Expectation Maximization

CMSC 691
UMBC
Outline

EM (Expectation Maximization)

Basic idea

Three coins example

Why EM works
Expectation Maximization (EM)

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty (compute expectations)

2. M-step: maximize log-likelihood, assuming these uncertain counts
Expectation Maximization (EM): E-step

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

\[ p(z_i) \] \[ \Rightarrow \] \[ \text{count}(z_i, w_i) \]

2. M-step: maximize log-likelihood, assuming these uncertain counts
Expectation Maximization (EM): E-step

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

\[ p(z_i) \quad \rightarrow \quad \text{count}(z_i, w_i) \]

2. M-step: maximize log-likelihood, assuming these uncertain counts

We’ve already seen this type of counting, when computing the gradient in maxent models.
Expectation Maximization (EM): M-step

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

2. M-step: maximize log-likelihood, assuming these uncertain counts

\[ p^{(t)}(z) \quad \begin{array}{c} \text{estimated} \\ \text{counts} \end{array} \quad \rightarrow \quad p^{(t+1)}(z) \]
$\max \mathbb{E} \frac{z \sim p(\theta(t))}{\log p(\theta(z, w))}$

the average log-likelihood of our complete data $(z, w)$, averaged across all $z$ and according to how likely our current model thinks $z$ is
maximize the average log-likelihood of our complete data \((z, w)\), averaged across all \(z\) and according to how likely our \textit{current} model thinks \(z\) is

\[
\max_{\theta} \mathbb{E}_{z \sim p_{\theta}(t) (\cdot | w)} \left[ \log p_{\theta} (z, w) \right]
\]
maximize the average log-likelihood of our complete data $(z, w)$, averaged across all $z$ and according to how likely our current model thinks $z$ is

$$\max_{\theta} \mathbb{E}_{z \sim p_{\theta}(t)(\cdot | w)} [\log p_{\theta}(z, w)]$$
maximize the average log-likelihood of our complete data \((z, w)\), averaged across all \(z\) and according to how likely our \textit{current} model thinks \(z\) is.

\[
\max_{\theta} \mathbb{E}_{z \sim p_{\theta}(t)\cdot|w)} \left[ \log p_{\theta}(z, w) \right]
\]
maximize the average log-likelihood of our complete data \((z, w)\), averaged across all \(z\) and according to how likely our *current* model thinks \(z\) is
EM Math

maximize the average log-likelihood of our complete data \((z, w)\), averaged across all \(z\) and according to how likely our current model thinks \(z\) is

\[
\max_{\theta} \mathbb{E}_{z \sim p_{\theta}(t)(\cdot|w)} \left[ \log p_\theta(z, w) \right]
\]

\(E\)-step: count under uncertainty

\(M\)-step: maximize log-likelihood
Why EM? Un-Supervised Learning

NO labeled data:
- human annotated
- relatively small/few examples

unlabeled data:
- raw; not annotated
- plentiful

EM/generative models in this case can be seen as a type of clustering
Why EM? Semi-Supervised Learning

- labeled data:
  - human annotated
  - relatively small/few examples

- unlabeled data:
  - raw; not annotated
  - plentiful
Why EM? Semi-Supervised Learning

- **labeled data:**
  - human annotated
  - relatively small/few examples

- **unlabeled data:**
  - raw; not annotated
  - plentiful
Why EM? Semi-Supervised Learning

labeled data:
- human annotated
- relatively small/few examples

unlabeled data:
- raw; not annotated
- plentiful
Why EM? Semi-Supervised Learning
EM (Expectation Maximization)

Basic idea

Three coins example

Why EM works
Imagine three coins

Flip 1\textsuperscript{st} coin (\textit{penny})

If heads: flip 2\textsuperscript{nd} coin (\textit{dollar coin})

If tails: flip 3\textsuperscript{rd} coin (\textit{dime})
Three Coins Example

Imagine three coins

Flip 1\textsuperscript{st} coin (penny)

If heads: flip 2\textsuperscript{nd} coin (dollar coin)

If tails: flip 3\textsuperscript{rd} coin (dime)

only observe these (record heads vs. tails outcome)

don’t observe this
Imagine three coins

Flip 1\textsuperscript{st} coin (penny)

If heads: flip 2\textsuperscript{nd} coin (dollar coin)

