4CSLL5
Parameter Estimation (Supervised and Unsupervised)

Martin Emms

September 26, 2018

The EM Algorithm

Outline

Hidden variable variant

Suppose you no longer see the outcome of Z; you still see the tosses of the chosen coin, but you can't tell which it is.

Hidden variable variant

Suppose you no longer see the outcome of Z; you still see the tosses of the chosen coin, but you can't tell which it is.

The data now looks like this

d	Z					**X**: tosses of chosen coin						H counts
1	?	H	H	H	H	H	H	H	H	T	T	(8H)
2	?	T	T	H	T	T	T	H	T	T	T	(2H)
3	?	H	T	H	H	T	H	H	H	H	T	(7H)
4	?	H	T	H	H	H	T	H	H	H	H	(8H)
5	?	T	T	T	T	T	T	H	T	T	T	(1H)
6	?	H	H	T	H	H	H	H	H	H	H	(9H)
7	?	T	H	H	T	H	H	H	H	H	T	(7H)
8	?	H	H	H	H	H	H	T	H	H	H	(9H)
9	?	H	H	T	T	T	T	T	H	T	T	(3H)

Hidden variable variant

Suppose you no longer see the outcome of Z; you still see the tosses of the chosen coin, but you can't tell which it is.

The data now looks like this

d	Z					**X**: tosses of chosen coin						H counts
1	?	H	H	H	H	H	H	H	H	T	T	(8H)
2	?	T	T	H	T	T	T	H	T	T	T	(2H)
3	?	H	T	H	H	T	H	H	H	H	T	(7H)
4	?	H	T	H	H	H	T	H	H	H	H	(8H)
5	?	T	T	T	T	T	T	H	T	T	T	(1H)
6	?	H	H	T	H	H	H	H	H	H	H	(9H)
7	?	T	H	H	T	H	H	H	H	H	T	(7H)
8	?	H	H	H	H	H	H	T	H	H	H	(9H)
9	?	H	H	T	T	T	T	T	H	T	T	(3H)

Z is so-called hidden variable in each case

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

For each d we just know the coin-tosses $\mathbf{X}^d$. Their probability is now a **sum**

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

For each d we just know the coin-tosses $\mathbf{X}^d$. Their probability is now a **sum**

$$p(\mathbf{X}^d) = p(Z=a)p(\mathbf{X}^{\mathbf{d}}|Z=a) + p(Z=b)p(\mathbf{X}^{\mathbf{d}}|Z=b)$$

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

For each d we just know the coin-tosses $\mathbf{X}^d$. Their probability is now a **sum**

$$
\begin{aligned}
p(\mathbf{X}^d) &= p(Z=a)p(\mathbf{X^d}|Z=a) + p(Z=b)p(\mathbf{X^d}|Z=b) \\
&= \theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)}
\end{aligned}
$$

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

For each d we just know the coin-tosses $\mathbf{X}^d$. Their probability is now a **sum**

$$\begin{aligned} p(\mathbf{X}^d) &= p(Z=a)p(\mathbf{X}^{\mathbf{d}}|Z=a) + p(Z=b)p(\mathbf{X}^{\mathbf{d}}|Z=b) \\ &= \theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)} \end{aligned}$$

and the entire data set's probability, $p(\mathbf{d})$ is the product:

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

For each d we just know the coin-tosses $\mathbf{X}^d$. Their probability is now a **sum**

$$\begin{aligned} p(\mathbf{X}^d) &= p(Z=a)p(\mathbf{X^d}|Z=a) + p(Z=b)p(\mathbf{X^d}|Z=b) \\ &= \theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)} \end{aligned}$$

and the entire data set's probability, $p(\mathbf{d})$ is the product:

$$\begin{aligned} p(\mathbf{d}) &= \prod_d p(\mathbf{X}^d) \qquad (11) \\ &= \end{aligned}$$

We still have the probability model for combinations $(Z, \mathbf{X})$, with the same parameters θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

We would still like to find values for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ which again **maximise** the probability of the observed data

For each d we just know the coin-tosses $\mathbf{X}^d$. Their probability is now a **sum**

$$\begin{aligned} p(\mathbf{X}^d) &= p(Z=a)p(\mathbf{X^d}|Z=a) + p(Z=b)p(\mathbf{X^d}|Z=b) \\ &= \theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)} \end{aligned}$$

and the entire data set's probability, $p(\mathbf{d})$ is the product:

$$p(\mathbf{d}) = \prod_d p(\mathbf{X}^d) \tag{11}$$

$$= \prod_d \left[\theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)}\right] \tag{12}$$

The 'product of sums' problem

$$p(\mathbf{d}) = \prod_{d} \left[\theta_a \theta_{h|a}^{\#(d,h)} (1 - \theta_{h|a})^{\#(d,t)} + (1 - \theta_a) \theta_{h|b}^{\#(d,h)} (1 - \theta_{h|b})^{\#(d,t)} \right]$$

so can we maximise (12), repeated above?

The 'product of sums' problem

$$p(\mathbf{d}) = \prod_{d} \left[\theta_a \theta_{h|a}^{\#(d,h)} (1 - \theta_{h|a})^{\#(d,t)} + (1 - \theta_a) \theta_{h|b}^{\#(d,h)} (1 - \theta_{h|b})^{\#(d,t)} \right]$$

so can we maximise (12), repeated above?

the preceding procedure of taking logs runs into a dead-end, because $p(\mathbf{d})$ is no longer all products, turning into sums.

The 'product of sums' problem

$$p(\mathbf{d}) = \prod_{d} \left[\theta_a \theta_{h|a}^{\#(d,h)} (1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)} (1-\theta_{h|b})^{\#(d,t)} \right]$$

so can we maximise (12), repeated above?

the preceding procedure of taking logs runs into a dead-end, because $p(\mathbf{d})$ is no longer all products, turning into sums. Instead the log is

$$\sum_{d} log \left[\theta_a \theta_{h|a}^{\#(d,h)} (1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)} (1-\theta_{h|b})^{\#(d,t)} \right]$$

and there is no known way to cleverly break this down as there was before

The 'product of sums' problem

$$p(\mathbf{d}) = \prod_d \left[\theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)}\right]$$

so can we maximise (12), repeated above?

the preceding procedure of taking logs runs into a dead-end, because $p(\mathbf{d})$ is no longer all products, turning into sums. Instead the log is

$$\sum_d log \left[\theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)}\right]$$

and there is no known way to cleverly break this down as there was before

this is essentially the problem we face if we want to do parameter estimation with hidden variables – this is done widely in eg. Machine Translation and Speech Recognition.

