Introduction to Expectation Maximization

CMSC 678

UMBC

March 28th, 2018
Outline

Administrivia & recap

Latent and probabilistic modeling

EM (Expectation Maximization)
  Basic idea
  Three coins example
  Why EM works
Announcement 1: Assignment 3

Due Wednesday April 11\textsuperscript{th}, 11:59 AM

Any questions?
Announcement 2: Progress Report on Project

Due Monday April 16\textsuperscript{th}, 11:59 AM

Build on the proposal:
  
  Update to address comments
  Discuss the progress you’ve made
  Discuss what remains to be done
  Discuss any new blocks you’ve experienced (or anticipate experiencing)

Any questions?
Recap(+) from last time...
Optimization for SVMs

Separable case: hard margin SVM

\[
\min_w \frac{1}{2} \|w\|^2 \\
\text{maximize margin}
\]

subject to: \( y_n w^T x_n \geq 1, \forall n \)

Non-separable case: soft margin SVM

\[
\min_w \frac{1}{2} \|w\|^2 + C \sum_n \xi_n \\
\text{maximize margin, minimize slack}
\]

subject to: \( y_n w^T x_n \geq 1 - \xi_n, \forall n \)

\( \xi_n \geq 0 \)
Subgradient is any direction that is below the function

For the hinge loss a possible subgradient is:

\[
\ell(y, \hat{y}) = \max\{0, 1 - yw^T x\}
\]

[Graph showing different subgradients and their slopes, indicating that the subgradient is not differentiable at \( z = 1 \).]

Subgradient is any direction that is below the function

For the hinge loss a possible subgradient is:

\[
\frac{d\ell^{\text{hinge}}}{dw} = \begin{cases} 
0 & \text{if } yw^T x > 1 \\
-yx & \text{otherwise}
\end{cases}
\]
Example: Hinge loss

$$\mathcal{L}(w) = \sum_{n} \max(0, 1 - y_n w^T x_n) + \frac{\lambda}{2} \|w\|^2 \quad \text{objective}$$

$$\frac{d\mathcal{L}}{dw} = \sum_{n} -1[y_n w^T x_n \leq 1] y_n x_n + \lambda w \quad \text{subgradient}$$

$$w \leftarrow w - \eta \left( \sum_{n} -1[y_n w^T x_n \leq 1] y_n x_n + \lambda w \right) \quad \text{update}$$

**Take-away:** the perceptron algorithm optimizes a hinge loss objective

**loss term**

$$w \leftarrow w + \eta y_n x_n$$

only for points $y_n w^T x_n \leq 1$

perceptron update $y_n w^T x_n \leq 0$
Lagrange multipliers

Assume an original optimization problem

\[
\min_x x^2 \\
\text{st. } x \geq b
\]

We convert it to a new optimization problem:

\[
\min_x \max_\alpha L(x, \alpha) \\
\text{st. } \alpha \geq 0 \\
L(x, \alpha) = x^2 - \alpha(x - b)
\]
Dual formulation for SVM

\[ \min_{\omega,b,\xi} \max_{\alpha \geq 0} \max_{\beta \geq 0} \mathcal{L}(\omega, b, \xi, \alpha, \beta) \]

\[ \mathcal{L}(\omega, b, \xi, \alpha, \beta) = \frac{1}{2} ||\omega||^2 + C \sum_n \xi_n - \sum_n \beta_n \xi_n \]

\[ - \sum_n \alpha_n [y_n (\omega \cdot x_n + b) - 1 + \xi_n] \]

\[ \min_{\alpha} -\mathcal{L}(\alpha) = \frac{1}{2} \sum_n \sum_m \alpha_n \alpha_m y_n y_m K(x_n, x_m) - \sum_n \alpha_n \]

subj. to \( 0 \leq \alpha_n \leq C \)
Course Overview (so far)

Basics of Probability
- Requirements to be a distribution (‘proportional to’, $\propto$)
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)
- Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradients
Remember from the first day: A Terminology Buffet

Classification
Regression
Clustering

Fully-supervised
Semi-supervised
Un-supervised

the task: what kind of problem are you solving?
the data: amount of human input/number of labeled examples
the approach: how any data are being used
Remember from the first day:
A Terminology Buffet

what we’ve currently sampled...