If tails: flip 3\textsuperscript{rd} coin (dime)

unobserved: part of speech? genre?

observed: $a, b, e$, etc.
We run the code, vs. The run failed
Three Coins Example

Imagine three coins

Flip 1\textsuperscript{st} coin (penny)

If heads: flip 2\textsuperscript{nd} coin (dollar coin)

If tails: flip 3\textsuperscript{rd} coin (dime)
Three Coins Example

Imagine three coins

\[ p(\text{heads}) = \lambda \]
\[ p(\text{tails}) = 1 - \lambda \]

\[ p(\text{heads}) = \gamma \]
\[ p(\text{tails}) = 1 - \gamma \]

\[ p(\text{heads}) = \psi \]
\[ p(\text{tails}) = 1 - \psi \]

Three parameters to estimate: \( \lambda, \gamma, \) and \( \psi \)
Generative Story for Three Coins

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Add complexity to better explain what we see

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N) = \prod_i p(w_i|z_i)p(z_i) \]

\[ p(\text{heads}) = \lambda \]
\[ p(\text{tails}) = 1 - \lambda \]

\[ p(\text{heads}) = \gamma \]
\[ p(\text{tails}) = 1 - \gamma \]

\[ p(\text{heads}) = \psi \]
\[ p(\text{tails}) = 1 - \psi \]

Generative Story

\[ \lambda = \text{distribution over penny} \]
\[ \gamma = \text{distribution for dollar coin} \]
\[ \psi = \text{distribution over dime} \]

for item \( i = 1 \) to \( N \):

\[ z_i \sim \text{Bernoulli}(\lambda) \]
\[ \text{if } z_i = H: w_i \sim \text{Bernoulli}(\gamma) \]
\[ \text{else: } w_i \sim \text{Bernoulli}(\psi) \]
Three Coins Example

If all flips were observed

\[ p(\text{heads}) = \lambda \quad p(\text{heads}) = \gamma \quad p(\text{heads}) = \psi \]

\[ p(\text{tails}) = 1 - \lambda \quad p(\text{tails}) = 1 - \gamma \quad p(\text{tails}) = 1 - \psi \]
Three Coins Example

If all flips were observed

\[ p(\text{heads}) = \lambda \]  \[ p(\text{heads}) = \gamma \]  \[ p(\text{heads}) = \psi \]

\[ p(\text{tails}) = 1 - \lambda \]  \[ p(\text{tails}) = 1 - \gamma \]  \[ p(\text{tails}) = 1 - \psi \]

\[ p(\text{heads}) = \frac{4}{6} \]  \[ p(\text{heads}) = \frac{1}{4} \]  \[ p(\text{heads}) = \frac{1}{2} \]

\[ p(\text{tails}) = \frac{2}{6} \]  \[ p(\text{tails}) = \frac{3}{4} \]  \[ p(\text{tails}) = \frac{1}{2} \]
Three Coins Example

But not all flips are observed → set parameter values

\[
p(\text{heads}) = \lambda = 0.6 \quad p(\text{heads}) = 0.8 \quad p(\text{heads}) = 0.6
\]
\[
p(\text{tails}) = 0.4 \quad p(\text{tails}) = 0.2 \quad p(\text{tails}) = 0.4
\]
Three Coins Example

But not all flips are observed → set parameter values

\[ p(\text{heads}) = \lambda = 0.6 \quad p(\text{heads}) = 0.8 \quad p(\text{heads}) = 0.6 \]
\[ p(\text{tails}) = 0.4 \quad p(\text{tails}) = 0.2 \quad p(\text{tails}) = 0.4 \]

Use these values to compute posteriors

\[ p(\text{heads} \mid \text{observed item } H) = \frac{p(\text{heads} \& H)}{p(H)} \]
\[ p(\text{heads} \mid \text{observed item } T) = \frac{p(\text{heads} \& T)}{p(T)} \]
Three Coins Example

But not all flips are observed → *set* parameter values

\[
p(\text{heads}) = \lambda = 0.6 \quad \quad p(\text{heads}) = 0.8 \quad \quad p(\text{heads}) = 0.6\]

\[
p(\text{tails}) = 0.4 \quad \quad p(\text{tails}) = 0.2 \quad \quad p(\text{tails}) = 0.4\]

