Expectation Maximization

What it is and how you use it

The Goal

The Goal

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

The Goal

- You have some observed data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

The Goal

- You have some observed data

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

The Goal

- You have some observed data
- You have a probabilistic model that "generated" your data

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

The Goal

- You have some observed data
- You have a probabilistic model that "generated" your data

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

Model

The Goal

- You have some observed data
- You have a probabilistic model that “generated” your data
- What are the most likely parameters of your model?

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

Model

The Goal

- You have some observed data
- You have a probabilistic model that “generated” your data
- What are the most likely parameters of your model?

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

Model

Parameters

The Goal

- You have some observed data
- You have a probabilistic model that "generated" your data
- What are the most likely parameters of your model?
- Let's find parameters that **maximize** the **expected** log likelihood of your data

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

Model

Parameters

The Goal

- You have some observed data
- You have a probabilistic model that "generated" your data
- What are the most likely parameters of your model?
- Let's find parameters that maximize the expected log likelihood of your data
- Why is this hard? Complex models, lots of parameters, and hidden data.

Data

$$\arg\max_{\theta} \mathbb{E}[\log p(x|\theta)]$$

Model

Parameters

Toys:
a random experiment

Toys:

a random experiment

Toys:
a random experiment

- We let n children each choose one of four toys, and keep a histogram y of their choices:

Toys:
a random experiment

- We let n children each choose one of four toys, and keep a histogram y of their choices:

$$\vec{y} = [y_1, y_2, y_3, y_4]; \sum_{i=1}^{4} y_i = n$$

Toys:
a random experiment

- We let n children each choose one of four toys, and keep a histogram y of their choices:

$$\vec{y} = [y_1, y_2, y_3, y_4]; \sum_{i=1}^{4} y_i = n$$

- Let's pick a Multinomial model with toy probabilities:

Toys:
a random experiment

- We let n children each choose one of four toys, and keep a histogram y of their choices:

$$\vec{y} = [y_1, y_2, y_3, y_4]; \sum_{i=1}^{4} y_i = n$$

- Let's pick a Multinomial model with toy probabilities:

$$\vec{p} = [p_1, p_2, p_3, p_4]; \sum_{i=1}^{4} p_i = 1$$

Toys:
a random experiment

- We let n children each choose one of four toys, and keep a histogram y of their choices:

$$\vec{y} = [y_1, y_2, y_3, y_4]; \sum_{i=1}^{4} y_i = n$$

- Let's pick a Multinomial model with toy probabilities:

$$\vec{p} = [p_1, p_2, p_3, p_4]; \sum_{i=1}^{4} p_i = 1$$

- Our model probability for any particular histogram is:

Toys:
a random experiment

- We let n children each choose one of four toys, and keep a histogram y of their choices:

$$\vec{y} = [y_1, y_2, y_3, y_4]; \sum_{i=1}^{4} y_i = n$$

- Let's pick a Multinomial model with toy probabilities:

$$\vec{p} = [p_1, p_2, p_3, p_4]; \sum_{i=1}^{4} p_i = 1$$

- Our model probability for any particular histogram is:

$$p(\vec{y}|\vec{p}, n) \sim Mu(\vec{y}|n, \vec{p})$$

$$= \frac{n!}{\prod_{i=1}^{4} y_i!} \prod_{i=1}^{4} p_i^{y_i}$$

Toys:
parameters and hidden data

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter □:

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter θ:

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter θ:

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- We also think that the kids really choose based on being in one of five hidden states of mind, and we want to count up how many kids are in each:

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter □:

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- We also think that the kids really choose based on being in one of five hidden states of mind, and we want to count up how many kids are in each:

$$\vec{x} \sim Mu(n, \vec{q}_\theta); \vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_4]$$

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter □:

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- We also think that the kids really choose based on being in one of five hidden states of mind, and we want to count up how many kids are in each:

$$\vec{x} \sim Mu(n, \vec{q}_\theta); \vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_4]$$

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- This gives us a new model based on the hidden X rather than the observed Y:

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter □:

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- We also think that the kids really choose based on being in one of five hidden states of mind, and we want to count up how many kids are in each:

$$\vec{x} \sim Mu(n, \vec{q}_\theta); \vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_4]$$

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- This gives us a new model based on the hidden X rather than the observed Y:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter :

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- We also think that the kids really choose based on being in one of five hidden states of mind, and we want to count up how many kids are in each:

$$\vec{x} \sim Mu(n, \vec{q}_\theta); \vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_4]$$

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- This gives us a new model based on the hidden X rather than the observed Y:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- All we need is to find a value of and values for X that fit our assumptions:

Toys:
parameters and hidden data

- Suppose we think the toy probabilities are related to some parameter :

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- We also think that the kids really choose based on being in one of five hidden states of mind, and we want to count up how many kids are in each:

$$\vec{x} \sim Mu(n, \vec{q}_\theta); \vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_4]$$

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- This gives us a new model based on the hidden X rather than the observed Y:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- All we need is to find a value of and values for X that fit our assumptions:

$$\arg\max_\theta \mathbb{E}[\log p(\vec{x}|\theta)]$$

Expectation Maximization:
The EM algorithm

EM finds the values of the parameters □ and hidden data X that maximize the likelihood of the observed data Y.