Classification
Regression
Clustering

Fully-supervised

Semi-supervised

Un-supervised

Probabilistic
Neural
Generative
Memory-based
Conditional
Exemplar
Spectral
...

the task: what kind of problem are you solving?
the data: amount of human input/number of labeled examples
the approach: how any data are being used
Remember from the first day: A Terminology Buffet

**Classification**
- Fully-supervised
- Semi-supervised
- Un-supervised

**Regression**

**Clustering**

---

**the task**: what kind of problem are you solving?

**the data**: amount of human input/number of labeled examples

**the approach**: how any data are being used

what we’ve currently sampled...
what we’ll be sampling next...
LATENT VARIABLE MODELS AND EXPECTATION MAXIMIZATION
Outline

Administrivia & recap

Latent and probabilistic modeling

EM (Expectation Maximization)
  Basic idea
  Three coins example
  Why EM works
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.
Is (Supervised) Classification “Latent?”
Not Obviously

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

best label = \( \arg \max_{\text{label}} p(\text{label} | \text{item}) \)

MAP classifier
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

Is (Supervised) Classification “Latent?”

Not Obviously

these values are unknown but the generation process (explanation) is transparent

Input:
an instance $d$
a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
A training set of $m$ hand-labeled instances $(d_1, c_1), \ldots, (d_m, c_m)$

Output:
a learned (probabilistic) classifier $\gamma$ that maps instances to classes
Latent Sequences: Part of Speech

British Left Waffles on Falkland Islands

British Left Waffles on Falkland Islands

Adjective  Noun  Verb

British Left Waffles on Falkland Islands

Noun  Verb  Noun
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse

Adapted from Jason Eisner, Noah Smith
Latent Modeling

explain what you see/annotate

with things “of importance” you don’t

orthography

morphology

lexemes

syntax

semantics

pragmatics

discourse

observed text

Adapted from Jason Eisner, Noah Smith
Latent Sequence Models: Part of Speech

\[ p(\text{British Left Waffles on Falkland Islands}) \]
Latent Sequence Models: Part of Speech

\( p(\text{British Left Waffles on Falkland Islands}) \)

(i): Adjective Noun Verb Prep Noun Noun

(ii): Noun Verb Noun Prep Noun Noun Noun
Latent Sequence Models: Part of Speech

\[ p(\text{British Left Waffles on Falkland Islands}) \]

1. Explain this sentence as a sequence of (likely?) latent (unseen) tags (labels): 

(i): Adjective Noun Verb Prep Noun Noun
(ii): Noun Verb Noun Prep Noun Noun
Latent Sequence Models: Part of Speech

p(British Left Waffles on Falkland Islands)

1. Explain this sentence as a sequence of (likely?) latent (unseen) tags (labels)

2. Produce a tag sequence for this sentence
Example: Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]
Example: Rolling a Die

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

N different (independent) rolls

\[
\begin{align*}
  w_1 &= 1 \\
  w_2 &= 5 \\
  w_3 &= 4 \\
  \cdots
\end{align*}
\]
Example: Rolling a Die

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

maximize (log-) likelihood to learn the probability parameters
Example: Rolling a Die

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

maximize (log-) likelihood to learn the probability parameters

p(1) = ?  p(2) = ?

p(3) = ?  p(4) = ?

p(5) = ?  p(6) = ?

maximum likelihood estimates
Example: Rolling a Die

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

maximize (log-) likelihood to learn the probability parameters

\[
\begin{align*}
p(1) &= 2/9 & p(2) &= 1/9 \\
p(3) &= 1/9 & p(4) &= 3/9 \\
p(5) &= 1/9 & p(6) &= 1/9
\end{align*}
\]
Example: Conditionally Rolling a Die

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

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

\[= \prod_i p(w_i|z_i)p(z_i)\]
Example: Conditionally Rolling a Die

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

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

\[ z_1 = A \quad w_1 = 1 \]
\[ z_2 = B \quad w_2 = 5 \]
\[ \ldots \]
Example: Conditionally Rolling a Die

\[
p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i)
\]

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

\[
= \prod_{i} p(w_i|z_i)p(z_i)
\]

examples of latent classes z:
- part of speech tag
- topic ("sports" vs. "politics")
Example: Conditionally Rolling a Die

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

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

goal: maximize (log-)likelihood
Example: Conditionally Rolling a Die

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

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

goal: maximize (log-)likelihood

we don’t actually observe these \( z \) values

we just see the observed items (rolls) \( w \)
Example: Conditionally Rolling a Die

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

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

\[
= \prod_i p(w_i|z_i) p(z_i)
\]

Goal: maximize (log-)likelihood

*we don’t actually observe these z values*

*we just see the items w*

If we *did* observe z, estimating the probability parameters would be easy...