Use these values to compute posteriors

\[
p(\text{heads} | \text{observed item H}) = \frac{p(H | \text{heads})p(\text{heads})}{p(H)}
\]

*rewrite joint using Bayes rule*

*marginal likelihood*
Three Coins Example

\[ HHHTHTHH \]
\[ HTHHTTTT \]

But not all flips are observed \( \Rightarrow \) set parameter values

\[ p(\text{heads}) = \lambda = .6 \quad p(\text{heads}) = .8 \quad p(\text{heads}) = .6 \]
\[ p(\text{tails}) = .4 \quad p(\text{tails}) = .2 \quad p(\text{tails}) = .4 \]

Use these values to compute posteriors

\[
p(\text{heads} \mid \text{observed item } H) = \frac{p(H \mid \text{heads})p(\text{heads})}{p(H)}
\]

\[ p(H \mid \text{heads}) = .8 \quad p(T \mid \text{heads}) = .2 \]
Three Coins Example

But not all flips are observed → set parameter values

\[
p(\text{heads}) = \lambda = .6 \quad p(\text{heads}) = .8 \quad p(\text{heads}) = .6 \\
p(\text{tails}) = .4 \quad p(\text{tails}) = .2 \quad p(\text{tails}) = .4
\]

Use these values to compute posteriors

\[
p(\text{heads} | \text{observed item } H) = \frac{p(H | \text{heads})p(\text{heads})}{p(H)}
\]

\[
p(H | \text{heads}) = .8 \quad p(T | \text{heads}) = .2
\]

\[
p(H) = p(H | \text{heads}) \times p(\text{heads}) + p(H | \text{tails}) \times p(\text{tails}) \\
= .8 \times .6 + .6 \times .4
\]
Three Coins Example

\[
\begin{array}{ccccccc}
H & H & T & H & T & H & H \\
H & T & H & T & T & T & T \\
\end{array}
\]

Use posteriors to update parameters

\[
p(\text{heads} | \text{obs. } H) = \frac{p(H | \text{heads})p(\text{heads})}{p(H)} = \frac{.8 \times .6}{.8 \times .6 + .6 \times .4} \approx 0.667
\]

\[
p(\text{heads} | \text{obs. } T) = \frac{p(T | \text{heads})p(\text{heads})}{p(T)} = \frac{.2 \times .6}{.2 \times .6 + .6 \times .4} \approx 0.334
\]

Q: Is \( p(\text{heads} | \text{obs. } H) + p(\text{heads} | \text{obs. } T) = 1 \)?
Three Coins Example

Use posteriors to update parameters

$$p(\text{heads} \mid \text{obs. } H) = \frac{p(H \mid \text{heads})p(\text{heads})}{p(H)} = \frac{.8 \times .6}{.8 \times .6 + .6 \times .4} \approx 0.667$$

$$p(\text{heads} \mid \text{obs. } T) = \frac{p(T \mid \text{heads})p(\text{heads})}{p(T)} = \frac{.2 \times .6}{.2 \times .6 + .6 \times .4} \approx 0.334$$

Q: Is $p(\text{heads} \mid \text{obs. } H) + p(\text{heads} \mid \text{obs. } T) = 1$?

A: No.
Three Coins Example

Use posteriors to update parameters

\[
p(\text{heads} | \text{obs. H}) = \frac{p(H|\text{heads})p(\text{heads})}{p(H)} = \frac{.8 \times .6}{.8 \times .6 + .6 \times .4} \approx 0.667
\]

\[
p(\text{heads} | \text{obs. T}) = \frac{p(T|\text{heads})p(\text{heads})}{p(T)} = \frac{.2 \times .6}{.2 \times .6 + .6 \times .4} \approx 0.334
\]

(in general, \( p(\text{heads} | \text{obs. H}) \) and \( p(\text{heads} | \text{obs. T}) \) do NOT sum to 1)

fully observed setting  
\[
p(\text{heads}) = \frac{\# \text{ heads from penny}}{\# \text{ total flips of penny}}
\]

our setting: partially-observed  
\[
p(\text{heads}) = \frac{\# \text{ expected heads from penny}}{\# \text{ total flips of penny}}
\]
Three Coins Example