Expectation Maximization: The EM algorithm

EM finds the values of the parameters and hidden data X that maximize the likelihood of the observed data Y.

1. Guess initial parameter values $\theta^{(m=0)}$
2. Calculate the distribution over the data $p(\vec{x}|\vec{y},\theta^{(m)})$
3. Calculate the expected log probability for the data

$$\begin{aligned} Q(\theta|\theta^{(m)}) &\triangleq \mathbb{E}[\log p(\vec{x}|\theta)] \\ &= \sum_{\vec{x}} \log p(\vec{x}|\theta) p(\vec{x}|\vec{y},\theta^{(m)}) \end{aligned}$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_{\theta} \mathbb{E}[\log p(\vec{x}|\theta)] = \arg\max_{\theta} Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

Expectation Maximization:
The EM algorithm

EM finds the values of the parameters □ and hidden data X that maximize the likelihood of the observed data Y.

E-step

1. Guess initial parameter values $\theta^{(m=0)}$

2. Calculate the distribution over the data $p(\vec{x}|\vec{y},\theta^{(m)})$

3. Calculate the expected log probability for the data

$$
\begin{aligned}
Q(\theta|\theta^{(m)}) &\triangleq \mathbb{E}[\log p(\vec{x}|\theta)] \\
&= \sum_{\vec{x}} \log p(\vec{x}|\theta) p(\vec{x}|\vec{y},\theta^{(m)})
\end{aligned}
$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_{\theta} \mathbb{E}[\log p(\vec{x}|\theta)] = \arg\max_{\theta} Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

Expectation Maximization:
The EM algorithm

EM finds the values of the parameters and hidden data X that maximize the likelihood of the observed data Y.

1. Guess initial parameter values $\theta^{(m=0)}$

E-step

2. Calculate the distribution over the data $p(\vec{x}|\vec{y}, \theta^{(m)})$
3. Calculate the expected log probability for the data

$$
\begin{aligned}
Q(\theta|\theta^{(m)}) &\triangleq \mathbb{E}[\log p(\vec{x}|\theta)] \\
&= \sum_{\vec{x}} \log p(\vec{x}|\theta) p(\vec{x}|\vec{y}, \theta^{(m)})
\end{aligned}
$$

M-step

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$
\arg\max_{\theta} \mathbb{E}[\log p(\vec{x}|\theta)] = \arg\max_{\theta} Q(\theta|\theta^{(m)})
$$

5. Repeat steps 2-4 until convergence

EM for Toys:

1. Guess initial parameter values

EM for Toys:
1. Guess initial parameter values

- For toys, our only parameter is □, on which all our probabilities depend

EM for Toys:
1. Guess initial parameter values

- For toys, our only parameter is θ, on which all our probabilities depend

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

EM for Toys:
1. Guess initial parameter values

- For toys, our only parameter is ▯, on which all our probabilities depend

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- Making a good guess doesn't matter for this simple example

EM for Toys:
1. Guess initial parameter values

- For toys, our only parameter is , on which all our probabilities depend

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- Making a good guess doesn't matter for this simple example

- It matters *a lot* in more complex cases – EM will find the nearest local maximum to your initial guess

EM for Toys:
1. Guess initial parameter values

- For toys, our only parameter is , on which all our probabilities depend

$$\vec{q}_\theta = \left[\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

- Making a good guess doesn't matter for this simple example

- It matters *a lot* in more complex cases – EM will find the nearest local maximum to your initial guess

- We will try several initial values to see what happens

EM for Toys:

2. Calculate a distribution over the data

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- This is a great time to wonder what happened to Y – our actual observations

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- This is a great time to wonder what happened to Y – our actual observations

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- This is a great time to wonder what happened to Y – our actual observations

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- We will need our observations very soon, but bear with me a little longer

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- This is a great time to wonder what happened to Y – our actual observations

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- We will need our observations very soon, but bear with me a little longer

- Let’s get ready for a little math

EM for Toys:
2. Calculate a distribution over the data

- We already know how to calculate the probability of seeing a particular X:

$$p(\vec{x}|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- This is a great time to wonder what happened to Y – our actual observations

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- We will need our observations very soon, but bear with me a little longer

- Let’s get ready for a little math

EM for Toys:
3. Calculate the expected log probability

EM for Toys:
3. Calculate the expected log probability

- Working through the calculation:

