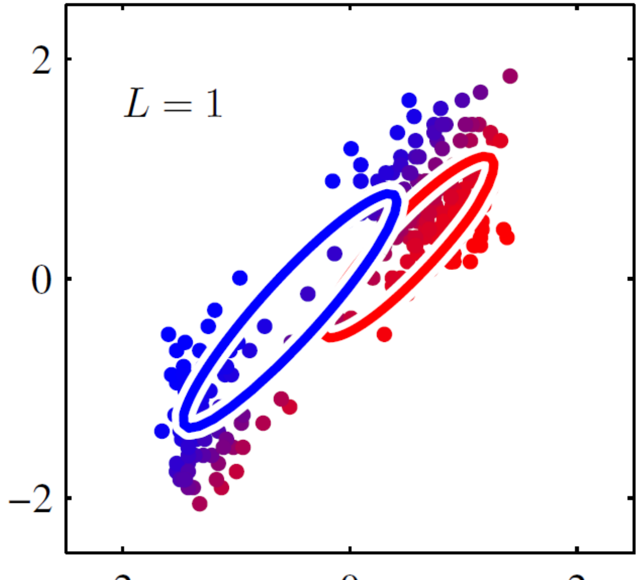
$$\log p(\mathcal{Y} \mid \pi, \mu, \Sigma) \geq \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \log \{\pi_k \mathcal{N}(y_n \mid \mu_k, \Sigma_k)\} = \mathcal{L}(q, \theta)$$

M-Step: $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q^{\text{new}}, \theta)$

Repeat for remaining parameters,

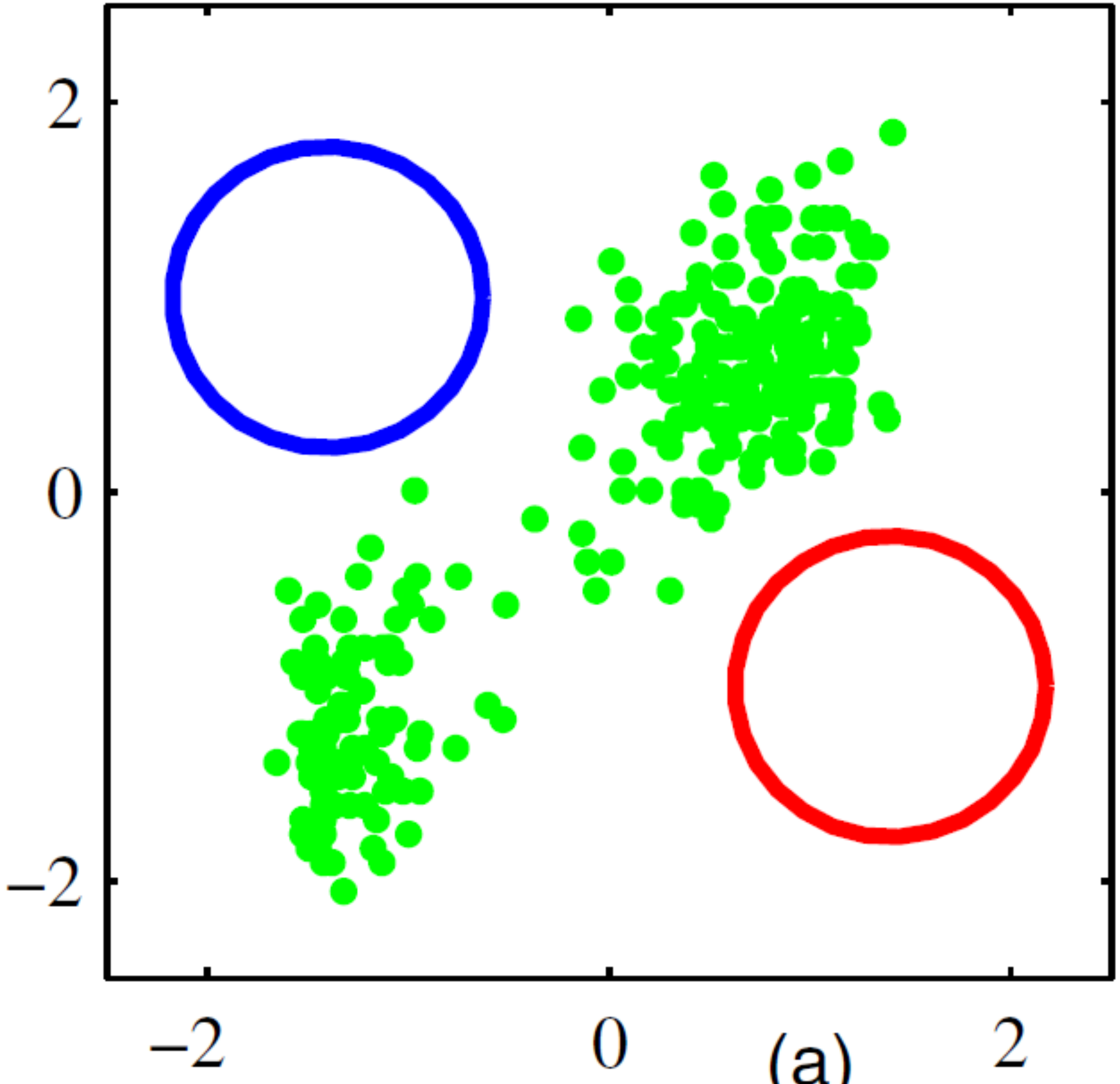
$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} q(z_n = k)(y_n - \mu_k^{\text{new}})(y_n - \mu_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$