Use posteriors to update parameters

\[
p(\text{heads} \mid \text{obs. } H) = \frac{p(H \mid \text{heads})p(\text{heads})}{p(H)} = \frac{.8 \times .6}{.8 \times .6 + .6 \times .4} \approx 0.667
\]

\[
p(\text{heads} \mid \text{obs. } T) = \frac{p(T \mid \text{heads})p(\text{heads})}{p(T)} = \frac{.2 \times .6}{.2 \times .6 + .6 \times .4} \approx 0.334
\]

our setting: partially-observed

\[
p^{(t+1)}(\text{heads}) = \frac{\# \text{ expected heads from penny}}{\# \text{ total flips of penny}} = \mathbb{E}_{p^{(t)}}[\# \text{ expected heads from penny}] / \# \text{ total flips of penny}
\]
Three Coins Example

Use posteriors to update parameters

\[
p(\text{heads} \mid \text{obs. H}) = \frac{p(H \mid \text{heads})p(\text{heads})}{p(H)} = \frac{.8 \ast .6}{.8 \ast .6 + .6 \ast .4} \approx 0.667
\]

\[
p(\text{heads} \mid \text{obs. T}) = \frac{p(T \mid \text{heads})p(\text{heads})}{p(T)} = \frac{.2 \ast .6}{.2 \ast .6 + .6 \ast .4} \approx 0.334
\]

\[
p^{(t+1)}(\text{heads}) = \frac{\# \text{ expected heads from penny}}{\# \text{ total flips of penny}} = \frac{\mathbb{E}_{p^{(t)}}[\# \text{ expected heads from penny}]}{\# \text{ total flips of penny}} = \frac{2 \ast p(\text{heads} \mid \text{obs. H}) + 4 \ast p(\text{heads} \mid \text{obs. T})}{6} \approx 0.444
\]
Expectation Maximization (EM)

0. Assume *some* value for your parameters

Two step, iterative algorithm:

1. E-step: count under uncertainty (compute expectations)

2. M-step: maximize log-likelihood, assuming these uncertain counts
Outline

EM (Expectation Maximization)

Basic idea

Three coins example

Why EM works
Why does EM work?

<table>
<thead>
<tr>
<th>$X$: observed data</th>
<th>$Y$: unobserved data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}(\theta) = \text{marginal log-likelihood of observed data} ; X$</td>
<td>$C(\theta) = \text{log-likelihood of complete data} ; (X,Y)$</td>
</tr>
<tr>
<td>$P(\theta) = \text{posterior log-likelihood of incomplete data} ; Y$</td>
<td></td>
</tr>
</tbody>
</table>

what do $C, \mathcal{M}, P$ look like?
Why does EM work?

$X$: observed data  
$Y$: unobserved data

$\mathcal{M}(\theta) = \text{marginal log-likelihood of observed data } X$

$\mathcal{C}(\theta) = \text{log-likelihood of complete data } (X,Y)$

$\mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y$

$\mathcal{C}(\theta) = \sum_i \log p(x_i, y_i)$
Why does EM work?

- **X**: observed data
- **Y**: unobserved data
- **$C(\theta)$**: log-likelihood of complete data (X,Y)
- **$M(\theta)$**: marginal log-likelihood of observed data X
- **$P(\theta)$**: posterior log-likelihood of incomplete data Y

Mathematical expressions:

$$C(\theta) = \sum_i \log p(x_i, y_i)$$

$$M(\theta) = \sum_i \log p(x_i) = \sum_i \log \sum_k p(x_i, y = k)$$
Why does EM work?

\[ \mathcal{C}(\theta) = \text{log-likelihood of complete data (X,Y)} \]

\[ \mathcal{M}(\theta) = \text{marginal log-likelihood of observed data X} \]

\[ \mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data Y} \]

\[ \mathcal{C}(\theta) = \sum_i \log p(x_i, y_i) \]

\[ \mathcal{M}(\theta) = \sum_i \log p(x_i) = \sum_i \log \sum_k p(x_i, y = k) \]

\[ \mathcal{P}(\theta) = \sum_i \log p(y_i | x_i) \]
Why does EM work?

\( X: \) observed data \hspace{1cm} \( Y: \) unobserved data

\( C(\theta) = \text{log-likelihood of complete data } (X,Y) \)

\( M(\theta) = \text{marginal log-likelihood of observed data } X \)

\( P(\theta) = \text{posterior log-likelihood of incomplete data } Y \)

\[
p_\theta(Y \mid X) = \frac{p_\theta(X, Y)}{p_\theta(X)} \quad \text{definition of conditional probability}
\]

\[ p_\theta(X) = \frac{p_\theta(X, Y)}{p_\theta(Y \mid X)} \quad \text{algebra} \]
Why does EM work?