EM for Toys:
3. Calculate the expected log probability

- Working through the calculation:

$$
\begin{aligned}
Q(\theta|\theta^{(m)}) &= \mathbb{E}[\log p(\vec{x}|\theta)] \\
&= \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\left[\log\left(\frac{n!}{\prod_{i=1}^{5} x_i!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2+x_5}\left(\frac{1-\theta}{4}\right)^{x_3+x_4}\right)\right] \\
&= \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\left[\log n! - \sum_{i=1}^{5}\log x_i! - x_1\log 2 + (x_2+x_5)\log\theta\right. \\
&\quad \left. - (x_2+x_5)\log 4 + (x_3+x_4)\log(1-\theta) - (x_3+x_4)\log 4\right]
\end{aligned}
$$

EM for Toys:
3. Calculate the expected log probability

- Working through the calculation:

$$
\begin{aligned}
Q(\theta|\theta^{(m)}) &= \mathbb{E}[\log p(\vec{x}|\theta)] \\
&= \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\left[\log\left(\frac{n!}{\prod_{i=1}^{5} x_i!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2+x_5}\left(\frac{1-\theta}{4}\right)^{x_3+x_4}\right)\right] \\
&= \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\left[\log n! - \sum_{i=1}^{5}\log x_i! - x_1 \log 2 + (x_2 + x_5)\log\theta\right. \\
&\left. - (x_2 + x_5)\log 4 + (x_3 + x_4)\log(1-\theta) - (x_3 + x_4)\log 4\right]
\end{aligned}
$$

- BUT we only want to find θ to maximize this expectation, not to calculate the maximum value. Let’s take out everything that doesn’t depend on θ.

EM for Toys:

3. Calculate the expected log probability

- Working through the calculation:

$$
\begin{aligned}
Q(\theta|\theta^{(m)}) &= \mathbb{E}[\log p(\vec{x}|\theta)] \\
&= \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\left[\log\left(\frac{n!}{\prod_{i=1}^{5} x_i!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2+x_5}\left(\frac{1-\theta}{4}\right)^{x_3+x_4}\right)\right] \\
&= \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\Bigg[\log n! - \sum_{i=1}^{5}\log x_i! - x_1\log 2 + (x_2+x_5)\log\theta \\
&\quad - (x_2+x_5)\log 4 + (x_3+x_4)\log(1-\theta) - (x_3+x_4)\log 4\Bigg]
\end{aligned}
$$

- BUT we only want to find □ to maximize this expectation, not to calculate the maximum value. Let’s take out everything that doesn’t depend on □.

$$
\begin{aligned}
\arg\max_{\theta\in(0,1)} Q(\theta|\theta^{(m)}) &= \arg\max_{\theta\in(0,1)} \mathbb{E}[\log p(\vec{x}|\theta)] \\
&\equiv \arg\max_{\theta\in(0,1)} \mathbb{E}_{\vec{x}|\vec{y},\theta^{(m)}}\left[(x_2+x_5)\log\theta + (x_3+x_4)\log(1-\theta)\right] \\
&= \arg\max_{\theta\in(0,1)} \left\{(\mathbb{E}[x_2]+\mathbb{E}[x_5])\log\theta + (\mathbb{E}[x_3]+\mathbb{E}[x_4])\log(1-\theta)\right\}
\end{aligned}
$$

I think we’ve earned a break.

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log \theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1 - \theta)\}$

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log \theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$
- Remember the observed data?

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log \theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$

- Remember the observed data?

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log \theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$

- Remember the observed data?

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- In order to get the expectations and tie us back to reality, we need to model the hidden data X in terms of the observed data Y. If we say the first two members of X are binomially distributed, given Y, then we have:

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta\in(0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log\theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$

- Remember the observed data?

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- In order to get the expectations and tie us back to reality, we need to model the hidden data X in terms of the observed data Y. If we say the first two members of X are binomially distributed, given Y, then we have:

$$p(\vec{x}|\vec{y},\theta) = \frac{y_1!}{x_1!x_2!}\left(\frac{2}{2+\theta}\right)^{x_1}\left(\frac{\theta}{2+\theta}\right)^{x_2}\mathbb{I}[x_1+x_2=y_1]\prod_{i=3}^{5}\mathbb{I}[x_i=y_{i-1}]$$

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log\theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$

- Remember the observed data?

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- In order to get the expectations and tie us back to reality, we need to model the hidden data X in terms of the observed data Y. If we say the first two members of X are binomially distributed, given Y, then we have:

$$p(\vec{x}|\vec{y},\theta) = \frac{y_1!}{x_1!x_2!}\left(\frac{2}{2+\theta}\right)^{x_1}\left(\frac{\theta}{2+\theta}\right)^{x_2}\mathbb{I}[x_1 + x_2 = y_1]\prod_{i=3}^{5}\mathbb{I}[x_i = y_{i-1}]$$

- Now we can get the expected values using the binomial mean:

EM for Toys:
3. Calculate the expected log probability (cont.)

