CSC535: Probabilistic Graphical Models

Parameter Learning and Expectation Maximization

Prof. Jason Pacheco
Parameter Estimation

We have a model in the form of a probability distribution, with unknown **parameters of interest** $\theta$,

$$p(X; \theta)$$

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

$$\{x_i\}_{i}^{N} \overset{iid}{\sim} p(\cdot; \theta)$$

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 students at UA, and we model them as Gaussian:

\[ \{x_i\}_{i=1}^N \sim \mathcal{N}(\mu, \sigma^2) \]

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…

**Likelihood Principle** Given a statistical model, the likelihood function describes all evidence of a parameter that is contained in the data.
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 $\theta$ -- the data are fixed
- Measure of how well parameter $\theta$ describes data (goodness of fit)

How could we use this to estimate a parameter $\theta$?
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 \( \mathcal{L}_N(\theta) \) convex?

<table>
<thead>
<tr>
<th>Yes</th>
<th>Unique, closed-form solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Gradient-based optimization</td>
</tr>
</tbody>
</table>

Approach

- Compute derivative \( \frac{d\mathcal{L}_N}{d\theta} \)
- Set to zero and solve \( \frac{d\mathcal{L}_N}{d\theta} = 0 \Rightarrow \hat{\theta}^{\text{MLE}} \)

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)

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

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 $\theta$.

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

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.
Marginal Likelihood Calculation

Recall the Gaussian Mixture Model…

\[ \theta = \{\mu_1, \sigma_1, \ldots, \mu_K, \sigma_K\} \]

Marginal Likelihood (likelihood function):

\[ p(\mathcal{Y} | \theta) = \sum_{z_1} \ldots \sum_{z_N} p(z_1, \ldots, z_N, \mathcal{Y} | \theta) \]

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

**Motivation** Approximate MLE / MAP when we cannot compute the marginal likelihood in closed-form
Lower Bounding Marginal Likelihood

Conditionally-independent model with partial observations...

\[ \log p(\mathcal{Y} | \theta) = \log \sum_{z_1} \ldots \sum_{z_N} p(z_1, \ldots, z_N, \mathcal{Y} | \theta) \]

( Multiply by \( q(z)/q(z)=1 \) )

\[ = \log \sum_z p(z, \mathcal{Y} | \theta) \left( \frac{q(z)}{q(z)} \right) \]

( Definition of Expected Value )

\[ = \log \mathbb{E}_q \left[ \frac{p(z, \mathcal{Y} | \theta)}{q(z)} \right] \]

( Jensen's Inequality )

\[ \geq \mathbb{E}_q \left[ \log \frac{p(z, \mathcal{Y} | \theta)}{q(z)} \right] \]

Shorthand
\[ z = z_1, \ldots, z_N \]

q(z) is any distribution with support over Z
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 \( f \).

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

The logarithm is concave.
Expectation Maximization

Find tightest lower bound of marginal likelihood,

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

Solve by coordinate ascent...

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

At iteration \( t \) do:

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

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

Until convergence
Expectation Maximization

Find tightest lower bound of marginal likelihood,

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

Solve by coordinate ascent…

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

At iteration $t$ do:

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

Until convergence

Fix $\theta$ and $q$
E-Step

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

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

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

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

Set \( q(z) \) to posterior with current parameters
\[ \theta^{(t)} = \arg \max_{\theta} \mathcal{L}(q^{(t)}, \theta) = \arg \max_{\theta} \mathbb{E}_{q^{(t)}} \left[ \log \frac{p(z, y | \theta)}{q^{(t)}} \right] \]

Adding / subtracting constants we have,

\[ \theta^{(t)} = \arg \max_{\theta} \sum_{z} q^{(t)}(z) \log p(z, y | \theta) \]

**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:

\[E\text{-Step}: \quad q^{(t)}(z) = p(z | y, \theta^{(t-1)})\]

\[M\text{-Step}: \quad \theta^{(t)} = \arg\max_{\theta} \mathcal{L}(q^{(t)}, \theta)\]

Until convergence

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

\[ q^{(t)}(z) = p(z | y, \theta^{(t-1)}) \quad \mathbb{E}_{q^{(t)}}[\log p(z, y | \theta)] = \mathcal{L}(q^{(t)}, \theta) \]

**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:** \(q^{\text{new}} = \arg \max_q \mathcal{L}(q, \theta^{\text{old}})\)

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

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

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

Commonly refer to \(q(z_n)\) as *responsibility*
Example: Gaussian Mixture Model

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

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

Start with mean parameter \(\mu_k\),

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

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

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

\[
\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} q^{\text{new}}(z_n = k) y_n \quad \text{where} \quad N_k = \sum_{n=1}^{N} q(z_n = k)
\]
Example: Gaussian Mixture Model