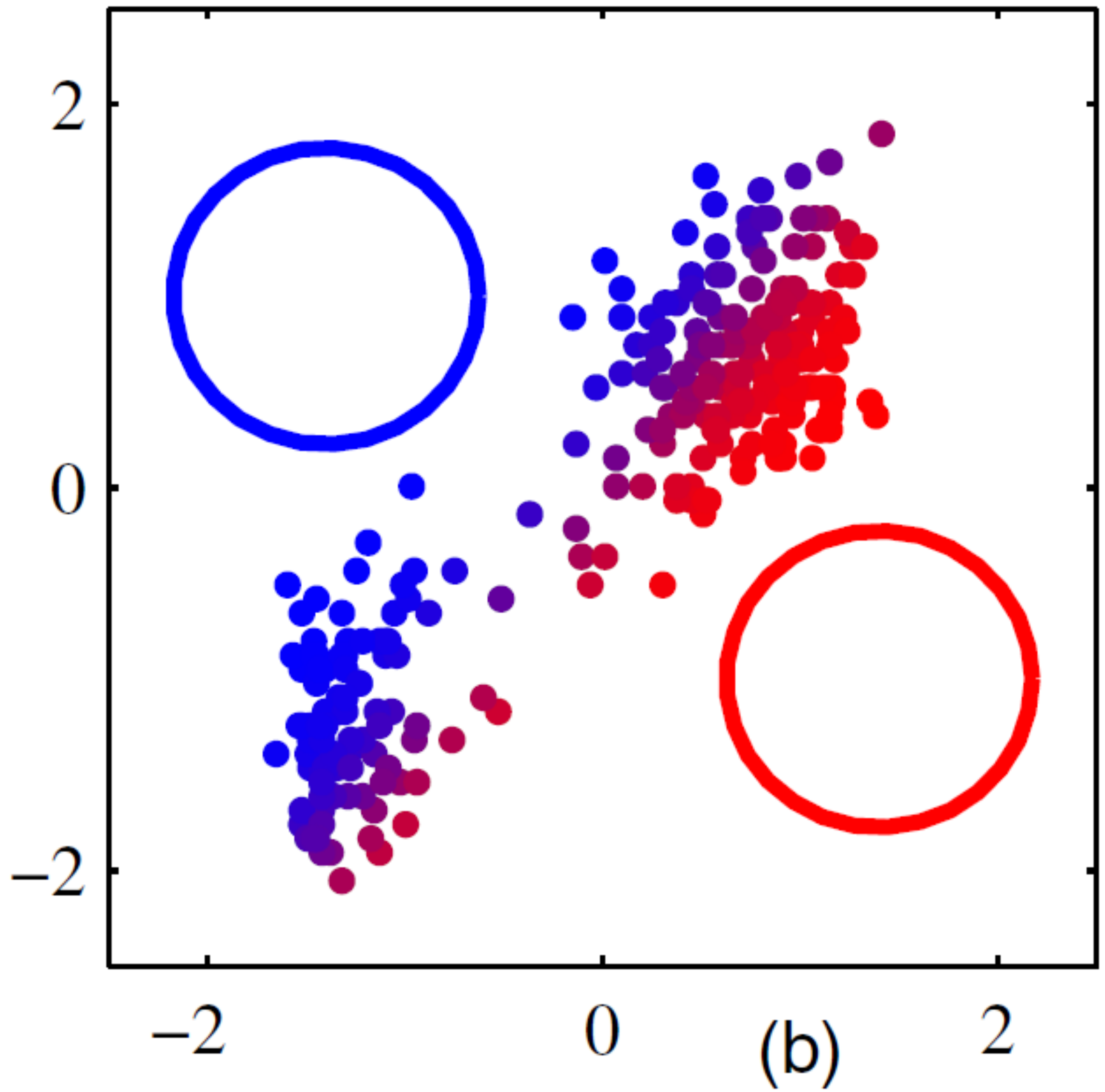
(c)

- Solving for mixture weights requires a bit more work
- Need constraint $\sum_k \pi_k = 1$
- Use Lagrange multiplier approach