- Our goal: $\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log\theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$

- Remember the observed data?

$$\vec{y} \triangleq [x_1 + x_2, x_3, x_4, x_5]$$

- In order to get the expectations and tie us back to reality, we need to model the hidden data X in terms of the observed data Y. If we say the first two members of X are binomially distributed, given Y, then we have:

$$p(\vec{x}|\vec{y},\theta) = \frac{y_1!}{x_1!x_2!}\left(\frac{2}{2+\theta}\right)^{x_1}\left(\frac{\theta}{2+\theta}\right)^{x_2} \mathbb{I}[x_1 + x_2 = y_1] \prod_{i=3}^{5} \mathbb{I}[x_i = y_{i-1}]$$

- Now we can get the expected values using the binomial mean:

$$\mathbb{E}_{\vec{x}|\vec{y},\theta}[\vec{x}] = \left[\frac{2}{2+\theta}y_1, \frac{\theta}{2+\theta}y_1, y_2, y_3, y_4\right]$$

EM for Toys:
4. Choose new parameters

EM for Toys:
4. Choose new parameters

- With everything we've learned, we can simplify our objective function:

EM for Toys:
4. Choose new parameters

- With everything we've learned, we can simplify our objective function:

$$\arg\max_{\theta \in (0,1)} \left\{ \left(\mathbb{E}[x_2] + \mathbb{E}[x_5] \right) \log \theta + \left(\mathbb{E}[x_3] + \mathbb{E}[x_4] \right) \log(1-\theta) \right\}$$

EM for Toys:
4. Choose new parameters

- With everything we've learned, we can simplify our objective function:

$$\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log \theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$$

$$= \arg\max_{\theta \in (0,1)} \left\{ \left(\frac{\theta}{2+\theta} y_1 + y_4 \right) \log \theta + (y_2 + y_3) \log(1-\theta) \right\}$$

EM for Toys:
4. Choose new parameters

- With everything we've learned, we can simplify our objective function:

$$\arg\max_{\theta \in (0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log \theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$$

$$= \arg\max_{\theta \in (0,1)} \left\{ \left(\frac{\theta}{2+\theta} y_1 + y_4 \right) \log \theta + (y_2 + y_3) \log(1-\theta) \right\}$$

$$= \frac{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_4}{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_2 + y_3 + y_4}$$

EM for Toys:
4. Choose new parameters

- With everything we've learned, we can simplify our objective function:

$$\arg\max_{\theta\in(0,1)} \{(\mathbb{E}[x_2] + \mathbb{E}[x_5]) \log\theta + (\mathbb{E}[x_3] + \mathbb{E}[x_4]) \log(1-\theta)\}$$

$$= \arg\max_{\theta\in(0,1)} \left\{ \left(\frac{\theta}{2+\theta} y_1 + y_4 \right) \log\theta + (y_2 + y_3) \log(1-\theta) \right\}$$

$$= \frac{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_4}{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_2 + y_3 + y_4}$$

- We're done! To find the value of that maximizes the expected log probability of Y, just run that single equation until it converges.