The 'product of sums' problem

$$p(\mathbf{d}) = \prod_d \left[\theta_a \theta_{h|a}^{\#(d,h)} (1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)} (1-\theta_{h|b})^{\#(d,t)}\right]$$

so can we maximise (12), repeated above?

the preceding procedure of taking logs runs into a dead-end, because $p(\mathbf{d})$ is no longer all products, turning into sums. Instead the log is

$$\sum_d log \left[\theta_a \theta_{h|a}^{\#(d,h)} (1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)} (1-\theta_{h|b})^{\#(d,t)}\right]$$

and there is no known way to cleverly break this down as there was before

this is essentially the problem we face if we want to do parameter estimation with hidden variables – this is done widely in eg. Machine Translation and Speech Recognition. The EM or 'Expectation Maximisation' algorithm will turn out to be the solution

The general hidden variable set-up

before proceeding lets try to make clear the general case of a hidden variable problem

The general hidden variable set-up

before proceeding lets try to make clear the general case of a hidden variable problem

You have D data items

The general hidden variable set-up

before proceeding lets try to make clear the general case of a hidden variable problem

You have D data items

In the fully observed case, each data item d is represented by the values of a set of variables, which we'll split into two sets $\langle \mathbf{z}^d, \mathbf{x}^d \rangle$

The general hidden variable set-up

before proceeding lets try to make clear the general case of a hidden variable problem

You have D data items

In the fully observed case, each data item d is represented by the values of a set of variables, which we'll split into two sets $\langle \mathbf{z}^d, \mathbf{x}^d \rangle$

and you have a probability model – ie. formula – spelling how likely any such fully observed case is $P(\langle \mathbf{z}^d, \mathbf{x}^d \rangle; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are all the *parameters* of the model

The general hidden variable set-up

before proceeding lets try to make clear the general case of a hidden variable problem

You have D data items

In the fully observed case, each data item d is represented by the values of a set of variables, which we'll split into two sets $\langle \mathbf{z}^d, \mathbf{x}^d \rangle$

and you have a probability model – ie. formula – spelling how likely any such fully observed case is $P(\langle \mathbf{z}^d, \mathbf{x}^d \rangle; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are all the *parameters* of the model

In the hidden case, for each data item d you just have values on a subset of the variables $\mathbf{x}^d$; the other variables $\mathbf{z}^d$ are *hidden*

The general hidden variable set-up

before proceeding lets try to make clear the general case of a hidden variable problem

You have D data items

In the fully observed case, each data item d is represented by the values of a set of variables, which we'll split into two sets $\langle \mathbf{z}^d, \mathbf{x}^d \rangle$

and you have a probability model – ie. formula – spelling how likely any such fully observed case is $P(\langle \mathbf{z}^d, \mathbf{x}^d \rangle; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are all the *parameters* of the model

In the hidden case, for each data item d you just have values on a subset of the variables $\mathbf{x}^d$; the other variables $\mathbf{z}^d$ are *hidden*

If $\mathcal{A}(\mathbf{z})$ represents the space of all possible values for the variables $\mathbf{z}$, then the probability of each partial data item is

$$P(\mathbf{x}^d; \boldsymbol{\theta}) = \sum_{\mathbf{k} \in \mathcal{A}(\mathbf{z})} P(\mathbf{z} = \mathbf{k}, \mathbf{x}^d \ ; \boldsymbol{\theta})$$

Outline

Taking stock: what kinds of thing *can* we calculate?

Taking stock: what kinds of thing *can* we calculate?

- **parameters given visible data**:

Taking stock: what kinds of thing *can* we calculate?

- **parameters given visible data**: we have seen illustrations where **z** is known for each datum, and where finding parameter values maximising the data likelihood was easy: its relative frequencies all the way *(scenario 1: 1 vis var; scenario 2: 2 vis vars, one for coin-choice, and one for the coin-tosses on whatever coin was chosen). In fact to do the parameter estimation we really just needed* numbers *about how often types of outcomes occurred.*

Taking stock: what kinds of thing *can* we calculate?

- **parameters given visible data**: we have seen illustrations where $\mathbf{z}$ is known for each datum, and where finding parameter values maximising the data likelihood was easy: its relative frequencies all the way *(scenario 1: 1 vis var; scenario 2: 2 vis vars, one for coin-choice, and one for the coin-tosses on whatever coin was chosen). In fact to do the parameter estimation we really just needed* numbers *about how often types of outcomes occurred.*
- **posterior probs on hidden vars**: if we have all the *parameters* $\boldsymbol{\theta}$, for datum d we can 'easily' work out $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d; \boldsymbol{\theta})$. In our third scenario where the coin choice was hidden, for $Z = a$ the formula is

$$P(Z = a|\mathbf{X}^d; \theta_a, \theta_{h|a}, \theta_{h|b}) = \frac{\theta_a \theta_{h|a}^{\#(d,h)} (1 - \theta_{h|a})^{\#(d,t)}}{\theta_a \theta_{h|a}^{\#(d,h)} (1 - \theta_{h|a})^{\#(d,t)} + (1 - \theta_a) \theta_{h|b}^{\#(d,h)} (1 - \theta_{h|b})^{\#(d,t)}}$$

Taking stock: what kinds of thing *can* we calculate?

- **parameters given visible data**: we have seen illustrations where $\mathbf{z}$ is known for each datum, and where finding parameter values maximising the data likelihood was easy: its relative frequencies all the way *(scenario 1: 1 vis var; scenario 2: 2 vis vars, one for coin-choice, and one for the coin-tosses on whatever coin was chosen). In fact to do the parameter estimation we really just needed* numbers *about how often types of outcomes occurred.*
- **posterior probs on hidden vars**: if we have all the *parameters* $\boldsymbol{\theta}$, for datum d we can 'easily' work out $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d; \boldsymbol{\theta})$. In our third scenario where the coin choice was hidden, for $Z = a$ the formula is

$$P(Z = a|\mathbf{X}^d; \theta_a, \theta_{h|a}, \theta_{h|b}) = \frac{\theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)}}{\theta_a \theta_{h|a}^{\#(d,h)}(1-\theta_{h|a})^{\#(d,t)} + (1-\theta_a)\theta_{h|b}^{\#(d,h)}(1-\theta_{h|b})^{\#(d,t)}}$$

- EM methods put those two abilities to use in iterative procedures to re-estimate parameters

EM sketch

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**:

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$;

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus;

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data.

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**:

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$;

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pretend these $\gamma_d(\mathbf{k})$ are counts in a virtual corpus of completions of d;

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pretend these $\gamma_d(\mathbf{k})$ are counts in a virtual corpus of completions of d; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data.

EM sketch

Let's use the notation $\gamma_d(\mathbf{k})$ for $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$ – which is something of a convention in EM methods

we will describe EM for the moment as just a kind of procedure or recipe. Later we will consider how to show that the procedure does something sensible.

- **Viterbi EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pick the best $\mathbf{z} = \mathbf{k}$ and 'complete' d with this value for $\mathbf{z}$ making a virtual complete corpus; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using *some* values for $\boldsymbol{\theta}$, for each d work out $\gamma_d(\mathbf{k})$ for each value $\mathbf{k} \in \mathcal{A}(\mathbf{z})$; (ii) pretend these $\gamma_d(\mathbf{k})$ are counts in a virtual corpus of completions of d; (iii) *re-estimate* $\boldsymbol{\theta}$ on this virtual data. If you back to (i) and do this over and over again you would be doing what is called **EM**

EM sketch specific for scenario 3 (hidden coin choice)

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**:

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$;

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus;

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data.

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**:

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$;

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pretend these $\gamma_d(k)$ are counts in a virtual corpus of completions of d;

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pretend these $\gamma_d(k)$ are counts in a virtual corpus of completions of d; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data.