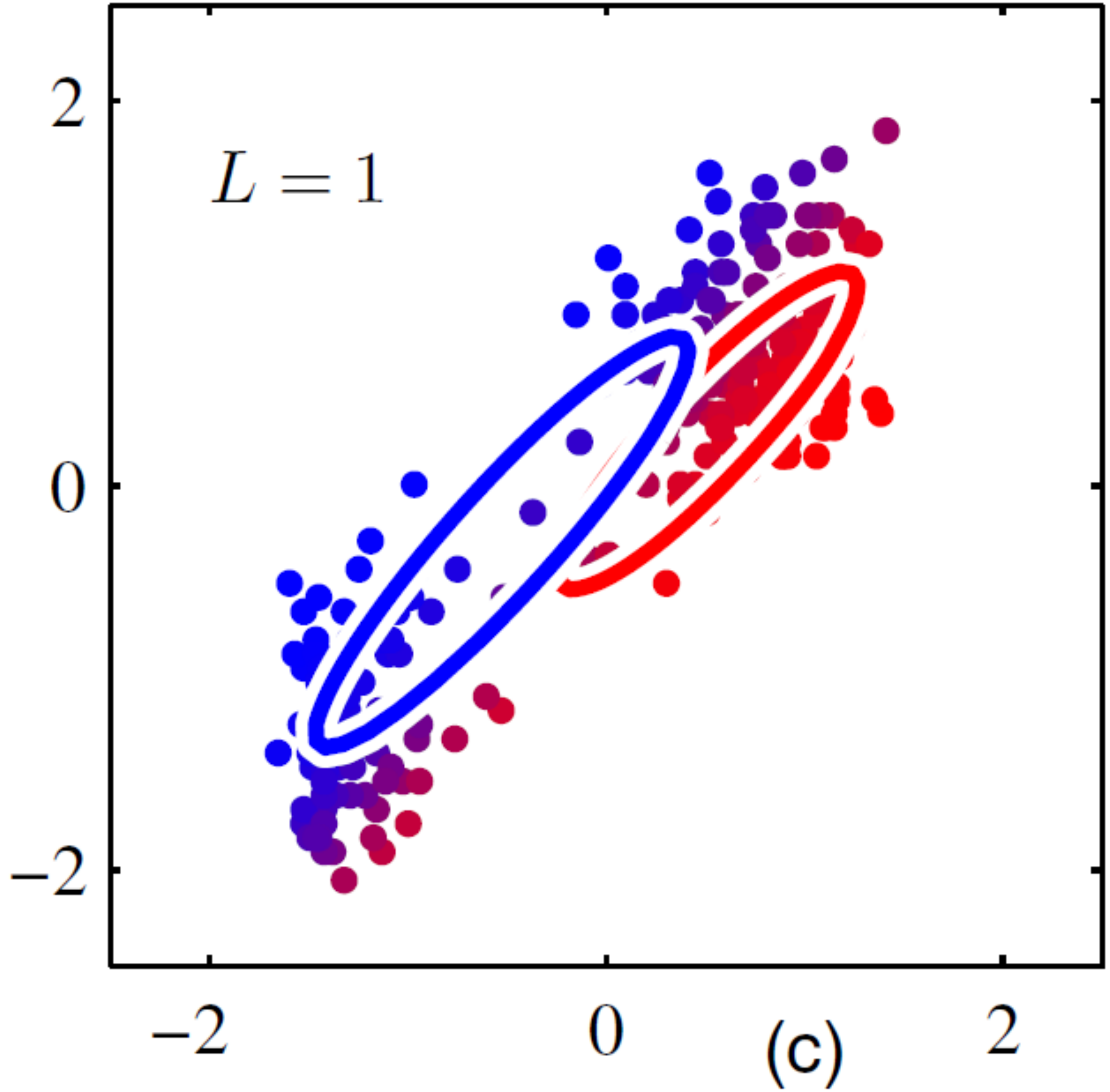


(a)