Let's look at some data

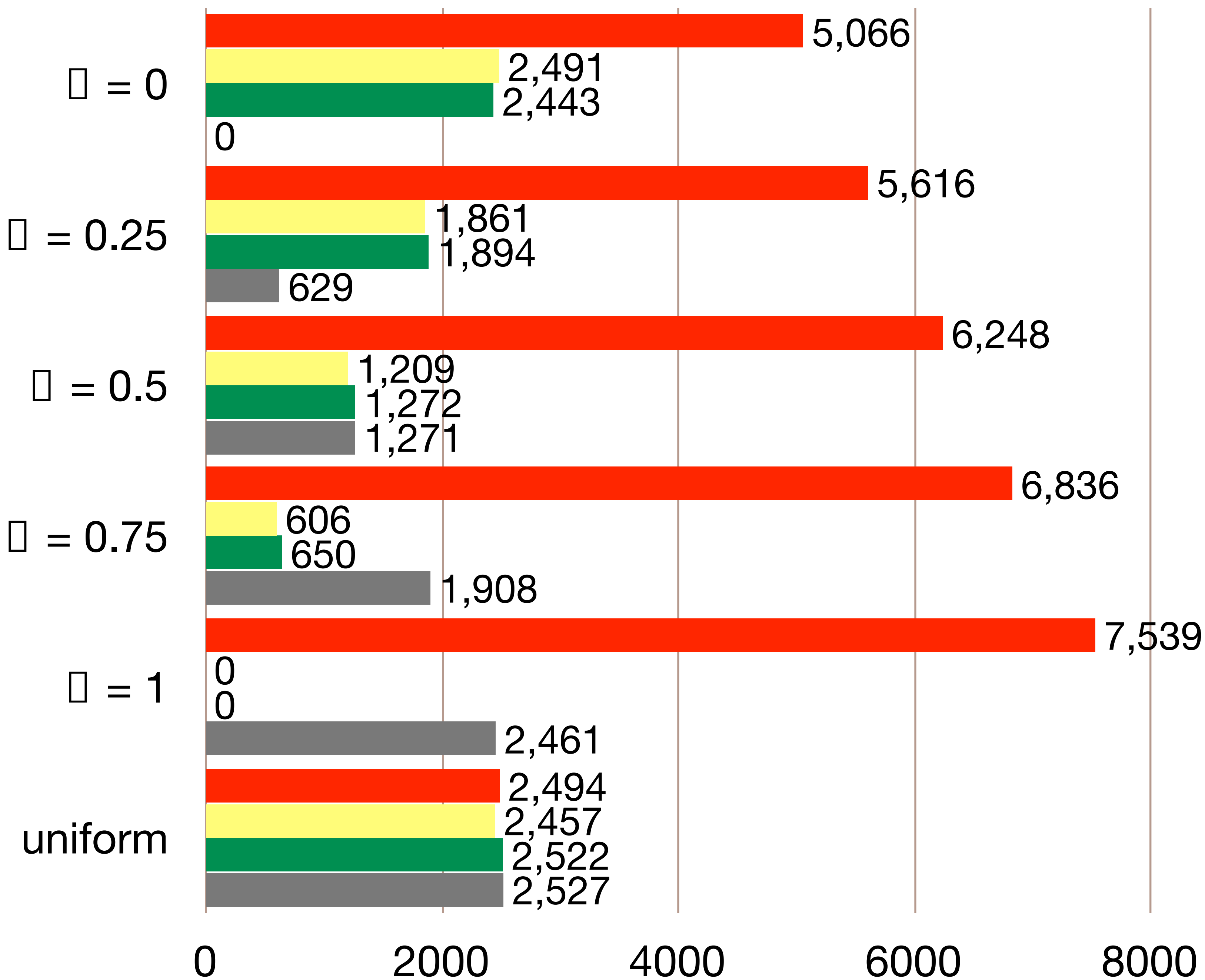
- Let's test this on fake data:

$$\theta \in \{0, 1/4, 1/2, 3/4, 1\}$$
$$n \in \{100, 1000, 10000\}$$

- Plus a uniform distribution, to see what happens when our model is wrong

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

Let’s look at some data



- Let’s test this on fake data:

$$\theta \in \{0, 1/4, 1/2, 3/4, 1\}$$

$$n \in \{100, 1000, 10000\}$$

- Plus a uniform distribution, to see what happens when our model is wrong

$$\vec{p}_\theta = \left[\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right]$$

```python
def run(self):

  # Initialize the observed histogram y and the first guess theta
  y = self.y
  theta = self.theta
  print ("Initial theta: {:0.6f}".format(theta))

  # Run up to some maximum number of rounds
  for round in range(1, self.max_rounds + 1):

    # Calculate the new parameter estimate for this round
    new_theta = (((theta / (2 + theta)) * y[0] + y[3]) /
                 ((theta / (2 + theta)) * y[0] + y[3] + y[2] + y[1]))
    delta = new_theta - theta
    theta = new_theta

    # Print our status and check for convergence
    print ("Round {} theta: {:0.9f} diff: {:0.3e}".format(round, theta, delta))
    if abs(delta) < 1e-12:
      print("Converged!")
      return
```

EM for Toys

Python implementation

```
Initial theta: 0.300000
Round 1 theta: 0.726310044 diff: 4.263e-01
Round 2 theta: 0.778614638 diff: 5.230e-02
Round 3 theta: 0.782829617 diff: 4.215e-03
Round 4 theta: 0.783155558 diff: 3.259e-04
Round 5 theta: 0.783180681 diff: 2.512e-05
Round 6 theta: 0.783182617 diff: 1.936e-06
Round 7 theta: 0.783182766 diff: 1.492e-07
Round 8 theta: 0.783182777 diff: 1.150e-08
Round 9 theta: 0.783182778 diff: 8.858e-10
Round 10 theta: 0.783182778 diff: 6.826e-11
Round 11 theta: 0.783182778 diff: 5.260e-12
Round 12 theta: 0.783182778 diff: 4.053e-13
Converged!
Theta: 0.7832
Predicted toy probs: [0.6958, 0.0542, 0.0542, 0.1958]
Empirical toy probs: [0.6920, 0.0430, 0.0660, 0.1990]
Y: [692, 43, 66, 199]
E[X]: [497.27, 194.73, 43, 66, 199]
KL(empirical||predicted): 0.002483
```