But we don’t! :(

Example: Conditionally Rolling a Die

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

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N) \]

\[ = \prod_i p(w_i|z_i) p(z_i) \]

**Goal:** maximize (log-)likelihood

*we don’t actually observe these z values*

*we just see the items w*

if we *did* observe z, estimating the probability parameters would be easy...

but we don’t! :( if we *knew* the probability parameters then we could estimate z and evaluate likelihood... but we don’t! :(/

add complexity to better explain what we see
Example: Conditionally Rolling a Die

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

\[ = \prod_i p(w_i|z_i) \, p(z_i) \]

*we don’t actually observe these z values*

goal: maximize *marginalized* (log-)likelihood
Example: Conditionally Rolling a Die

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

*we don’t actually observe these z values*

goal: maximize *marginalized* (log-)likelihood
Example: Conditionally Rolling a Die

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

\[ = \prod_i p(w_i|z_i) p(z_i) \]

*we don’t actually observe these z values*

goal: maximize *marginalized* (log-)likelihood
Example: Conditionally Rolling a Die

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

\[ = \prod_i p(w_i|z_i) p(z_i) \]

we don’t actually observe these \( z \) values

goal: maximize \textbf{marginalized} (log-)likelihood

\[ p(w_1, w_2, \ldots, w_N) = \left( \sum_{z_1} p(z_1, w) \right) \left( \sum_{z_2} p(z_2, w) \right) \cdots \left( \sum_{z_N} p(z_N, w) \right) \]
Example: Conditionally Rolling a Die

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

Goal: maximize \textit{marginalized} (log-)likelihood

\[
p(w_1, w_2, \ldots, w_N) = \left( \sum_{z_1} p(z_1, w_1) \right) \left( \sum_{z_2} p(z_2, w_2) \right) \cdots \left( \sum_{z_N} p(z_N, w_N) \right)
\]

if we \textit{did} observe \( z \), estimating the probability parameters would be easy... but we don’t! :(

if we \textit{knew} the probability parameters then we could estimate \( z \) and evaluate likelihood... but we don’t! :(
if we *did* observe $z$, estimating the probability parameters would be easy... but we don’t! :(

if we *knew* the probability parameters then we could estimate $z$ and evaluate likelihood... but we don’t! :(
if we did observe $z$, estimating the probability parameters would be easy...

but we don't! :(

if we knew the probability parameters then we could estimate $z$ and evaluate likelihood... but we don't! :(
Outline

Administrivia & recap

Latent and probabilistic modeling

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

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

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

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

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

\[ p^{(t+1)}(z) \]

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

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

*E-step: count under uncertainty*

*M-step: maximize log-likelihood*
EM Math

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

*E-step: count under uncertainty*

*old parameters*

*posterior distribution*

*M-step: maximize log-likelihood*
EM Math

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

*E-step: count under uncertainty*

*old parameters*

*posterior distribution*

*new parameters*

*M-step: maximize log-likelihood*
Outline

Administrivia & recap

Latent and probabilistic modeling

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})
Imagine three coins

Flip 1st coin (penny)

If heads: flip 2nd coin (dollar coin)

If tails: flip 3rd coin (dime)
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 \textit{run} failed
Imagine three coins

Flip 1st coin (penny)
$p(\text{heads}) = \lambda$
$p(\text{tails}) = 1 - \lambda$

If heads: flip 2nd coin (dollar coin)
$p(\text{heads}) = \gamma$
$p(\text{tails}) = 1 - \gamma$

If tails: flip 3rd coin (dime)
$p(\text{heads}) = \psi$
$p(\text{tails}) = 1 - \psi$
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 \( \psi \)
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

\[
\begin{align*}
\text{H H T H T H} \\
\text{H T H T T T}
\end{align*}
\]

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

\[
\begin{align*}
p(\text{heads}) &= \frac{4}{6} & \quad p(\text{heads}) &= \frac{1}{4} & \quad p(\text{heads}) &= \frac{1}{2} \\
p(\text{tails}) &= \frac{2}{6} & \quad p(\text{tails}) &= \frac{3}{4} & \quad p(\text{tails}) &= \frac{1}{2}
\end{align*}
\]
Three Coins Example