EM sketch specific for scenario 3 (hidden coin choice)

- **Viterbi EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pick the best $Z = k$ and 'complete' d with this value for Z making a virtual complete corpus; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you go back to (i) and do this over and over again you would be doing what is called **Viterbi EM**
- **real EM**: (i) using some values for $\theta_a, \theta_{h|a}, \theta_{h|b}$, for each d work out $\gamma_d(k)$ for each value $k \in \{a, b\}$; (ii) pretend these $\gamma_d(k)$ are counts in a virtual corpus of completions of d; (iii) *re-estimate* $\theta_a, \theta_{h|a}, \theta_{h|b}$ on this virtual data. If you back to (i) and do this over and over again you would be doing what is called **EM**

The EM algorithm

The EM algorithm is a parameter (re)-estimation procedure, which starting from some original setting of parameters $\boldsymbol{\theta}^0$, generates a converging sequence of re-estimates:

$$\boldsymbol{\theta}^0 \rightarrow \ldots \rightarrow \boldsymbol{\theta}^n \rightarrow \boldsymbol{\theta}^{n+1} \rightarrow \ldots \rightarrow \boldsymbol{\theta}^{final}$$

where each $\boldsymbol{\theta}^n$ goes to $\boldsymbol{\theta}^{n+1}$ by a so-called **E**-step, followed by a **M** step:

The EM algorithm

The EM algorithm is a parameter (re)-estimation procedure, which starting from some original setting of parameters $\boldsymbol{\theta}^0$, generates a converging sequence of re-estimates:

$$\boldsymbol{\theta}^0 \rightarrow \ldots \rightarrow \boldsymbol{\theta}^n \rightarrow \boldsymbol{\theta}^{n+1} \rightarrow \ldots \rightarrow \boldsymbol{\theta}^{final}$$

where each $\boldsymbol{\theta}^n$ goes to $\boldsymbol{\theta}^{n+1}$ by a so-called **E**-step, followed by a **M** step:

E step

generate a virtual complete data corpus by treating each incomplete data item $(\mathbf{x}^d)$ *as standing for* all *possible completions with values for* $\mathbf{z}$, $(\mathbf{z} = \mathbf{k}, \mathbf{x}^d)$, *weighting* each *by its conditional probability* $P(\mathbf{z} = \mathbf{k} | \mathbf{x}^d; \boldsymbol{\theta}^n)$, *under current parameters* $\boldsymbol{\theta}^n$*: often this quantity is called the* responsibility.

The EM algorithm

The EM algorithm is a parameter (re)-estimation procedure, which starting from some original setting of parameters $\boldsymbol{\theta}^0$, generates a converging sequence of re-estimates:

$$\boldsymbol{\theta}^0 \rightarrow \ldots \rightarrow \boldsymbol{\theta}^n \rightarrow \boldsymbol{\theta}^{n+1} \rightarrow \ldots \rightarrow \boldsymbol{\theta}^{final}$$

where each $\boldsymbol{\theta}^n$ goes to $\boldsymbol{\theta}^{n+1}$ by a so-called **E**-step, followed by a **M** step:

E step

generate a virtual complete data corpus by treating each incomplete data item $(\mathbf{x}^d)$ *as standing for* all *possible completions with values for* $\mathbf{z}$, $(\mathbf{z} = \mathbf{k}, \mathbf{x}^d)$, *weighting* each *by its conditional probability* $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d; \boldsymbol{\theta}^n)$, *under current parameters* $\boldsymbol{\theta}^n$*: often this quantity is called the* responsibility. *Use* $\gamma_d(\mathbf{k})$ *for* $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$.

The EM algorithm

The EM algorithm is a parameter (re)-estimation procedure, which starting from some original setting of parameters $\boldsymbol{\theta}^0$, generates a converging sequence of re-estimates:

$$\boldsymbol{\theta}^0 \rightarrow \ldots \rightarrow \boldsymbol{\theta}^n \rightarrow \boldsymbol{\theta}^{n+1} \rightarrow \ldots \rightarrow \boldsymbol{\theta}^{final}$$

where each $\boldsymbol{\theta}^n$ goes to $\boldsymbol{\theta}^{n+1}$ by a so-called **E**-step, followed by a **M** step:

E step

generate a virtual complete data corpus by treating each incomplete data item $(\mathbf{x}^d)$ *as standing for* all *possible completions with values for* $\mathbf{z}$, $(\mathbf{z} = \mathbf{k}, \mathbf{x}^d)$, *weighting* each *by its conditional probability* $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d; \boldsymbol{\theta}^n)$, *under current parameters* $\boldsymbol{\theta}^n$*: often this quantity is called the* responsibility. *Use* $\gamma_d(\mathbf{k})$ *for* $P(\mathbf{z} = \mathbf{k}|\mathbf{x}^d)$.

M step

treating the 'responsibilities' $\gamma_d(k)$ *as if they were counts, apply maximum likelihood estimation to the virtual corpus to derive new estimates* $\boldsymbol{\theta}^{n+1}$.

The E step gives weighted guesses, $\gamma_d(k)$, for each way of completing each data point. These $\gamma_d(k)$ are then treated as counts of virtual completed data, so each data point x^d is split into virtual population

The E step gives weighted guesses, $\gamma_d(k)$, for each way of completing each data point. These $\gamma_d(k)$ are then treated as counts of virtual completed data, so each data point x^d is split into virtual population

$$x^d \begin{cases} \text{virtual data} & \text{virtual 'count'} \\ (z = 1, \mathbf{x}^d) & \gamma_d(1) \\ \vdots & \vdots \\ (z = k, \mathbf{x}^d) & \gamma_d(k) \end{cases}$$

recall the observed data for our 3rd scenario, with coin-choice hidden:

d	Z	**X**: tosses of chosen coin										H counts
1	?	H	H	H	H	H	H	H	H	T	T	(8H)
2	?	T	T	H	T	T	T	H	T	T	T	(2H)
3	?	H	T	H	H	T	H	H	H	H	T	(7H)
4	?	H	T	H	H	H	T	H	H	H	H	(8H)
5	?	T	T	T	T	T	T	H	T	T	T	(1H)
6	?	H	H	T	H	H	H	H	H	H	H	(9H)
7	?	T	H	H	T	H	H	H	H	H	T	(7H)
8	?	H	H	H	H	H	H	T	H	H	H	(9H)
9	?	H	H	T	T	T	T	T	H	T	T	(3H)

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$\mathbf{X^1}$: $(8H, 2T)$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X^1} : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X^1} : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X^6} : (9H, 1T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X^1} : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X^6} : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \ \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \ \gamma_6(b) = 0.08 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X^1} : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \qquad \mathbf{X^6} : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X^2} : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.66 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \ \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \ \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \; \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \; \gamma_7(b) = 0.17 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \ \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \ \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \ \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \ \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \qquad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.66 \end{cases} \qquad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \; \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \; \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) \; \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) \; \gamma_3(b) = 0.17 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \ \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \ \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \ \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \ \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) \ \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) \ \gamma_3(b) = 0.17 \end{cases} \quad \mathbf{X}^8 : (9H, 1T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \; \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \; \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) \; \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) \; \gamma_3(b) = 0.17 \end{cases} \quad \mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8) \; \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8) \; \gamma_8(b) = 0.08 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1)\ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1)\ \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6)\ \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6)\ \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2)\ \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2)\ \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7)\ \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7)\ \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3)\ \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3)\ \gamma_3(b) = 0.17 \end{cases} \quad \mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8)\ \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8)\ \gamma_8(b) = 0.08 \end{cases}$$

$$\mathbf{X}^4 : (8H, 2T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \; \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \; \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) \; \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) \; \gamma_3(b) = 0.17 \end{cases} \quad \mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8) \; \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8) \; \gamma_8(b) = 0.08 \end{cases}$$

$$\mathbf{X}^4 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^4) \; \gamma_4(a) = 0.88 \\ (z = b, \mathbf{X}^4) \; \gamma_4(b) = 0.12 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \; \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \; \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \; \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \; \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) \; \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) \; \gamma_3(b) = 0.17 \end{cases} \quad \mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8) \; \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8) \; \gamma_8(b) = 0.08 \end{cases}$$

$$\mathbf{X}^4 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^4) \; \gamma_4(a) = 0.88 \\ (z = b, \mathbf{X}^4) \; \gamma_4(b) = 0.12 \end{cases} \quad \mathbf{X}^9 : (3H, 7T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) & \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) & \gamma_1(b) = 0.12 \end{cases} \quad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) & \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) & \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) & \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) & \gamma_2(b) = 0.66 \end{cases} \quad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) & \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) & \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) & \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) & \gamma_3(b) = 0.17 \end{cases} \quad \mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8) & \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8) & \gamma_8(b) = 0.08 \end{cases}$$