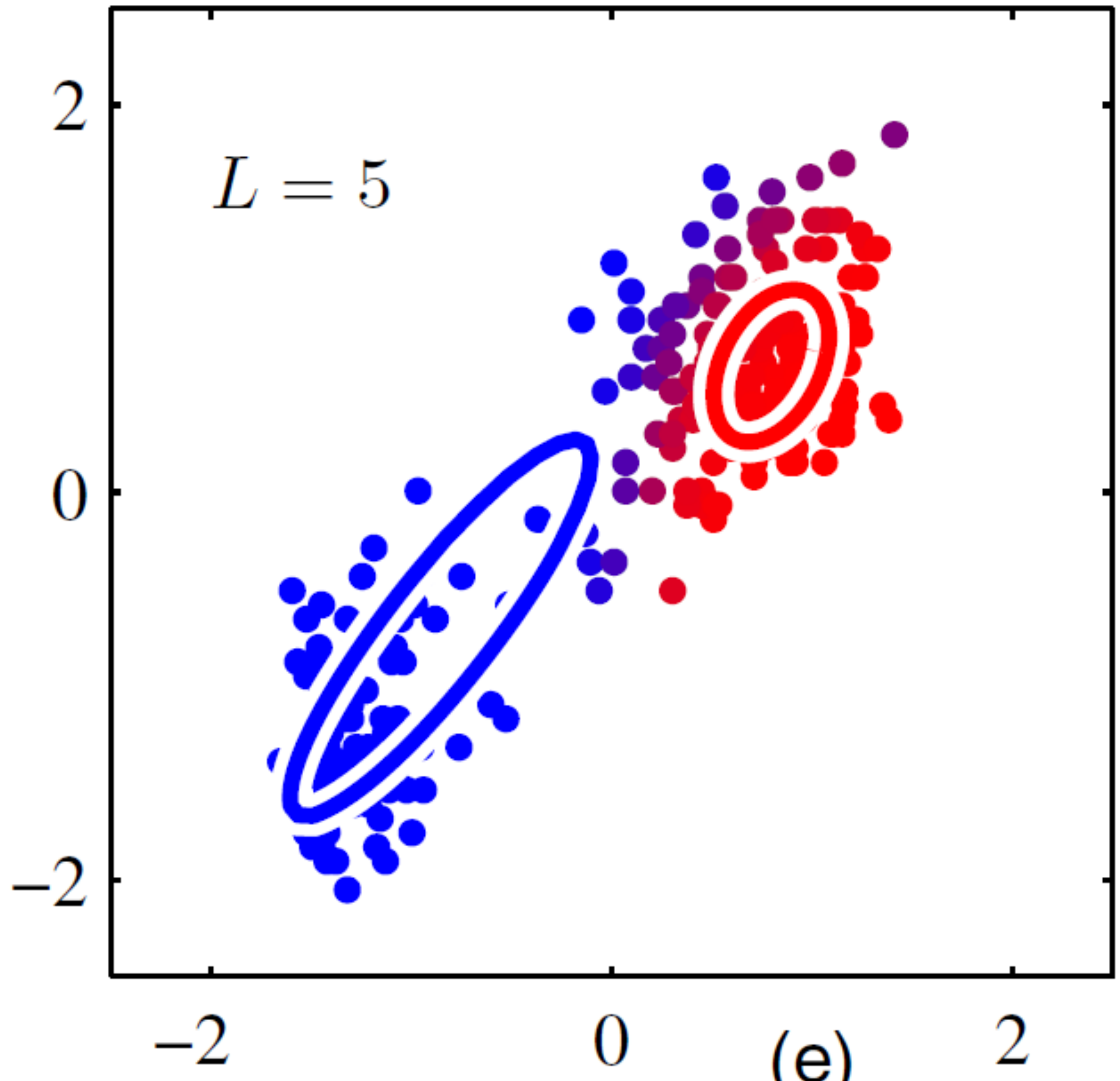
Source: Chris Bishop, PRML



Source: Chris Bishop, PRML

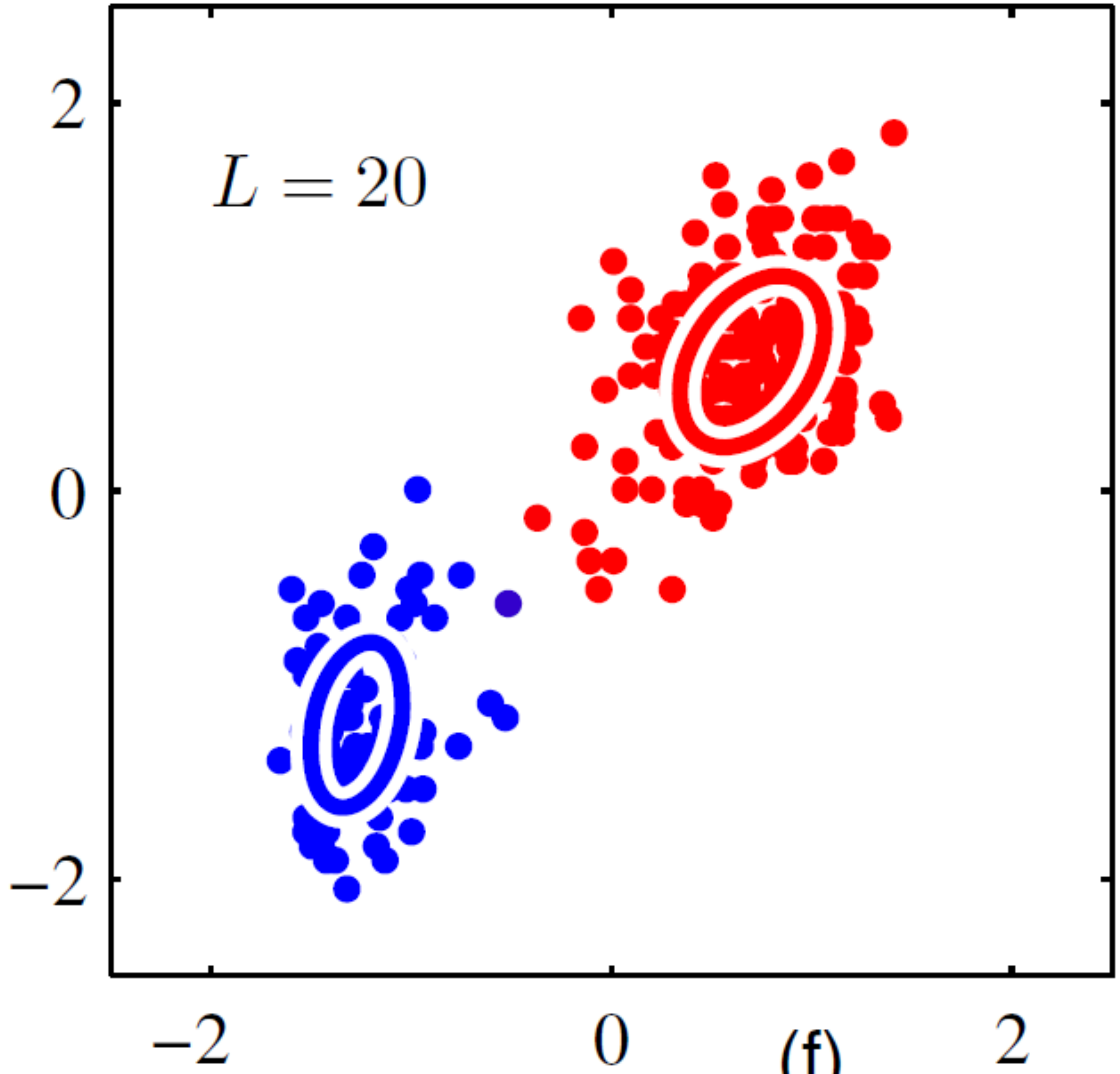


Source: Chris Bishop, PRML



(e)