Example EM Output

□ = 0.75; n = 1,000; Guess = 0.3

```
Initial theta: 0.750000
Round 1 theta: 0.780563690 diff: 3.056e-02
Round 2 theta: 0.782980580 diff: 2.417e-03
Round 3 theta: 0.783167195 diff: 1.866e-04
Round 4 theta: 0.783181577 diff: 1.438e-05
Round 5 theta: 0.783182686 diff: 1.108e-06
Round 6 theta: 0.783182771 diff: 8.540e-08
Round 7 theta: 0.783182778 diff: 6.581e-09
Round 8 theta: 0.783182778 diff: 5.071e-10
Round 9 theta: 0.783182778 diff: 3.908e-11
Round 10 theta: 0.783182778 diff: 3.011e-12
Round 11 theta: 0.783182778 diff: 2.320e-13
Converged!
Theta: 0.7832
Predicted toy probs: [0.6958, 0.0542, 0.0542, 0.1958]
Empirical toy probs: [0.6920, 0.0430, 0.0660, 0.1990]
Y: [692, 43, 66, 199]
E[X]: [497.27, 194.73, 43, 66, 199]
KL(empirical||predicted): 0.002483
```

Example EM Output

 = 0.75; n = 1,000; Guess = 0.75

```
Initial theta: 0.250000
Round 1 theta: 0.331221198 diff: 8.122e-02
Round 2 theta: 0.338049688 diff: 6.828e-03
Round 3 theta: 0.338596066 diff: 5.464e-04
Round 4 theta: 0.338639608 diff: 4.354e-05
Round 5 theta: 0.338643076 diff: 3.469e-06
Round 6 theta: 0.338643353 diff: 2.763e-07
Round 7 theta: 0.338643375 diff: 2.201e-08
Round 8 theta: 0.338643377 diff: 1.754e-09
Round 9 theta: 0.338643377 diff: 1.397e-10
Round 10 theta: 0.338643377 diff: 1.113e-11
Round 11 theta: 0.338643377 diff: 8.866e-13
Converged!
Theta: 0.3386
Predicted toy probs: [0.5847, 0.1653, 0.1653, 0.0847]
Empirical toy probs: [0.2570, 0.2330, 0.2830, 0.2270]
Y: [257, 233, 283, 227]
E[X]: [219.79, 37.21, 233, 283, 227]
KL(empirical||predicted): 0.244673
```

Example EM Output

uniform; n = 1,000; Guess = 0.25

What is the data telling us?

- EM is finding the local maximum closest to the initialization point

What is the data telling us?

- EM is finding the local maximum closest to the initialization point
- If we initialize to the “right answer,” it will move away from that to the maximum for the observed data

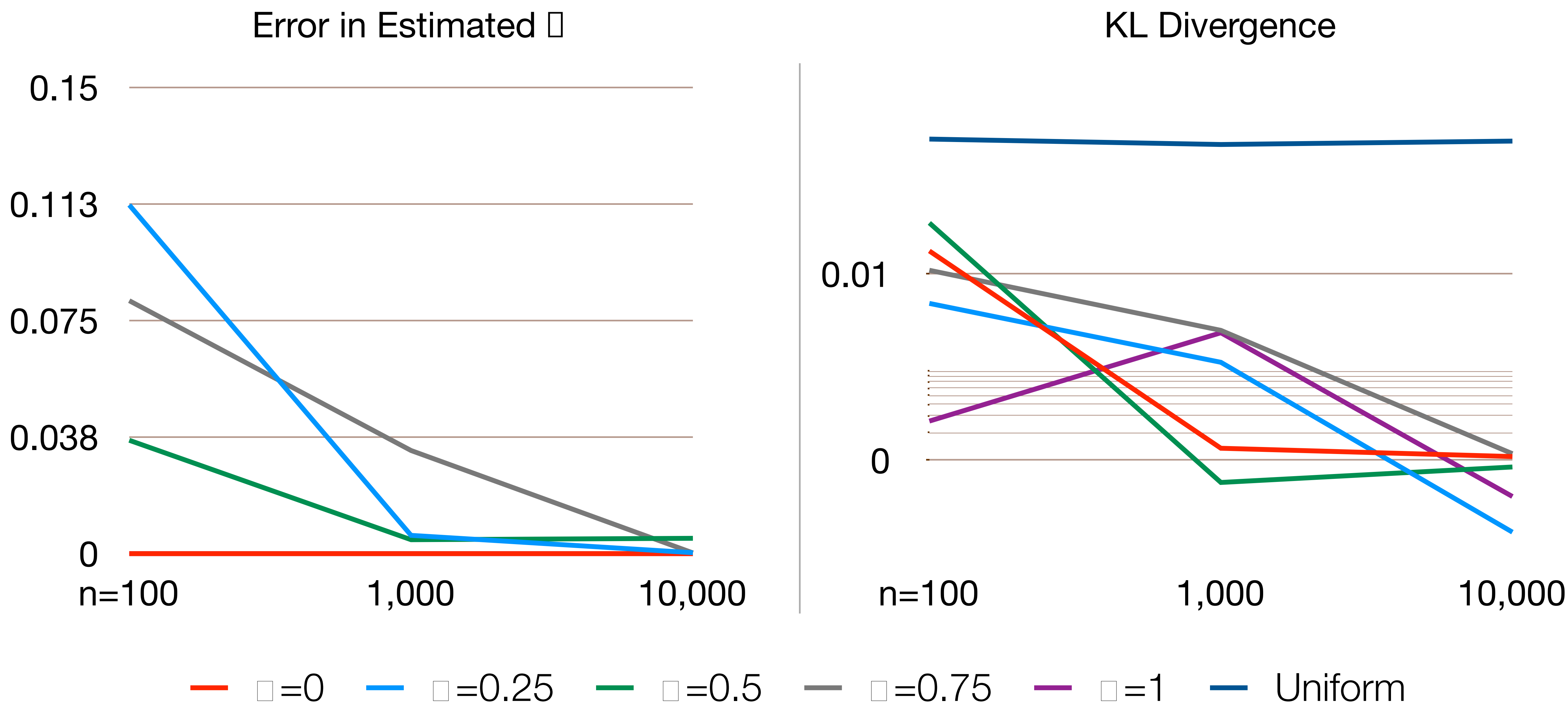
What is the data telling us?