$$\mathbf{X}^4 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^4) & \gamma_4(a) = 0.88 \\ (z = b, \mathbf{X}^4) & \gamma_4(b) = 0.12 \end{cases} \quad \mathbf{X}^9 : (3H, 7T) \begin{cases} (z = a, \mathbf{X}^9) & \gamma_9(a) = 0.22 \\ (z = b, \mathbf{X}^9) & \gamma_9(b) = 0.78 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) & \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) & \gamma_1(b) = 0.12 \end{cases} \qquad \mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) & \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) & \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) & \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) & \gamma_2(b) = 0.66 \end{cases} \qquad \mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) & \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) & \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) & \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) & \gamma_3(b) = 0.17 \end{cases} \qquad \mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8) & \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8) & \gamma_8(b) = 0.08 \end{cases}$$

$$\mathbf{X}^4 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^4) & \gamma_4(a) = 0.88 \\ (z = b, \mathbf{X}^4) & \gamma_4(b) = 0.12 \end{cases} \qquad \mathbf{X}^9 : (3H, 7T) \begin{cases} (z = a, \mathbf{X}^9) & \gamma_9(a) = 0.22 \\ (z = b, \mathbf{X}^9) & \gamma_9(b) = 0.78 \end{cases}$$

$$\mathbf{X}^5 : (1H, 9T)$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

E step for coin example

In the **E**-step you should picture each data point $\mathbf{X}^d$ as split into virtual population of $Z = a$ and $Z = b$ versions, with $\gamma_d(Z)$ as the virtual counts [1] :

$$\mathbf{X}^1 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.88 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.12 \end{cases}$$

$$\mathbf{X}^2 : (2H, 8T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.34 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.66 \end{cases}$$

$$\mathbf{X}^3 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^3) \ \gamma_3(a) = 0.83 \\ (z = b, \mathbf{X}^3) \ \gamma_3(b) = 0.17 \end{cases}$$

$$\mathbf{X}^4 : (8H, 2T) \begin{cases} (z = a, \mathbf{X}^4) \ \gamma_4(a) = 0.88 \\ (z = b, \mathbf{X}^4) \ \gamma_4(b) = 0.12 \end{cases}$$

$$\mathbf{X}^5 : (1H, 9T) \begin{cases} (z = a, \mathbf{X}^5) \ \gamma_5(a) = 0.45 \\ (z = b, \mathbf{X}^5) \ \gamma_5(b) = 0.55 \end{cases}$$

$$\mathbf{X}^6 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^6) \ \gamma_6(a) = 0.92 \\ (z = b, \mathbf{X}^6) \ \gamma_6(b) = 0.08 \end{cases}$$

$$\mathbf{X}^7 : (7H, 3T) \begin{cases} (z = a, \mathbf{X}^7) \ \gamma_7(a) = 0.83 \\ (z = b, \mathbf{X}^7) \ \gamma_7(b) = 0.17 \end{cases}$$

$$\mathbf{X}^8 : (9H, 1T) \begin{cases} (z = a, \mathbf{X}^8) \ \gamma_8(a) = 0.92 \\ (z = b, \mathbf{X}^8) \ \gamma_8(b) = 0.08 \end{cases}$$

$$\mathbf{X}^9 : (3H, 7T) \begin{cases} (z = a, \mathbf{X}^9) \ \gamma_9(a) = 0.22 \\ (z = b, \mathbf{X}^9) \ \gamma_9(b) = 0.78 \end{cases}$$

[1] the $\gamma_d(Z)$ numbers above assume $\theta_a = 0.5, \theta_{h|a} = 0.4, \theta_{h|b} = 0.3$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2$$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2 = 1.17965 \times 10^{-4}$$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2 = 1.17965 \times 10^{-4}$$

$$d = 1 : p(Z = b, HHHHHHHHTT) = 0.5 \times (0.3)^8 \times (0.7)^2$$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2 = 1.17965 \times 10^{-4}$$

$$d = 1 : p(Z = b, HHHHHHHHTT) = 0.5 \times (0.3)^8 \times (0.7)^2 = 1.60744 \times 10^{-5}$$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2 = 1.17965 \times 10^{-4}$$

$$d = 1 : p(Z = b, HHHHHHHHTT) = 0.5 \times (0.3)^8 \times (0.7)^2 = 1.60744 \times 10^{-5}$$

$$d = 1 : sum = 0.000134039$$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2 = 1.17965 \times 10^{-4}$$

$$d = 1 : p(Z = b, HHHHHHHHTT) = 0.5 \times (0.3)^8 \times (0.7)^2 = 1.60744 \times 10^{-5}$$

$$d = 1 : sum = 0.000134039$$

$$\gamma_1(a) = \frac{p(Z = a, HHHHHHHHTT)}{\sum_z P(Z = z, HHHHHHHHTT)} = \frac{1.17965 \times 10^{-4}}{sum} = 0.880077$$

Example calc of $\gamma_1(Z)$

$$d = 1 : p(Z = a, HHHHHHHHTT) = 0.5 \times (0.4)^8 \times (0.6)^2 = 1.17965 \times 10^{-4}$$

$$d = 1 : p(Z = b, HHHHHHHHTT) = 0.5 \times (0.3)^8 \times (0.7)^2 = 1.60744 \times 10^{-5}$$

$$d = 1 : sum = 0.000134039$$

$$\gamma_1(a) = \frac{p(Z = a, HHHHHHHHTT)}{\sum_z P(Z = z, HHHHHHHHTT)} = \frac{1.17965 \times 10^{-4}}{sum} = 0.880077$$

$$\gamma_1(b) = \frac{p(Z = b, HHHHHHHHTT)}{\sum_z P(Z = z, HHHHHHHHTT)} = \frac{1.60744 \times 10^{-5}}{sum} = 0.119923$$

M step for coin example

M step for coin example

In the **M** step you treat the $\gamma_d(Z)$ values as if they were genuine counts and re-estimate parameters in the usual common-sense fashion based on relative frequencies.

M step for coin example

In the **M** step you treat the $\gamma_d(Z)$ values as if they were genuine counts and re-estimate parameters in the usual common-sense fashion based on relative frequencies.

As a mental trick to help visualize you might consider all the preceding $\gamma_d(Z)$ as multiplied by 100 – effectively each single d is being treated as split out into 100 virtual versions, with $\gamma_d(Z) \times 100$ for each Z alternative

M step for coin example

In the **M** step you treat the $\gamma_d(Z)$ values as if they were genuine counts and re-estimate parameters in the usual common-sense fashion based on relative frequencies.

As a mental trick to help visualize you might consider all the preceding $\gamma_d(Z)$ as multiplied by 100 – effectively each single d is being treated as split out into 100 virtual versions, with $\gamma_d(Z) \times 100$ for each Z alternative

the 'common-sense' re-estimation of the parameters obtained this way represent a maximum likelihood estimate for any complete corpus that exhibits the same *ratios* as the obtained virtual corpus.