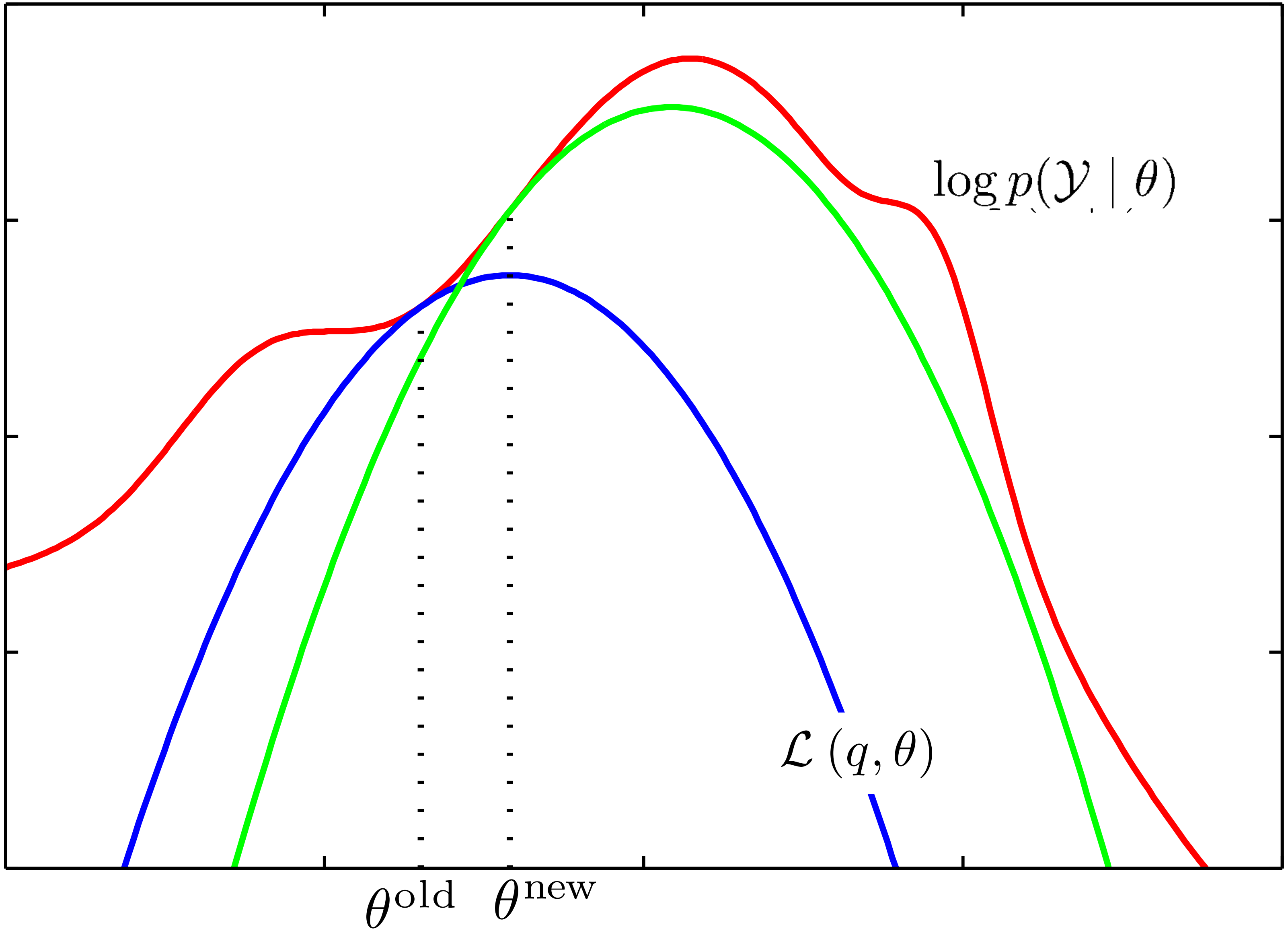
Source: Chris Bishop, PRML



(f)

Source: Chris Bishop, PRML

EM: A Sequence of Lower Bounds



C. Bishop, Pattern Recognition & Machine Learning

EM Lower Bound

$$\mathbf{E}_q\left[\log \frac{p(z,y \mid \theta)}{q(z)}\right] = \mathbf{E}_q\left[\log \frac{p(z,y \mid \theta)}{q(z)} \frac{p(y \mid \theta)}{p(y \mid \theta)}\right]$$ **(Multiply by 1)**

$$= \log p(y \mid \theta) - \mathrm{KL}(q(z) \| p(z \mid y, \theta))$$ **(Definition of KL)**

Bound gap is the Kullback-Leibler divergence KL(q||p),

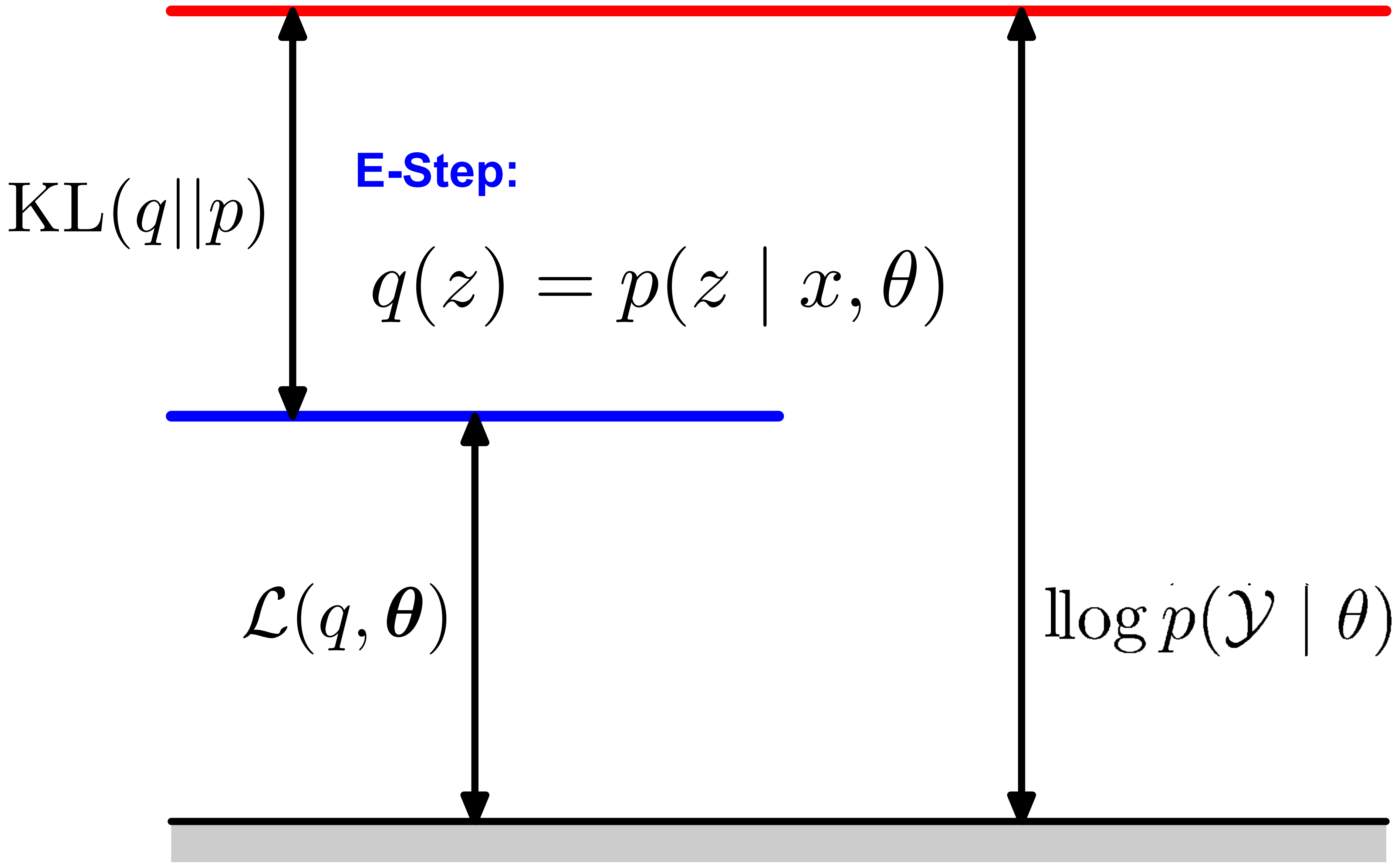
$$\mathrm{KL}(q(z) \| p(z \mid y, \theta)) = \sum_z q(z) \log \frac{q(z)}{p(z \mid y, \theta)}$$