- EM is finding the local maximum closest to the initialization point
- If we initialize to the “right answer,” it will move away from that to the maximum for the observed data
- EM can’t fix a bad model: if your modeling assumptions are bad, it will find the best answer *consistent with those assumptions*

What is the data telling us?

- EM is finding the local maximum closest to the initialization point
- If we initialize to the “right answer,” it will move away from that to the maximum for the observed data
- EM can’t fix a bad model: if your modeling assumptions are bad, it will find the best answer *consistent with those assumptions*
- As you’d expect, EM is also sensitive to the amount of data you give it

Results of all data runs for Toys



Results of all data runs for Toys

So what?

The EM Algorithm:
a second look

Let's think about how to do this in general.

1. Guess initial parameter values $\theta^{(m=0)}$
2. Calculate the distribution over the data $p(\vec{x}|\vec{y},\theta^{(m)})$
3. Calculate the expected log probability for the data

$$\begin{aligned} Q(\theta|\theta^{(m)}) &\triangleq \mathbb{E}[\log p(\vec{x}|\theta)] \\ &= \sum_{\vec{x}} \log p(\vec{x}|\theta) p(\vec{x}|\vec{y},\theta^{(m)}) \end{aligned}$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_\theta \mathbb{E}[\log p(\vec{x}|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

E-step

M-step

The EM Algorithm: a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

E-step

M-step

1. Guess initial parameter values $\theta^{(m=0)}$

2. Calculate the distribution over the data $p(\vec{x}|\vec{y},\theta^{(m)})$

3. Calculate the expected log probability for the data

$$Q(\theta|\theta^{(m)}) \triangleq \mathbb{E}[\log p(\vec{x}|\theta)]$$
$$= \sum_{\vec{x}} \log p(\vec{x}|\theta) p(\vec{x}|\vec{y},\theta^{(m)})$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_{\theta} \mathbb{E}[\log p(\vec{x}|\theta)] = \arg\max_{\theta} Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

The EM Algorithm: a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

E-step

M-step

1. Guess initial parameter values $\theta^{(m=0)}$
2. Calculate the distribution over the data $p(x|y, \theta^{(m)})$
3. Calculate the expected log probability for the data

$$Q(\theta|\theta^{(m)}) \triangleq \mathbb{E}[\log p(x|\theta)]$$
$$= \sum_x \log p(x|\theta) p(x|y, \theta^{(m)})$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_\theta \mathbb{E}[\log p(x|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

The EM Algorithm: a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

Variables

E-step

M-step

1. Guess initial parameter values $\theta^{(m=0)}$

2. Calculate the distribution over the data $p(x|y, \theta^{(m)})$

3. Calculate the expected log probability for the data

$$Q(\theta|\theta^{(m)}) \triangleq \mathbb{E}[\log p(x|\theta)]$$
$$= \sum_x \log p(x|\theta) p(x|y, \theta^{(m)})$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_\theta \mathbb{E}[\log p(x|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

The EM Algorithm:
a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

Variables

1. Guess initial parameter values $\theta^{(m=0)}$ $\theta \in \Theta$

2. Calculate the distribution over the data $p(x|y,\theta^{(m)})$

3. Calculate the expected log probability for the data

$$Q(\theta|\theta^{(m)}) \triangleq \mathbb{E}[\log p(x|\theta)] = \sum_x \log p(x|\theta) p(x|y,\theta^{(m)})$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_\theta \mathbb{E}[\log p(x|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

E-step (steps 2–3)

M-step (step 4)

The EM Algorithm: a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

Variables

$\theta \in \Theta$

$y, Y \in \mathbb{R}^{d_1}$

E-step

M-step

1. Guess initial parameter values $\theta^{(m=0)}$
2. Calculate the distribution over the data $p(x|y, \theta^{(m)})$
3. Calculate the expected log probability for the data

$$
\begin{aligned}
Q(\theta|\theta^{(m)}) &\triangleq \mathbb{E}[\log p(x|\theta)] \\
&= \sum_x \log p(x|\theta) p(x|y, \theta^{(m)})
\end{aligned}
$$