M step for coin example

For the coin scenario we can write down formulae for what the new round of estimates will be

M step for coin example

For the coin scenario we can write down formulae for what the new round of estimates will be
In (**??**), (**??**), (**??**) we had the estimation formulae for the fully observed case, making use of an indicator functions $\delta(d,.)$ – which for any given d are 1 for just one value of Z. The re-estimation formula for an M step are just these with the indicator function $\delta(d,.)$ replaced throughout by $\gamma_d(.)$

M step for coin example

For the coin scenario we can write down formulae for what the new round of estimates will be
In (**??**), (**??**), (**??**) we had the estimation formulae for the fully observed case, making use of an indicator functions $\delta(d,.)$ – which for any given d are 1 for just one value of Z. The re-estimation formula for an M step are just these with the indicator function $\delta(d,.)$ replaced throughout by $\gamma_d(.)$

$$est(\theta_a) = \frac{\sum_d \gamma_d(A)}{D} \tag{13}$$

M step for coin example

For the coin scenario we can write down formulae for what the new round of estimates will be
In (**??**), (**??**), (**??**) we had the estimation formulae for the fully observed case, making use of an indicator functions $\delta(d,.)$ – which for any given d are 1 for just one value of Z. The re-estimation formula for an M step are just these with the indicator function $\delta(d,.)$ replaced throughout by $\gamma_d(.)$

$$est(\theta_a) = \frac{\sum_d \gamma_d(A)}{D} \tag{13}$$

$$est(\theta_{h|a}) = \frac{\sum_d \gamma_d(A)\#(d,h)}{\sum_d \gamma_d(A)10} \tag{14}$$

M step for coin example

For the coin scenario we can write down formulae for what the new round of estimates will be

In (**??**), (**??**), (**??**) we had the estimation formulae for the fully observed case, making use of an indicator functions $\delta(d,.)$ – which for any given d are 1 for just one value of Z. The re-estimation formula for an M step are just these with the indicator function $\delta(d,.)$ replaced throughout by $\gamma_d(.)$

$$est(\theta_a) = \frac{\sum_d \gamma_d(A)}{D} \tag{13}$$

$$est(\theta_{h|a}) = \frac{\sum_d \gamma_d(A)\#(d,h)}{\sum_d \gamma_d(A)10} \tag{14}$$

$$est(\theta_{h|b}) = \frac{\sum_d \gamma_d(B)\#(d,h)}{\sum_d \gamma_d(B)10} \tag{15}$$

'common sense' M-step for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

in case that did not persuade, here's how to get to these re-estimation formulae by 'common sense' based on the virtual corpus

'common sense' M-step for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

in case that did not persuade, here's how to get to these re-estimation formulae by 'common sense' based on the virtual corpus

for θ_a, need $(cnt\ of\ virtual\ Z = A\ cases)/(cnt\ of\ all\ virtual\ Z\ cases)$, ie.

$$est(\theta_a) = \frac{\sum_d \gamma_d(a)}{\sum_d \gamma_d(a) + \sum_d \gamma_d(b)} = \frac{\sum_d \gamma_d(a)}{\sum_d(\gamma_d(a) + \gamma_d(b))} = \frac{\sum_d \gamma_d(A)}{D} \tag{16}$$

'common sense' M-step for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

in case that did not persuade, here's how to get to these re-estimation formulae by 'common sense' based on the virtual corpus

for θ_a, need *(cnt of virtual Z = A cases)/(cnt of all virtual Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_d \gamma_d(a)}{\sum_d \gamma_d(a) + \sum_d \gamma_d(b)} = \frac{\sum_d \gamma_d(a)}{\sum_d(\gamma_d(a) + \gamma_d(b))} = \frac{\sum_d \gamma_d(A)}{D} \quad (16)$$

for $\theta_{h|a}$, need
(cnt of H in virtual Z = a cases)/(cnt of all tosses in virtual Z = a cases), ie.

$$est(\theta_{h|a}) = \frac{\sum_d \gamma_d(a)\#(d,h)}{\sum_d \gamma_d(a)(\#(d,h) + \#(d,t))} = \frac{\sum_d \gamma_d(a)\#(d,h)}{\sum_d \gamma_d(a)10} \quad (17)$$

'common sense' M-step for θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

in case that did not persuade, here's how to get to these re-estimation formulae by 'common sense' based on the virtual corpus

for θ_a, need *(cnt of virtual Z = A cases)/(cnt of all virtual Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_d \gamma_d(a)}{\sum_d \gamma_d(a) + \sum_d \gamma_d(b)} = \frac{\sum_d \gamma_d(a)}{\sum_d(\gamma_d(a) + \gamma_d(b))} = \frac{\sum_d \gamma_d(A)}{D} \qquad (16)$$

for $\theta_{h|a}$, need
(cnt of H in virtual Z = a cases)/(cnt of all tosses in virtual Z = a cases), ie.

$$est(\theta_{h|a}) = \frac{\sum_d \gamma_d(a)\#(d,h)}{\sum_d \gamma_d(a)(\#(d,h) + \#(d,t))} = \frac{\sum_d \gamma_d(a)\#(d,h)}{\sum_d \gamma_d(a)10} \qquad (17)$$

for $\theta_{h|b}$, need
(cnt of H in virtual Z = b cases)/(cnt of all tosses in virtual Z = b cases), ie.

$$est(\theta_{h|b}) = \frac{\sum_d \gamma_d(b)\#(d,h)}{\sum_d \gamma_d(b)(\#(d,h) + \#(d,t))} = \frac{\sum_d \gamma_d(b)\#(d,h)}{\sum_d \gamma_d(b)10} \qquad (18)$$

Properties of EM re-estimation

EM starts with some setting $\boldsymbol{\theta}^0$ of the parameters and one E-M cycle takes one setting $\boldsymbol{\theta}^n$ into another $\boldsymbol{\theta}^{n+1}$.

Properties of EM re-estimation

EM starts with some setting $\boldsymbol{\theta}^0$ of the parameters and one E-M cycle takes one setting $\boldsymbol{\theta}^n$ into another $\boldsymbol{\theta}^{n+1}$.

the data gets likelier over the iterations

$$P(\mathbf{d}; \boldsymbol{\theta}^n) \leq P(\mathbf{d}; \boldsymbol{\theta}^{n+1})$$

Properties of EM re-estimation

EM starts with some setting $\boldsymbol{\theta}^0$ of the parameters and one E-M cycle takes one setting $\boldsymbol{\theta}^n$ into another $\boldsymbol{\theta}^{n+1}$.

the data gets likelier over the iterations

$$P(\mathbf{d}; \boldsymbol{\theta}^n) \leq P(\mathbf{d}; \boldsymbol{\theta}^{n+1})$$

because the data cannot just get likelier and likelier, the procedure converges to a final setting $\boldsymbol{\theta}^{final}$

Properties of EM re-estimation

EM starts with some setting $\boldsymbol{\theta}^0$ of the parameters and one E-M cycle takes one setting $\boldsymbol{\theta}^n$ into another $\boldsymbol{\theta}^{n+1}$.

the data gets likelier over the iterations

$$P(\mathbf{d}; \boldsymbol{\theta}^n) \leq P(\mathbf{d}; \boldsymbol{\theta}^{n+1})$$

because the data cannot just get likelier and likelier, the procedure converges to a final setting $\boldsymbol{\theta}^{final}$

so whatever values $\boldsymbol{\theta}^0$ you start with, running the algorithm will give you better values $\boldsymbol{\theta}^{final}$

some provisos though ...

- Caveat One: there may be many local maxima, so there is no guarantee that the re-estimation will converge to *the* best values

some provisos though ...

- Caveat One: there may be many local maxima, so there is no guarantee that the re-estimation will converge to *the* best values
- Caveat Two: if the data set **d** is rather small the derived parameters may fit fresh data only poorly – this the classic over-fitting problem.

some provisos though ...