\( X: \) observed data \hspace{1cm} \( Y: \) unobserved data

\( C(\theta) = \) log-likelihood of complete data \((X,Y)\)

\( \mathcal{M}(\theta) = \) marginal log-likelihood of observed data \(X\)

\( \mathcal{P}(\theta) = \) posterior log-likelihood of incomplete data \(Y\)

\[
\begin{align*}
\mathcal{M}(\theta) &= \mathcal{P}(\theta) - \mathcal{P}(\theta) \\
\mathcal{P}(\theta) &= \sum_i \log p(y_i | x_i)
\end{align*}
\]
Why does EM work?

- **X**: observed data
- **Y**: unobserved data
- \( \mathcal{M}(\theta) = \) marginal log-likelihood of observed data \( X \)
- \( \mathcal{C}(\theta) = \) log-likelihood of complete data \((X, Y)\)
- \( \mathcal{P}(\theta) = \) posterior log-likelihood of incomplete data \( Y \)

\[
p_\theta(Y \mid X) = \frac{p_\theta(X, Y)}{p_\theta(X)} \quad \Rightarrow \quad p_\theta(X) = \frac{p_\theta(X, Y)}{p_\theta(Y \mid X)}
\]

\[
\mathcal{M}(\theta) = \mathcal{C}(\theta) - \mathcal{P}(\theta)
\]

\[
\mathbb{E}_{Y \sim \theta(t)}[\mathcal{M}(\theta) \mid X] = \mathbb{E}_{Y \sim \theta(t)}[\mathcal{C}(\theta) \mid X] - \mathbb{E}_{Y \sim \theta(t)}[\mathcal{P}(\theta) \mid X]
\]

*take a conditional expectation
* (why? we’ll cover this more in variational inference)
Why does EM work?

\[ X: \text{observed data} \quad Y: \text{unobserved data} \]

\[ \mathcal{M}(\theta) = \text{marginal log-likelihood of observed data } X \]

\[ \mathcal{C}(\theta) = \text{log-likelihood of complete data } (X,Y) \]

\[ \mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y \]

\[ p_\theta(Y | X) = \frac{p_\theta(X, Y)}{p_\theta(X)} \quad \rightarrow \quad p_\theta(X) = \frac{p_\theta(X, Y)}{p_\theta(Y | X)} \]

\[ \mathcal{M}(\theta) = \mathcal{C}(\theta) - \mathcal{P}(\theta) \]

\[ \mathbb{E}_{Y \sim \theta(t)}[\mathcal{M}(\theta)|X] = \mathbb{E}_{Y \sim \theta(t)}[\mathcal{C}(\theta)|X] - \mathbb{E}_{Y \sim \theta(t)}[\mathcal{P}(\theta)|X] \]

\[ \mathcal{M}(\theta) = \mathbb{E}_{Y \sim \theta(t)}[\mathcal{C}(\theta)|X] - \mathbb{E}_{Y \sim \theta(t)}[\mathcal{P}(\theta)|X] \]

\[ \mathcal{M}(\theta) = \sum_i \log p(x_i) = \sum_i \log \sum_k p(x_i, y = k) \]

\[ \mathcal{M} \text{ already sums over } Y \]
Why does EM work?

\(X: \text{observed data}\)

\(Y: \text{unobserved data}\)

\(\mathcal{M}(\theta) = \text{marginal log-likelihood of observed data } X\)

\(\mathcal{C}(\theta) = \log\text{-likelihood of complete data } (X,Y)\)

\(\mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y\)

\[
\mathcal{M}(\theta) = \mathbb{E}_{Y \sim \theta(t)}[\mathcal{C}(\theta) | X] - \mathbb{E}_{Y \sim \theta(t)}[\mathcal{P}(\theta) | X]
\]

\[
\mathbb{E}_{Y \sim \theta(t)}[\mathcal{C}(\theta) | X] = \sum_i \sum_k p_{\theta(t)}(y = k | x_i) \log p(x_i, y = k)
\]
Why does EM work?