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

**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

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

**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
$L = 5$
$L = 20$
EM: A Sequence of Lower Bounds

\[ \log p(Y \mid \theta) \]

\[ \mathcal{L}(q, \theta) \]

\[ \theta^{\text{old}} \quad \theta^{\text{new}} \]
EM Lower Bound

\[ E_q \left[ \log \frac{p(z, y | \theta)}{q(z)} \right] = E_q \left[ \log \frac{p(z, y | \theta)}{q(z)} \frac{p(y | \theta)}{p(y | \theta)} \right] \]

\[ = \log p(y | \theta) - KL(q(z) || p(z | y, \theta)) \]

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

\[ KL(q(z) || p(z | y, \theta)) = \sum_z q(z) \log \frac{q(z)}{p(z | y, \theta)} \]

- Similar to a “distance” between q and p

\[ KL(q || p) \geq 0 \text{ and } KL(q || p) = 0 \text{ if and only if } q = p \]

- This is why solution to E-step is \( q(z) = p(z | y, \theta) \)
Lower Bounds on Marginal Likelihood

E-Step:

\[ q(z) = p(z \mid x, \theta) \]

\[ \text{KL}(q \| p) \]

\[ \mathcal{L}(q, \theta) \]

\[ \ln p(\mathcal{Y} \mid \theta) \]

C. Bishop, Pattern Recognition & Machine Learning
**Expectation Maximization Algorithm**

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

\[ KL(q||p) = 0 \]

\[ \mathcal{L}(q, \theta^{\text{old}}) \]

\[ \log p(\mathcal{Y} | \theta^{\text{old}}) \]

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

\[ \mathcal{L}(q, \theta^{\text{new}}) \]

\[ \log p(\mathcal{Y} | \theta^{\text{new}}) \]
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 \( \theta^{old} \) 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,

\[
\theta^{\text{MLE}} = \arg \max_{\theta} \log p(\mathcal{Y} \mid \theta)
\]

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,

\[
\theta^{\text{MAP}} = \arg \max_{\theta} \log p(\theta) + \log p(\mathcal{Y} \mid \theta)
\]
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 )

\[
= \log \mathbb{E}_q \left[ \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right]
\]

( Jensen's Inequality )

\[
\geq \mathbb{E}_q \left[ \log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right]
\]
Bound holds with addition of log-prior

\[
\begin{align*}
\arg\max_{\theta} \log p(\theta \mid \mathcal{Y}) &= \arg\max_{\theta} \log \sum_z p(z, \mathcal{Y} \mid \theta) + \log p(\theta) \\
&= \log \sum_z p(z, \mathcal{Y} \mid \theta) \left( \frac{q(z)}{q(z)} \right) + \log p(\theta) \\
&= \log \mathbb{E}_q \left[ \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] + \log p(\theta) \\
&\geq \mathbb{E}_q \left[ \log \frac{p(z, \mathcal{Y} \mid \theta)}{q(z)} \right] + \log p(\theta)
\end{align*}
\]
MAP EM

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

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

\[
q^{\text{new}} = \arg \max_q \mathbb{E}_q \left[ \log \frac{p(z, \mathcal{Y} | \theta^{\text{old}})}{q(z)} \right] + \log p(\theta^{\text{old}}) \quad \text{(Constant in } q(z)\text{)}
\]

Same solution as standard maximum likelihood EM,

\[
q^{\text{new}} = p(z | \mathcal{Y}, \theta^{\text{old}})
\]

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

\[
\theta^{\text{new}} = \arg \max_{\theta} \mathbb{E}_{q^{\text{new}}} \left[ \log p(z, \mathcal{Y} | \theta) \right] + \log p(\theta)
\]
Initialize Parameters: $\theta^{(0)}$

At iteration $t$ do:

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

**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,

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

**M-Step Maximize** expected log-likelihood,

$\theta^{(t)} = \arg \max_{\theta} \mathcal{L}(q^{(t)}, \theta) + \log p(\theta)$
Maximum likelihood estimation (MLE) maximizes (log-)likelihood func,
\[ \theta^\text{MLE} = \arg \max_{\theta} \log p(\mathcal{Y} \mid \theta) \equiv \mathcal{L}(\theta) \]
Where parameters are unknown non-random quantities

Maximum a posteriori (MAP) maximizes posterior probability,
\[ \theta^\text{MAP} = \arg \max_{\theta} \log p(\theta \mid \mathcal{Y}) = \arg \max_{\theta} \mathcal{L}(\theta) + \log p(\theta) \]
Parameters are random quantities with prior \( p(\theta) \).
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
Approximate MLE for intractable marginal likelihood via lower bound,

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

Coordinate ascent alternately maximizes $q(z)$ and $\theta$,

**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,

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

M-step is problem-dependent, requires gradient calculation
Easily extends to (approximate) MAP estimation,

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

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)