- Similar to a “distance” between q and p

$$\mathrm{KL}(q \mid\mid p) \geq 0 \text{ and } \mathrm{KL}(q \mid\mid p) = 0 \text{ if and only if } q = p$$

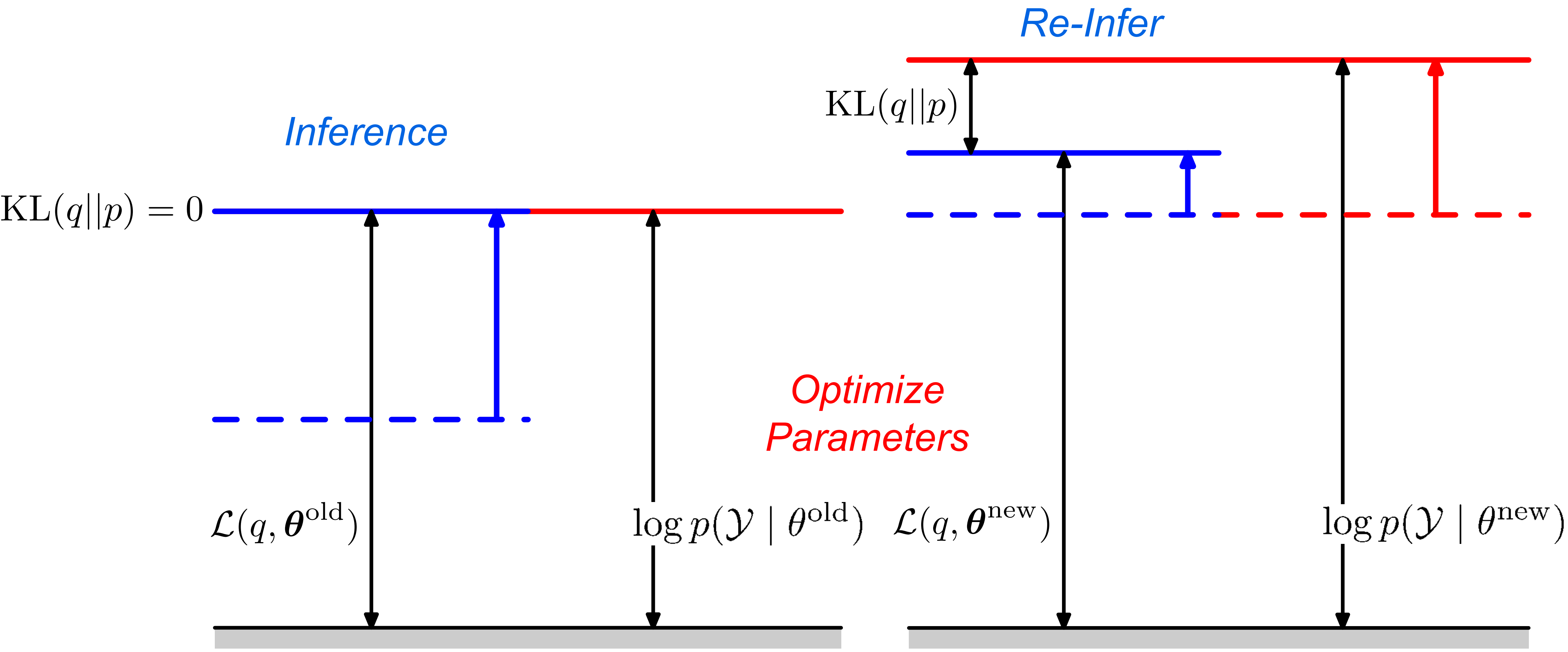
- This is why solution to E-step is $q(z) = p(z \mid y, \theta)$