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$
\arg\max_\theta \mathbb{E}[\log p(x|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})
$$

5. Repeat steps 2-4 until convergence

The EM Algorithm: a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

Variables

$\theta \in \Theta$

$y, Y \in \mathbb{R}^{d_1}$

$z, Z \in \mathbb{R}^{d_2}$

E-step

1. Guess initial parameter values $\theta^{(m=0)}$
2. Calculate the distribution over the data $p(x|y, \theta^{(m)})$
3. Calculate the expected log probability for the data

$$Q(\theta|\theta^{(m)}) \triangleq \mathbb{E}[\log p(x|\theta)]$$
$$= \sum_x \log p(x|\theta) p(x|y, \theta^{(m)})$$

M-step

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_\theta \mathbb{E}[\log p(x|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

The EM Algorithm:
a second look

Let's think about how to do this in general.

To start with, let's allow X and Y to be anything.

Variables

$\theta \in \Theta$

$y, Y \in \mathbb{R}^{d_1}$

$z, Z \in \mathbb{R}^{d_2}$

$x \triangleq (y, z)$

$X \triangleq (Y, Z)$

E-step

1. Guess initial parameter values $\theta^{(m=0)}$
2. Calculate the distribution over the data $p(x|y, \theta^{(m)})$
3. Calculate the expected log probability for the data

$$Q(\theta|\theta^{(m)}) \triangleq \mathbb{E}[\log p(x|\theta)] = \sum_x \log p(x|\theta) p(x|y, \theta^{(m)})$$

M-step

4. Choose new parameter values to maximize $Q(\theta|\theta^{(m)})$

$$\arg\max_\theta \mathbb{E}[\log p(x|\theta)] = \arg\max_\theta Q(\theta|\theta^{(m)})$$

5. Repeat steps 2-4 until convergence

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
	Model for observations, given params
	Model for complete data in one round

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y\|\theta)$	Model for observations, given params
	Model for complete data in one round

The EM Algorithm: known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y \vert \theta)$	Model for observations, given params

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y \vert \theta)$	Model for observations, given params
$p(X = x \vert y, \theta)$	

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y \mid \theta)$	Model for observations, given params
$p(X = x \mid y, \theta)$	Model for complete data in one round

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y \vert \theta)$	Model for observations, given params
$p(X = x \vert y, \theta)$	Model for complete data in one round
$\mathbb{E}\left[p(X = x \vert y, \theta)\right]$	

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y \mid \theta)$	Model for observations, given params
$p(X = x \mid y, \theta)$	Model for complete data in one round
$\mathbb{E}\left[p(X = x \mid y, \theta)\right]$	

The EM Algorithm:
known unknowns

Variables	Meaning
$\theta \in \Theta$	Parameters (unknown)
$y, Y \in \mathbb{R}^{d_1}$	Observed data and R.V. (known)
$z, Z \in \mathbb{R}^{d_2}$	Hidden data and R.V. (unknown)
$x \triangleq (y, z)\ X \triangleq (Y, Z)$	Complete data and R.V.
$p(Y = y \vert \theta)$	Model for observations, given params
$p(X = x \vert y, \theta)$	Model for complete data in one round
$\mathbb{E}\left[p(X = x \vert y, \theta)\right]$	$\int_{x:p(x \vert y,\theta)>0} xp(X = x \vert y, \theta)dx$

The EM Algorithm: the tricky part

$$\int_{x:p(x|y,\theta)>0} xp(X = x|y,\theta)dx \quad \sum_{x \in X} xp(X = x|y,\theta)$$

- How do we maximize this?

The EM Algorithm: the tricky part

$$\int_{x:p(x|y,\theta)>0} xp(X = x|y,\theta)dx \quad \sum_{x \in X} xp(X = x|y,\theta)$$

- How do we maximize this?
- It depends on what's hiding inside your model

The EM Algorithm: the tricky part

$$\int_{x:p(x|y,\theta)>0} xp(X = x|y,\theta)dx \quad \sum_{x \in X} xp(X = x|y,\theta)$$

- How do we maximize this?
- It depends on what's hiding inside your model
- Toys has a discrete model; we solved it algebraically

The EM Algorithm: the tricky part

$$\int_{x:p(x|y,\theta)>0} xp(X = x|y,\theta)dx \quad \sum_{x \in X} xp(X = x|y,\theta)$$

- How do we maximize this?
- It depends on what's hiding inside your model
- Toys has a discrete model; we solved it algebraically
- You typically differentiate, set it to zero, and solve

That’s all for now!

That's all for now!

Coming up:

- Proof of convergence
- Actually useful models
- An information theoretical look
- 100% fewer birds