- Caveat One: there may be many local maxima, so there is no guarantee that the re-estimation will converge to *the* best values
- Caveat Two: if the data set $\mathbf{d}$ is rather small the derived parameters may fit fresh data only poorly – this the classic over-fitting problem.
- Caveat Three: it will be prohibitively expensive to calculate all $\gamma_d(\mathbf{k})$ *if* the set $\mathcal{A}(\mathbf{z})$ of the possible values of $\mathbf{z}$ is exponentially big. This does not apply to our hidden coin choice scenario – size of $\mathcal{A}(\mathbf{z})$ is 2 – but definitely applies to applications we are going to look at (eg. in Machine Translation and Speech Recognition) and requires algorithic ingenuity to make it still work.

Outline

A numerically worked example

To keep things manageable on slides lets suppose a minute data set

d	Z	**X**: tosses of chosen coin	
1	?	H	H
2	?	T	T

A numerically worked example

To keep things manageable on slides lets suppose a minute data set

d	Z	**X**: tosses of chosen coin	
1	?	H	H
2	?	T	T

looks like having A be entirely biased one way, and B entirely the other will give maximum prob to this.

A numerically worked example

To keep things manageable on slides lets suppose a minute data set

d	Z	**X**: tosses of chosen coin	
1	?	H	H
2	?	T	T

looks like having A be entirely biased one way, and B entirely the other will give maximum prob to this. The outcomes when EM is run from start $\theta_a = 0.5$, $\theta_{h|a} = 0.75$ and $\theta_{h|b} = 0.5$ is:

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.75	0.5	-3.97763	0.0634766
0.446154	0.775862	0.277778	-3.36722	0.0969094
0.467361	0.922972	0.128866	-2.59395	0.165632
0.49254	0.993083	0.0214144	-2.08205	0.236179
:	:	:	:	:
0.5	1	0	-2	0.25

A numerically worked example

To keep things manageable on slides lets suppose a minute data set

d	Z	**X**: tosses of chosen coin	
1	?	H	H
2	?	T	T

looks like having A be entirely biased one way, and B entirely the other will give maximum prob to this. The outcomes when EM is run from start $\theta_a = 0.5$, $\theta_{h|a} = 0.75$ and $\theta_{h|b} = 0.5$ is:

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.75	0.5	-3.97763	0.0634766
0.446154	0.775862	0.277778	-3.36722	0.0969094
0.467361	0.922972	0.128866	-2.59395	0.165632
0.49254	0.993083	0.0214144	-2.08205	0.236179
:	:	:	:	:
0.5	1	0	-2	0.25

so EM finds the intuitive solution. Next few slides trace the first iteration of the algorithm

Let $\mathbf{X}^d$ be the coin toss outcomes for a particular trial. The probability of the version where the chosen coin was A is

$$\begin{aligned} P(Z = a, \mathbf{X}^d) &= P(Z = a) \times P(h|a)^{\#(d,h)} \times P(t|a)^{\#(d,t)} \\ &= \theta_a \times \theta_{h|a}^{\#(d,h)} \times \theta_{t|a}^{\#(d,t)} \end{aligned}$$

Let $\mathbf{X}^d$ be the coin toss outcomes for a particular trial. The probability of the version where the chosen coin was A is

$$\begin{aligned} P(Z=a,\mathbf{X}^d) &= P(Z=a) \times P(h|a)^{\#(d,h)} \times P(t|a)^{\#(d,t)} \\ &= \theta_a \times \theta_{h|a}^{\#(d,h)} \times \theta_{t|a}^{\#(d,t)} \end{aligned}$$

and likewise the probability of the version where the chosen coin was B is given by

$$\begin{aligned} P(Z=b,\mathbf{X}^d) &= P(Z=b) \times P(h|b)^{\#(d,h)} \times P(t|b)^{\#(d,t)} \\ &= \theta_b \times \theta_{h|b}^{\#(d,h)} \times \theta_{t|b}^{\#(d,t)} \end{aligned}$$

Let $\mathbf{X}^d$ be the coin toss outcomes for a particular trial. The probability of the version where the chosen coin was A is

$$\begin{aligned} P(Z=a,\mathbf{X}^d) &= P(Z=a) \times P(h|a)^{\#(d,h)} \times P(t|a)^{\#(d,t)} \\ &= \theta_a \times \theta_{h|a}^{\#(d,h)} \times \theta_{t|a}^{\#(d,t)} \end{aligned}$$

and likewise the probability of the version where the chosen coin was B is given by

$$\begin{aligned} P(Z=b,\mathbf{X}^d) &= P(Z=b) \times P(h|b)^{\#(d,h)} \times P(t|b)^{\#(d,t)} \\ &= \theta_b \times \theta_{h|b}^{\#(d,h)} \times \theta_{t|b}^{\#(d,t)} \end{aligned}$$

and from these joint probability formula the *conditional probabilities* for the hidden variable will be:

$$\begin{aligned} P(Z=a|\mathbf{X}^d) &= \frac{P(Z=a,\mathbf{X}^d)}{\sum_k P(Z=k,\mathbf{X}^d)} \\ P(Z=b|\mathbf{X}^d) &= \frac{P(Z=b,\mathbf{X}^d)}{\sum_k P(Z=k,\mathbf{X}^d)} \end{aligned}$$

On the particular data set at hand the joint probability formulae are particularly simple

On the particular data set at hand the joint probability formulae are particularly simple

$$
\begin{aligned}
P(Z=a,\mathbf{X}^1) &= \theta_a \times (\theta_{h|a})^2 \\
P(Z=b,\mathbf{X}^1) &= \theta_b \times (\theta_{h|b})^2 \\
P(Z=a,\mathbf{X}^2) &= \theta_a \times (\theta_{t|a})^2 \\
P(Z=b,\mathbf{X}^2) &= \theta_b \times (\theta_{t|b})^2
\end{aligned}
$$

On the particular data set at hand the joint probability formulae are particularly simple

$$
\begin{aligned}
P(Z = a, \mathbf{X}^1) &= \theta_a \times (\theta_{h|a})^2 \\
P(Z = b, \mathbf{X}^1) &= \theta_b \times (\theta_{h|b})^2 \\
P(Z = a, \mathbf{X}^2) &= \theta_a \times (\theta_{t|a})^2 \\
P(Z = b, \mathbf{X}^2) &= \theta_b \times (\theta_{t|b})^2
\end{aligned}
$$

and thus the formulae for $\gamma_d(Z)$ are:

On the particular data set at hand the joint probability formulae are particularly simple

$$
\begin{aligned}
P(Z = a, \mathbf{X}^1) &= \theta_a \times (\theta_{h|a})^2 \\
P(Z = b, \mathbf{X}^1) &= \theta_b \times (\theta_{h|b})^2 \\
P(Z = a, \mathbf{X}^2) &= \theta_a \times (\theta_{t|a})^2 \\
P(Z = b, \mathbf{X}^2) &= \theta_b \times (\theta_{t|b})^2
\end{aligned}
$$

and thus the formulae for $\gamma_d(Z)$ are:

$$
\begin{aligned}
\gamma_1(a) &= \frac{\theta_a \times (\theta_{h|a})^2}{\theta_a \times (\theta_{h|a})^2 + \theta_b \times (\theta_{h|b})^2} \\
\gamma_1(b) &= \frac{\theta_b \times (\theta_{h|b})^2}{\theta_a \times (\theta_{h|a})^2 + \theta_b \times (\theta_{h|b})^2} \\
\gamma_2(a) &= \frac{\theta_a \times (\theta_{t|a})^2}{\theta_a \times (\theta_{t|a})^2 + \theta_b \times (\theta_{t|b})^2} \\
\gamma_2(b) &= \frac{\theta_b \times (\theta_{t|b})^2}{\theta_a \times (\theta_{t|a})^2 + \theta_b \times (\theta_{t|b})^2}
\end{aligned}
$$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$
$d = 1 :\rightarrow \gamma_1(A) = 0.692308$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$
$d = 1 :\rightarrow \gamma_1(A) = 0.692308$
$d = 1 :\rightarrow \gamma_1(B) = 0.307692$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$
$d = 1 :\rightarrow \gamma_1(A) = 0.692308$
$d = 1 :\rightarrow \gamma_1(B) = 0.307692$
$d = 2 : p(Z = A, TT) = 0.5 \times 0.25 \times 0.25 = 0.03125$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$
$d = 1 :\rightarrow \gamma_1(A) = 0.692308$
$d = 1 :\rightarrow \gamma_1(B) = 0.307692$
$d = 2 : p(Z = A, TT) = 0.5 \times 0.25 \times 0.25 = 0.03125$
$d = 2 : p(Z = B, TT) = 0.5 \times 0.5 \times 0.5 = 0.125$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$$
$$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$$
$$d = 1 :\rightarrow sum = 0.40625$$
$$d = 1 :\rightarrow \gamma_1(A) = 0.692308$$
$$d = 1 :\rightarrow \gamma_1(B) = 0.307692$$
$$d = 2 : p(Z = A, TT) = 0.5 \times 0.25 \times 0.25 = 0.03125$$
$$d = 2 : p(Z = B, TT) = 0.5 \times 0.5 \times 0.5 = 0.125$$
$$d = 2 :\rightarrow sum = 0.15625$$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$
$d = 1 :\rightarrow \gamma_1(A) = 0.692308$
$d = 1 :\rightarrow \gamma_1(B) = 0.307692$
$d = 2 : p(Z = A, TT) = 0.5 \times 0.25 \times 0.25 = 0.03125$
$d = 2 : p(Z = B, TT) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 2 :\rightarrow sum = 0.15625$
$d = 2 :\rightarrow \gamma_2(A) = 0.2$

To carry out an EM estimation of the parameters given the data we need some initial setting of the parameters. We will suppose this is:

$$\theta_a = \frac{1}{2}, \theta_b = \frac{1}{2}, \theta_{h|a} = \frac{3}{4}, \theta_{t|a} = \frac{1}{4}, \theta_{h|b} = \frac{1}{2}, \theta_{t|b} = \frac{1}{2}$$

ITERATION 1

For each piece of data have to first compute the conditional probabilities of the hidden variable given the data:

$d = 1 : p(Z = A, HH) = 0.5 \times 0.75 \times 0.75 = 0.28125$
$d = 1 : p(Z = B, HH) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 1 :\rightarrow sum = 0.40625$
$d = 1 :\rightarrow \gamma_1(A) = 0.692308$
$d = 1 :\rightarrow \gamma_1(B) = 0.307692$
$d = 2 : p(Z = A, TT) = 0.5 \times 0.25 \times 0.25 = 0.03125$
$d = 2 : p(Z = B, TT) = 0.5 \times 0.5 \times 0.5 = 0.125$
$d = 2 :\rightarrow sum = 0.15625$
$d = 2 :\rightarrow \gamma_2(A) = 0.2$
$d = 2 :\rightarrow \gamma_2(B) = 0.8$

Armed with these γ values we now treat each data item $\mathbf{X}^d$ as if it splits into two versions, one filling out Z as a, and with 'count' $\gamma_d(a)$, and one filling out Z as b, and with 'count' $\gamma_d(b)$.

Armed with these γ values we now treat each data item $\mathbf{X}^d$ as if it splits into two versions, one filling out Z as a, and with 'count' $\gamma_d(a)$, and one filling out Z as b, and with 'count' $\gamma_d(b)$.

$\mathbf{X}^1 : (2H, 0T)$

Armed with these γ values we now treat each data item $\mathbf{X}^d$ as if it splits into two versions, one filling out Z as a, and with 'count' $\gamma_d(a)$, and one filling out Z as b, and with 'count' $\gamma_d(b)$.

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$

Armed with these γ values we now treat each data item $\mathbf{X}^d$ as if it splits into two versions, one filling out Z as a, and with 'count' $\gamma_d(a)$, and one filling out Z as b, and with 'count' $\gamma_d(b)$.

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$

$$\mathbf{X}^2 : (0H, 2T)$$

Armed with these γ values we now treat each data item $\mathbf{X}^d$ as if it splits into two versions, one filling out Z as a, and with 'count' $\gamma_d(a)$, and one filling out Z as b, and with 'count' $\gamma_d(b)$.

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$

$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.8 \end{cases}$$

$$\mathbf{X^1} : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) & \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) & \gamma_1(b) = 0.307692 \end{cases}$$

$$\mathbf{X^2} : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) & \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) & \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event.

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a)$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b)$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) & \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) & \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) & \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) & \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a) \#(d, h)$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \; \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \; \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \; \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \; \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1)\ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1)\ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2)\ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2)\ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$
$$E(A, T) = \sum_d \gamma_d(a)\#(d, t)$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$
$$E(A, T) = \sum_d \gamma_d(a)\#(d, t) = (0.692308 \times 0 + 0.2 \times 2) = 0.4$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$
$$E(A, T) = \sum_d \gamma_d(a)\#(d, t) = (0.692308 \times 0 + 0.2 \times 2) = 0.4$$
$$E(B, H) = \sum_d \gamma_d(b)\#(d, h)$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1)\ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1)\ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2)\ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2)\ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$
$$E(A, T) = \sum_d \gamma_d(a)\#(d, t) = (0.692308 \times 0 + 0.2 \times 2) = 0.4$$
$$E(B, H) = \sum_d \gamma_d(b)\#(d, h) = (0.307692 \times 2 + 0.8 \times 0) = 0.615385$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$
$$E(A, T) = \sum_d \gamma_d(a)\#(d, t) = (0.692308 \times 0 + 0.2 \times 2) = 0.4$$
$$E(B, H) = \sum_d \gamma_d(b)\#(d, h) = (0.307692 \times 2 + 0.8 \times 0) = 0.615385$$
$$E(B, T) = \sum_d \gamma_d(b)\#(d, t)$$

$$\mathbf{X}^1 : (2H, 0T) \begin{cases} (z = a, \mathbf{X}^1) \ \gamma_1(a) = 0.692308 \\ (z = b, \mathbf{X}^1) \ \gamma_1(b) = 0.307692 \end{cases}$$
$$\mathbf{X}^2 : (0H, 2T) \begin{cases} (z = a, \mathbf{X}^2) \ \gamma_2(a) = 0.2 \\ (z = b, \mathbf{X}^2) \ \gamma_2(b) = 0.8 \end{cases}$$

We then go through this virtual corpus accumulating counts of certain kinds of event. For events of hidden variable being $Z = a$ and $Z = b$ we get

$$E(A) = \gamma_1(a) + \gamma_2(a) = 0.692308 + 0.2 = 0.892308$$
$$E(B) = \gamma_1(b) + \gamma_2(b) = 0.307692 + 0.8 = 1.10769$$

Then we need to go through the $Z = a$ cases and count types of coin toss, and likewise for $Z = b$ cases

$$E(A, H) = \sum_d \gamma_d(a)\#(d, h) = (0.692308 \times 2 + 0.2 \times 0) = 1.38462$$
$$E(A, T) = \sum_d \gamma_d(a)\#(d, t) = (0.692308 \times 0 + 0.2 \times 2) = 0.4$$
$$E(B, H) = \sum_d \gamma_d(b)\#(d, h) = (0.307692 \times 2 + 0.8 \times 0) = 0.615385$$
$$E(B, T) = \sum_d \gamma_d(b)\#(d, t) = (0.307692 \times 0 + 0.8 \times 2) = 1.6$$

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

$$est(\theta_a) = E(A)/2 = 0.892308/2 = 0.446154$$

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

$$\begin{aligned} est(\theta_a) &= E(A)/2 = 0.892308/2 = 0.446154 \\ est(\theta_b) &= E(B)/2 = 1.10769/2 = 0.553846 \end{aligned}$$

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

$$
\begin{aligned}
est(\theta_a) &= E(A)/2 = 0.892308/2 = 0.446154 \\
est(\theta_b) &= E(B)/2 = 1.10769/2 = 0.553846
\end{aligned}
$$

Note the denominator 2 in the re-estimation formula for θ_a.