Lower Bounds on Marginal Likelihood



C. Bishop, Pattern Recognition & Machine Learning

Expectation Maximization Algorithm



E Step: *Optimize distribution on hidden variables given parameters*

M Step: *Optimize parameters given distribution on hidden variables*

Properties of Expectation Maximization Algorithm

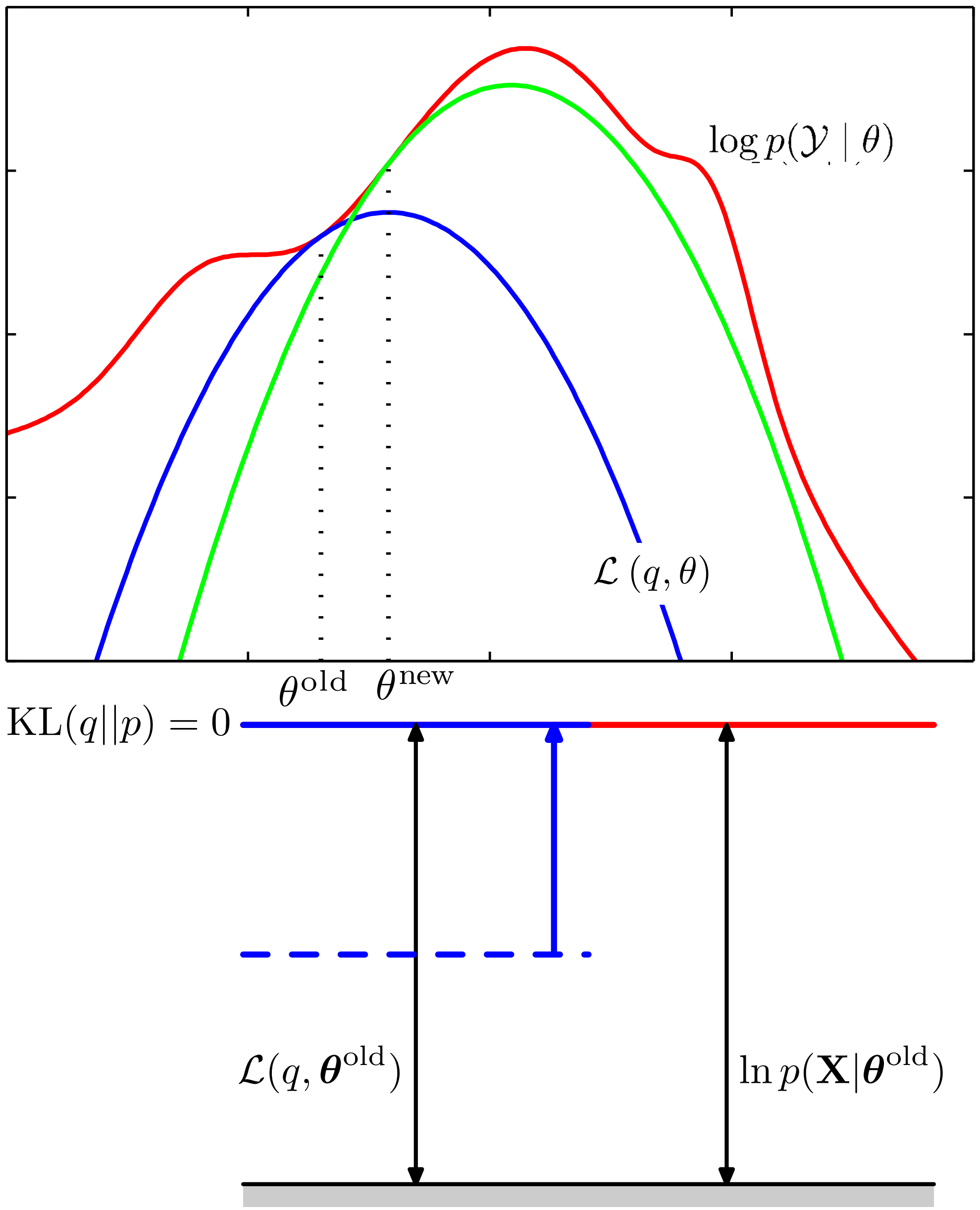
Sequence of bounds is monotonic,

$$\mathcal{L}(q^{(1)}, \theta^{(1)}) \leq \mathcal{L}(q^{(2)}, \theta^{(2)}) \leq \ldots \leq \mathcal{L}(q^{(T)}, \theta^{(T)})$$

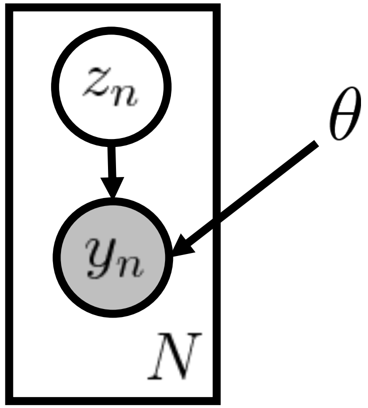
Guaranteed to converge
(**Pf.** Monotonic sequence bounded above.)

Converges to a local maximum of the marginal likelihood

After each E-step bound is tight at θ^{old} so likelihood calculation is exact (for those parameters)



MLE vs. MAP Estimation

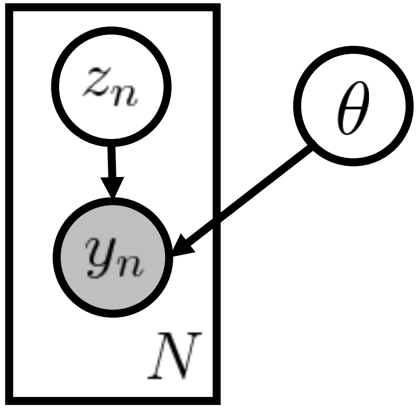


Conditional model,

$$p(z, y \mid \theta) = \prod_{n=1}^{N} p(z_n) p(y_n \mid z_n, \theta)$$

MLE estimate of unknown non-random parameters,

$$\theta^{\mathrm{MLE}} = \arg\max_{\theta} \log p(\mathcal{Y} \mid \theta)$$



Generative model,

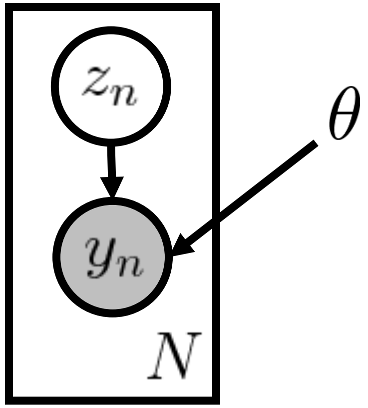
$$p(z, y, \theta) = p(\theta) \prod_{n=1}^{N} p(z_n) p(y_n \mid z_n, \theta)$$

MAP estimate of random parameters,

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta} \log p(\theta) + \log p(\mathcal{Y} \mid \theta)$$

