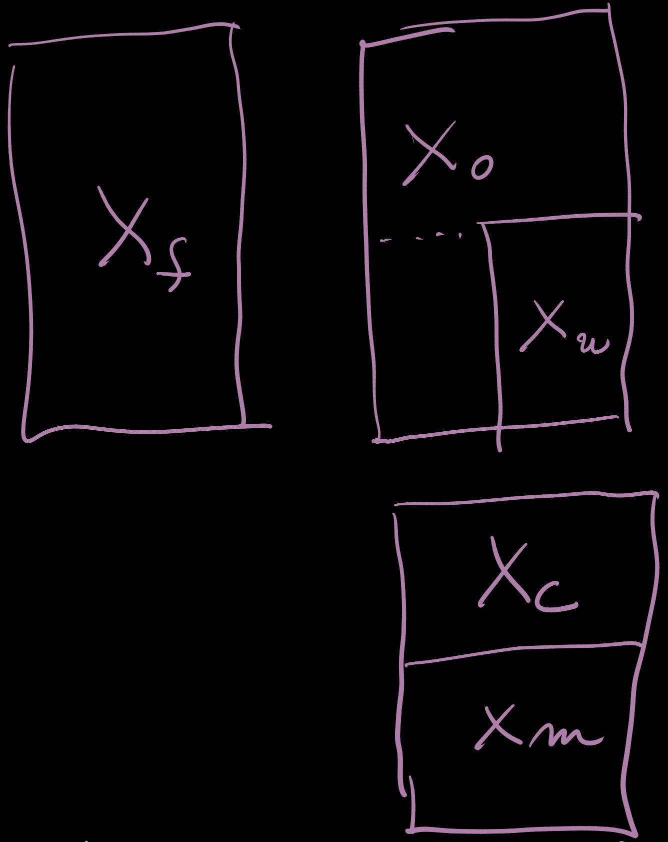
# Why the E-M algorithm "works":

Recall:

| | | Text |
|---|---|---|
| $X_f$ | full data | $Y$ |
| $X_o$ | observed | $X$ |
| $X_u$ | un-observed (missing data or latent parameters) | $Z$ |

An example




Likelihood: $f(X_o \mid \underset{\sim}{\theta})$

I'll forget to write this.

$$\underbrace{f(x_o \mid \theta)} = \int \underbrace{f(x_o, x_u \mid \theta)}\, dx_u$$

Often " is complicated but $f(x_f \mid \theta)$ is simple

Note also $f(x_o \mid \theta) = \int \dfrac{f(x_o, x_u \mid \theta)}{f(x_u \mid \theta)} f(x_u \mid \theta)\, dx_u$

$$= E_{x_u \mid \theta}\left( f(x_o \mid x_u, \theta) \right)$$

marginal density = mean conditional density

" mean = " " mean

" variance $\neq$ mean conditional variance

BUT = mean conditional variance

+ variance of conditional mean

# E-M algorithm

Remember: we really want: $\arg\max_{\theta} \ell(X_o|\theta)$

Use $\ell(X_f|\theta) = \boxed{\ell(X_o, X_u|\theta)}$

E step: Let $Q(\theta, \theta^t) = E(\boxed{\ell(X_o, X_u|\theta)} \,|\, X_o, \theta^t)$

$= E_{X_u|X_o,\theta^t}(\ell(X_o, X_u|\theta))$

M step: $\theta_{t+1} = \arg\max_{\theta} Q(\theta, \theta^t)$

Now we show how going from $\theta^t \to \theta^{t+1}$ cannot decrease $l(X_0 \mid \theta)$

---

$$l(x_0 \mid \theta) = \log f(x_0 \mid \theta)$$

$$= \log f(x_0, x_u \mid \theta) - \log f(x_u \mid x_0, \theta)$$

Take $E(\ldots \mid X_0, \theta^t)$

A $f(x_u \mid x_0, \theta)$

$$= \frac{\text{B } f(x_0, x_u \mid \theta)}{\text{C } f(x_0 \mid \theta)}$$

C = B / A

$\log$ C = $\log$ B $- \log$ A

$$E\left(\ell(x_o\mid\theta)\mid x_o,\theta^t\right)$$

$$= E\left(\log f(x_o,x_u\mid\theta)\mid x_o,\theta^t\right)$$

$$- \underbrace{E\left(\log f(x_u\mid x_o,\theta)\mid x_o,\theta^t\right)}_{\text{call this } H(\theta,\theta^t)}$$

why?

We get: $\ell(x_o\mid\theta)$

$$= Q(\theta,\theta^t)$$

$$- H(\theta,\theta^t)$$

(also $x_o$ but that stays the same)

Consider $H(\theta, \theta^t)$: (from text p. 103)

$H(\theta^t, \theta^t) - H(\theta, \theta^t)$

only difference

$$= E\left(\log f(x_u \mid x_o, \theta^t) \mid x_o, \theta^t\right)$$

$$- E\left(\log f(x_u \mid x_o, \theta) \mid x_o, \theta^t\right)$$

$$= E\left(-\log\left(\frac{f(x_u \mid x_o, \theta)}{f(x_u \mid x_o, \theta^t)}\right) \middle| x_o, \theta^t\right)$$

$$= \int -\log\left(\frac{f(x_n \mid x_0, \theta)}{f(x_n \mid x_0, \theta^t)}\right) f(x_n \mid x_0, \theta^t)\, dX_n$$

$$\geq -\log \int \frac{f(x_n \mid x_0, \theta)}{\cancel{f(x_n \mid x_0, \theta^t)}} \cancel{f(x_n \mid x_0, \theta^t)}\, dX_n$$

$$= -\log \underbrace{\int f(x_n \mid x_0, \theta)\, dX_n}_{=1} = 0$$

So $H(\theta^t, \theta^t) \geq H(\theta, \theta^t)$

**Recall:** $$\ell(x_0 \mid \theta) = Q(\theta, \theta^t) - H(\theta, \theta^t)$$

So: Since $Q(\theta^{t+1}, \theta^t) \geq Q(\theta^t, \theta^t)$ why?

and $H(\theta^*, \theta^*) \geq H(\theta, \theta^*)$

$$\ell(x_0 \mid \theta^{t+1}) = Q(\theta \;, \theta^{t+1}) - H(\theta^{t+1}, \theta^t)$$

$$\ell(x_0|\theta^{t+1}) - \ell(x_0|\theta^{t})$$

$$= \underbrace{Q(\theta^{t+1};\theta^{t}) - Q(\theta^{t},\theta^{t})}_{\geq 0 \text{ why?}}$$

$$\underbrace{-\Big(\underbrace{H(\theta^{t+1};\theta^{t}) - H(\theta^{t};\theta^{t})}_{\leq 0}\Big)}_{\geq 0}$$

$$\geq 0$$

$\boxed{\text{Q.E.D.}}$