Let $\theta^*$ be the value that maximizes $Q(\theta, \theta^{(t)})$
Why does EM work?

\[ X: \text{observed data} \quad Y: \text{unobserved data} \]

\[ \mathcal{M}(\theta) = \text{marginal log-likelihood of observed data } X \]

\[ \mathcal{C}(\theta) = \text{log-likelihood of complete data } (X,Y) \]

\[ \mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y \]

\[ \mathcal{M}(\theta) = \mathbb{E}_{Y \sim \theta^{(t)}}[\mathcal{C}(\theta)|X] - \mathbb{E}_{Y \sim \theta^{(t)}}[\mathcal{P}(\theta)|X] \]

\[ Q(\theta, \theta^{(t)}) = \mathbb{E}_{Y \sim \theta^{(t)}}[\mathcal{C}(\theta)|X] - \mathbb{E}_{Y \sim \theta^{(t)}}[\mathcal{P}(\theta)|X] \]

\[ R(\theta, \theta^{(t)}) = \mathbb{E}_{Y \sim \theta^{(t)}}[\mathcal{P}(\theta)|X] - \mathbb{E}_{Y \sim \theta^{(t)}}[\mathcal{P}(\theta)|X] \]

Let \( \theta^* \) be the value that maximizes \( Q(\theta, \theta^{(t)}) \)

\[ \mathcal{M}(\theta^*) - \mathcal{M}(\theta^{(t)}) = (Q(\theta^*, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})) - (R(\theta^*, \theta^{(t)}) - R(\theta^{(t)}, \theta^{(t)})) \]
Why does EM work?

- **X**: observed data
- **Y**: unobserved data
- \( C(\theta) = \) log-likelihood of complete data \((X,Y)\)
- \( P(\theta) = \) posterior log-likelihood of incomplete data \(Y\)
- \( M(\theta) = \) marginal log-likelihood of observed data \(X\)

\[
M(\theta) = \mathbb{E}_{Y \sim \theta(t)}[C(\theta)|X] - \mathbb{E}_{Y \sim \theta(t)}[P(\theta)|X]
\]

Let \( \theta^* \) be the value that maximizes \( Q(\theta, \theta^{(t)}) \)

\[
M(\theta^*) - M(\theta^{(t)}) = (Q(\theta^*, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})) - (R(\theta^*, \theta^{(t)}) - R(\theta^{(t)}, \theta^{(t)})) \geq 0
\]

\( \leq 0 \) (we’ll see why with Jensen’s inequality, in variational inference)
Why does EM work?

- **X**: observed data
- **Y**: unobserved data
- $\mathcal{M}(\theta) = \text{marginal log-likelihood of observed data } X$
- $\mathcal{C}(\theta) = \text{log-likelihood of complete data } (X,Y)$
- $\mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y$
- $\mathcal{Q}(\theta, \theta(t))$
- $\mathcal{R}(\theta, \theta(t))$

\[
\mathcal{M}(\theta) = \mathbb{E}_{Y \sim \theta(t)}[\mathcal{C}(\theta)|X] - \mathbb{E}_{Y \sim \theta(t)}[\mathcal{P}(\theta)|X]
\]

Let $\theta^*$ be the value that maximizes $Q(\theta, \theta(t))$

\[
\mathcal{M}(\theta^*) - \mathcal{M}(\theta(t)) = (Q(\theta^*, \theta(t)) - Q(\theta(t), \theta(t))) - (R(\theta^*, \theta(t)) - R(\theta(t), \theta(t)))
\]

\[
\mathcal{M}(\theta^*) - \mathcal{M}(\theta(t)) \geq 0
\]

EM does not decrease the marginal log-likelihood
Generalized EM

Partial M step: find a $\theta$ that simply increases, rather than maximizes, $Q$

Partial E step: only consider some of the variables (an online learning algorithm)
EM has its pitfalls

Objective is not convex $\rightarrow$ converge to a bad local optimum

Computing expectations can be hard: the E-step could require clever algorithms

How well does log-likelihood correlate with an end task?
A Maximization-Maximization Procedure

\[ F(\theta, q) = \mathbb{E}[C(\theta)] - \mathbb{E}[\log q(Z)] \]

*any* distribution over \( Z \)

we’ll see this again with variational inference
EM (Expectation Maximization)

Basic idea

Three coins example

Why EM works