EM Lower Bound

Recall EM lower bound of marginal likelihood



$$\arg\max_{\theta} \log p(\mathcal{Y} \mid \theta) = \arg\max_{\theta} \log \sum_{z} p(z, \mathcal{Y} \mid \theta)$$

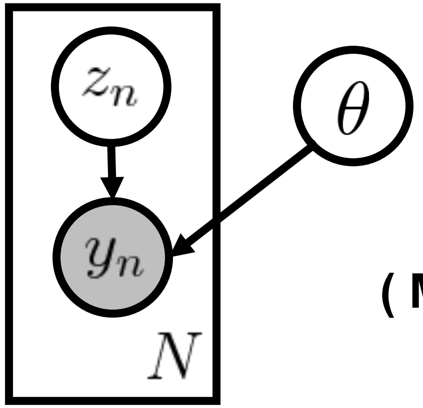
(Multiply by q(z)/q(z)=1)
$$= \log \sum_{z} p(z, \mathcal{Y} \mid \theta) \left(\frac{q(z)}{q(z)} \right)$$

(Definition of Expected Value)
$$= \log \mathbf{E}_q \left[\frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right]$$

(Jensen's Inequality)
$$\geq \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right]$$

MAP EM

Bound holds with addition of log-prior



$$\arg\max_\theta \log p(\theta \mid \mathcal{Y}) = \arg\max_\theta \log \sum_z p(z, \mathcal{Y} \mid \theta) + \log p(\theta)$$

(Multiply by q(z)/q(z)=1) $$= \log \sum_z p(z, \mathcal{Y} \mid \theta) \left(\frac{q(z)}{q(z)} \right) + \log p(\theta)$$

(Definition of Expected Value) $$= \log \mathbf{E}_q \left[\frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] + \log p(\theta)$$

(Jensen's Inequality) $$\geq \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] + \log p(\theta)$$

MAP EM

$$\max_{\theta} \log p(\theta, \mathcal{Y}) \geq \max_{q,\theta} \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] + \log p(\theta)$$

E-Step: Fix parameters and maximize w.r.t. q(z),

$$q^{\text{new}} = \arg\max_{q} \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta^{\text{old}})}{q(z)} \right] + \boxed{\log p(\theta^{\text{old}})}$$

Constant in q(z)

Same solution as standard maximum likelihood EM,

$$q^{\text{new}} = p(z \mid \mathcal{Y}, \theta^{\text{old}})$$

M-Step: Fix q(z) and optimize parameters,

$$\theta^{\text{new}} = \arg\max_{\theta} \mathbf{E}_{q^{\text{new}}} \left[\log p(z, \mathcal{Y} \mid \theta) \right] + \log p(\theta)$$

MAP EM

Initialize Parameters: $\theta^{(0)}$

At iteration t do:

E-Step: $q^{(t)}(z) = p(z \mid y, \theta^{(t-1)})$

M-Step: $\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta) + \log p(\theta)$

Until convergence

E-Step Compute **expected** log-likelihood under the posterior distribution,

$$q^{(t)}(z) = p(z \mid y, \theta^{(t-1)}) \qquad \mathbf{E}_{q^{(t)}}[\log p(z, y \mid \theta)] = \mathcal{L}(q^{(t)}, \theta)$$

M-Step Maximize expected log-likelihood,

$$\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta) + \log p(\theta)$$

Learning Summary

Maximum likelihood estimation (MLE) maximizes (log-)likelihood func,

$$\theta^{\mathrm{MLE}} = \arg\max_{\theta} \log p(\mathcal{Y} \mid \theta) \equiv \mathcal{L}(\theta)$$

Where parameters are *unknown non-random* quantities

Maximum a posteriori (MAP) maximizes posterior probability,

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta} \log p(\theta \mid \mathcal{Y}) = \arg\max_{\theta} \mathcal{L}(\theta) + \log p(\theta)$$

Parameters are *random* quantities with prior $p(\theta)$.

Learning Summary

- Most models will not yield closed-form MLE/MAP estimates
- Gradient-based methods optimize log-likelihood function

$$\theta^{k+1} = \theta^k + \beta \nabla_\theta \mathcal{L}(\theta^k)$$

- Expectation Maximization (EM) alternative to gradient methods
- Both approaches approximate for non-convex models

EM Summary

Approximate MLE for intractable marginal likelihood via lower bound,

$$\max_{\theta} \log p(\mathcal{Y} \mid \theta) \geq \max_{q,\theta} \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] \equiv \mathcal{L}(q, \theta)$$

Coordinate ascent alternately maximizes $q(z)$ and θ,

E-Step

$$q^{\text{new}} = \arg\max_{q} \mathcal{L}(q, \theta^{\text{old}})$$

M-Step

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q^{\text{new}}, \theta)$$

Solution to E-step sets q to posterior over hidden variables,

$$q^{\text{new}}(z) = p(z \mid \mathcal{Y}, \theta^{\text{old}})$$

M-step is problem-dependent, requires gradient calculation

EM Summary

Easily extends to (approximate) MAP estimation,

$$\max_{\theta} \log p(\theta \mid \mathcal{Y}) \geq \max_{q,\theta} \mathbf{E}_q \left[\log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] + \log p(\theta) + \text{const.}$$

E-step unchanged / Slightly modifies M-step,

E-Step

$$q^{\text{new}} = \arg\max_{q} \mathcal{L}(q, \theta^{\text{old}}) = p(z \mid \mathcal{Y}, \theta^{\text{old}})$$

M-Step

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q^{\text{new}}, \theta) + \log p(\theta)$$

Properties of both MLE / MAP EM

- Monotonic in $\mathcal{L}(q, \theta)$ or $\mathcal{L}(q, \theta) + \log p(\theta)$ (for MAP)
- Provably converge to local optima (hence approximate estimation)