But not all flips are observed → set parameter values

\[
\begin{align*}
    p(\text{heads}) &= \lambda = 0.6 &
    p(\text{heads}) &= 0.8 &
    p(\text{heads}) &= 0.6 \\
    p(\text{tails}) &= 0.4 &
    p(\text{tails}) &= 0.2 &
    p(\text{tails}) &= 0.4
\end{align*}
\]
Three Coins Example

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

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

But not all flips are observed → set parameter values

\[ \begin{align*}
  p(\text{heads}) &= \lambda = .6 & p(\text{heads}) &= .8 & p(\text{heads}) &= .6 \\
  p(\text{tails}) &= .4 & p(\text{tails}) &= .2 & p(\text{tails}) &= .4
\end{align*} \]

Use these values to compute posteriors

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

rewrite joint using Bayes rule
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

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

\[
p(H \mid \text{heads}) = 0.8 \quad p(T \mid \text{heads}) = 0.2
\]
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

\[ H \ H \ T \ H \ T \ H \]

\[ H \ T \ H \ T \ T \ T \]

Use posteriors to update parameters

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

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

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

\[\begin{align*}
\mathbb{H} & \mathbb{H} \mathbb{T} \mathbb{H} \mathbb{T} \mathbb{H} \\
\mathbb{H} & \mathbb{T} \mathbb{H} \mathbb{T} \mathbb{T} \mathbb{T}
\end{align*}\]

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

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

Use posteriors to update parameters

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

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

(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

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

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

our setting: partially-observed

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

Use posteriors to update parameters

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

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

our setting: partially-observed

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

\[
= \frac{2 \times p(\text{heads} \mid \text{obs. H}) + 4 \times p(\text{heads} \mid \text{obs. T})}{6} \approx 0.444
\]
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

Administrivia & recap

Latent and probabilistic modeling

EM (Expectation Maximization)
  Basic idea
  Three coins example
  Why EM works
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 \( Y \)

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

\( X: \) observed data  \( Y: \) unobserved data

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

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

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

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

<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>$\mathcal{C}(\theta) = \text{log-likelihood of complete data } (X,Y)$</td>
</tr>
<tr>
<td>$\mathcal{P}(\theta) = \text{posterior log-likelihood of incomplete data } Y$</td>
<td></td>
</tr>
</tbody>
</table>

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

\[ \mathcal{M}(\theta) = \sum_i \log p(x_i) = \sum_i \log \sum_k p(x_i, y = k) \]
Why does EM work?

\(X: \text{observed data} \quad Y: \text{unobserved data}\)

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

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

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

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

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

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

\(X\): observed data  \(Y\): unobserved data

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

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

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

\[p_\theta(Y \mid X) = \frac{p_\theta(X, Y)}{p_\theta(X)}\]  \(\text{algebra}\)  \[p_\theta(X) = \frac{p_\theta(X, Y)}{p_\theta(Y \mid X)}\]

definition of conditional probability
Why does EM work?

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

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

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

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

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

\[ C(\theta) = \sum_i \log p(x_i, y_i) \quad \quad \quad \quad \quad M(\theta) = \sum_i \log p(x_i) = \sum_i \log \sum_k p(x_i, y = k) \quad \quad \quad \quad \quad P(\theta) = \sum_i \log p(y_i \mid x_i) \]

\[ M(\theta) = C(\theta) - P(\theta) \]
Why does EM work?

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

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

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

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

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

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

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

*take a conditional expectation*

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

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

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

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

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

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

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

\( \mathcal{M} \text{ already sums over } Y \)

\[
\mathcal{M}(\theta) = \sum_{i} \log p(x_i) = \sum_{i} \log \sum_{k} p(x_i, y = k)
\]
Why does EM work?

$X$: observed data

$Y$: unobserved data

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

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

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

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

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

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

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

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

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

\( Q(\theta, \theta^{(t)}) \)

\( R(\theta, \theta^{(t)}) \)

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

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

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

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

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

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

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

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

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

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

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

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

\[ Q(\theta, \theta^{(t)}) \]

\[ R(\theta, \theta^{(t)}) \]

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

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

\[ \geq 0 \]

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

X: observed data  
Y: unobserved data

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

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

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

\[
M(\theta) = E_{Y \sim \theta^{(t)}} [C(\theta)|X] - E_{Y \sim \theta^{(t)}} [P(\theta)|X] = Q(\theta, \theta^{(t)}) - R(\theta, \theta^{(t)})
\]

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

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

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

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