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

$$
\begin{aligned}
est(\theta_a) &= E(A)/2 = 0.892308/2 = 0.446154 \\
est(\theta_b) &= E(B)/2 = 1.10769/2 = 0.553846
\end{aligned}
$$

Note the denominator 2 in the re-estimation formula for θ_a. We could have written the denominator as $E(A) + E(B)$, but this is

$$\sum_d \gamma_d(a) + \sum_d \gamma_d(b)$$

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

$$\begin{aligned} est(\theta_a) &= E(A)/2 = 0.892308/2 = 0.446154 \\ est(\theta_b) &= E(B)/2 = 1.10769/2 = 0.553846 \end{aligned}$$

Note the denominator 2 in the re-estimation formula for θ_a. We could have written the denominator as $E(A) + E(B)$, but this is

$$\sum_d \gamma_d(a) + \sum_d \gamma_d(b) = \sum_d [\gamma_d(a) + \gamma_d(b)]$$

re-estimating θ_a and θ_b

Then from these 'expected' counts we re-estimate parameters

$$est(\theta_a) = E(A)/2 = 0.892308/2 = 0.446154$$
$$est(\theta_b) = E(B)/2 = 1.10769/2 = 0.553846$$

Note the denominator 2 in the re-estimation formula for θ_a. We could have written the denominator as $E(A) + E(B)$, but this is

$$\sum_d \gamma_d(a) + \sum_d \gamma_d(b) = \sum_d [\gamma_d(a) + \gamma_d(b)] = \sum_d [1] = 2$$

re-estimating $\theta_{h|a}$

re-estimating $\theta_{h|a}$

$$
\begin{aligned}
est(\theta_{h|a}) = E(A,H)/\sum_X [E(A,X)] &= 1.38462/(1.38462+0.4) \\
&= 1.38462/1.78462 \\
&= 0.775862
\end{aligned}
$$

re-estimating $\theta_{h|a}$

$$
\begin{aligned}
est(\theta_{h|a}) = E(A,H)/\sum_X[E(A,X)] &= 1.38462/(1.38462+0.4) \\
&= 1.38462/1.78462 \\
&= 0.775862
\end{aligned}
$$

$$
\begin{aligned}
est(\theta_{t|a}) = E(A,T)/\sum_X[E(A,X)] &= 0.4/(1.38462+0.4) \\
&= 0.4/1.78462 \\
&= 0.224138
\end{aligned}
$$

re-estimating $\theta_{h|b}$

re-estimating $\theta_{h|b}$

$$
\begin{aligned}
est(\theta_{h|b}) = E(B,H)/\sum_X [E(B,X)] &= 0.615385/(0.615385 + 1.6) \\
&= 0.615385/2.21538 \\
&= 0.277778
\end{aligned}
$$

re-estimating $\theta_{h|b}$

$$
\begin{aligned}
est(\theta_{h|b}) = E(B,H)/\sum_X [E(B,X)] &= 0.615385/(0.615385 + 1.6) \\
&= 0.615385/2.21538 \\
&= 0.277778
\end{aligned}
$$

$$
\begin{aligned}
est(\theta_{t|b}) = E(B,T)/\sum_X [E(B,X)] &= 1.6/(0.615385 + 1.6) \\
&= 1.6/2.21538 \\
&= 0.722222
\end{aligned}
$$

the above traced through how the 2nd row of the table below comes from the first.

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.75	0.5	-3.97763	0.0634766
0.446154	0.775862	0.277778	-3.36722	0.0969094
0.467361	0.922972	0.128866	-2.59395	0.165632
0.49254	0.993083	0.0214144	-2.08205	0.236179
:	:	:	:	:
0.5	1	0	-2	0.25

the above traced through how the 2nd row of the table below comes from the first.

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.75	0.5	-3.97763	0.0634766
0.446154	0.775862	0.277778	-3.36722	0.0969094
0.467361	0.922972	0.128866	-2.59395	0.165632
0.49254	0.993083	0.0214144	-2.08205	0.236179
:	:	:	:	:
0.5	1	0	-2	0.25

In the end it converges to $\theta_a = 0.5$, $\theta_{h|a} = 1$, $\theta_{h|b} = 0$.

the above traced through how the 2nd row of the table below comes from the first.

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.75	0.5	-3.97763	0.0634766
0.446154	0.775862	0.277778	-3.36722	0.0969094
0.467361	0.922972	0.128866	-2.59395	0.165632
0.49254	0.993083	0.0214144	-2.08205	0.236179
:	:	:	:	:
0.5	1	0	-2	0.25

In the end it converges to $\theta_a = 0.5$, $\theta_{h|a} = 1$, $\theta_{h|b} = 0$.

also tracked in the table is the increasing prob of the data, and log-prob

Outline

More realistic run of EM

recall the data we had for our 2nd scenario, with the coin-choice observed:

d	Z	**X**: tosses of chosen coin										H counts
1	A	H	H	H	H	H	H	H	H	T	T	(8H)
2	B	T	T	H	T	T	T	H	T	T	T	(2H)
3	A	H	T	H	H	T	H	H	H	H	T	(7H)
4	A	H	T	H	H	H	T	H	H	H	H	(8H)
5	B	T	T	T	T	T	T	H	T	T	T	(1H)
6	A	H	H	T	H	H	H	H	H	H	H	(9H)
7	A	T	H	H	T	H	H	H	H	H	T	(7H)
8	A	H	H	H	H	H	H	T	H	H	H	(9H)
9	B	H	H	T	T	T	T	T	H	T	T	(3H)

recall supervised estimation gave: $\theta_a = 0.66, \theta_{h|a} = 0.8, \theta_{h|b} = 0.2$

More realistic run of EM continued

here's an outcome of running EM treating Z as hidden

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.4	0.3	-101.033	3.85587e-31
0.698501	0.70713	0.351806	-77.3507	5.18952e-24
0.666619	0.793432	0.213219	-73.2502	8.90206e-23
0.66705	0.799293	0.200725	-73.2201	9.08992e-23
0.667134	0.799354	0.200453	-73.2201	9.08999e-23

no further change

More realistic run of EM continued

here's an outcome of running EM treating Z as hidden

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.4	0.3	-101.033	3.85587e-31
0.698501	0.70713	0.351806	-77.3507	5.18952e-24
0.666619	0.793432	0.213219	-73.2502	8.90206e-23
0.66705	0.799293	0.200725	-73.2201	9.08992e-23
0.667134	0.799354	0.200453	-73.2201	9.08999e-23
		no further change		

these are very close to the numbers obtained when Z was not hidden.

More realistic run of EM continued

here's an outcome of running EM treating Z as hidden

θ_a	$\theta_{h\|a}$	$\theta_{h\|b}$	logprob	prob
0.5	0.4	0.3	-101.033	3.85587e-31
0.698501	0.70713	0.351806	-77.3507	5.18952e-24
0.666619	0.793432	0.213219	-73.2502	8.90206e-23
0.66705	0.799293	0.200725	-73.2201	9.08992e-23
0.667134	0.799354	0.200453	-73.2201	9.08999e-23

no further change

these are very close to the numbers obtained when Z was not hidden.

On this data set also the final outcome is not very dependent on the initial values