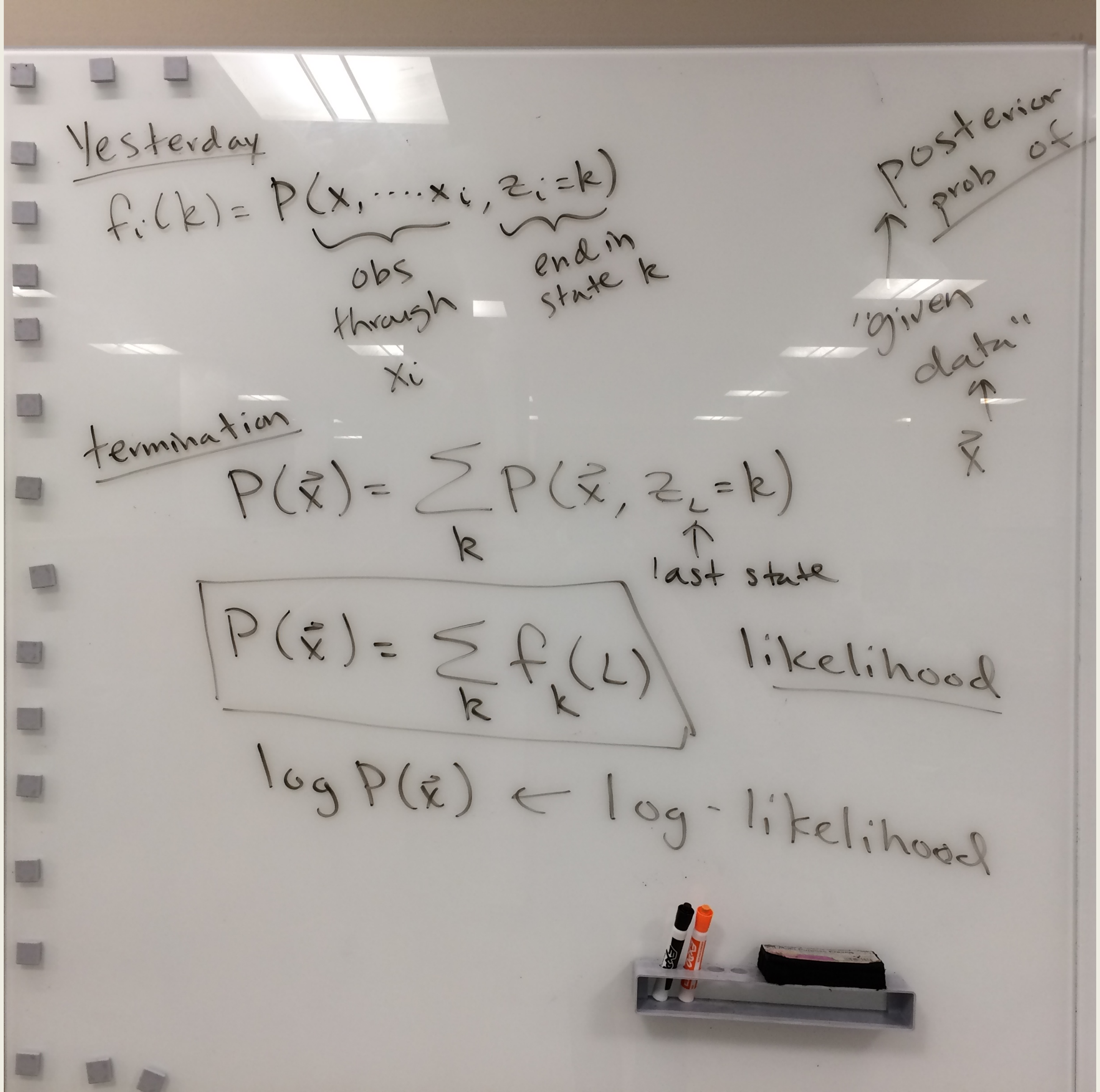
CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

In-lab notes: April 5

- Posterior decoding
- Posterior mean
- Working in log-space

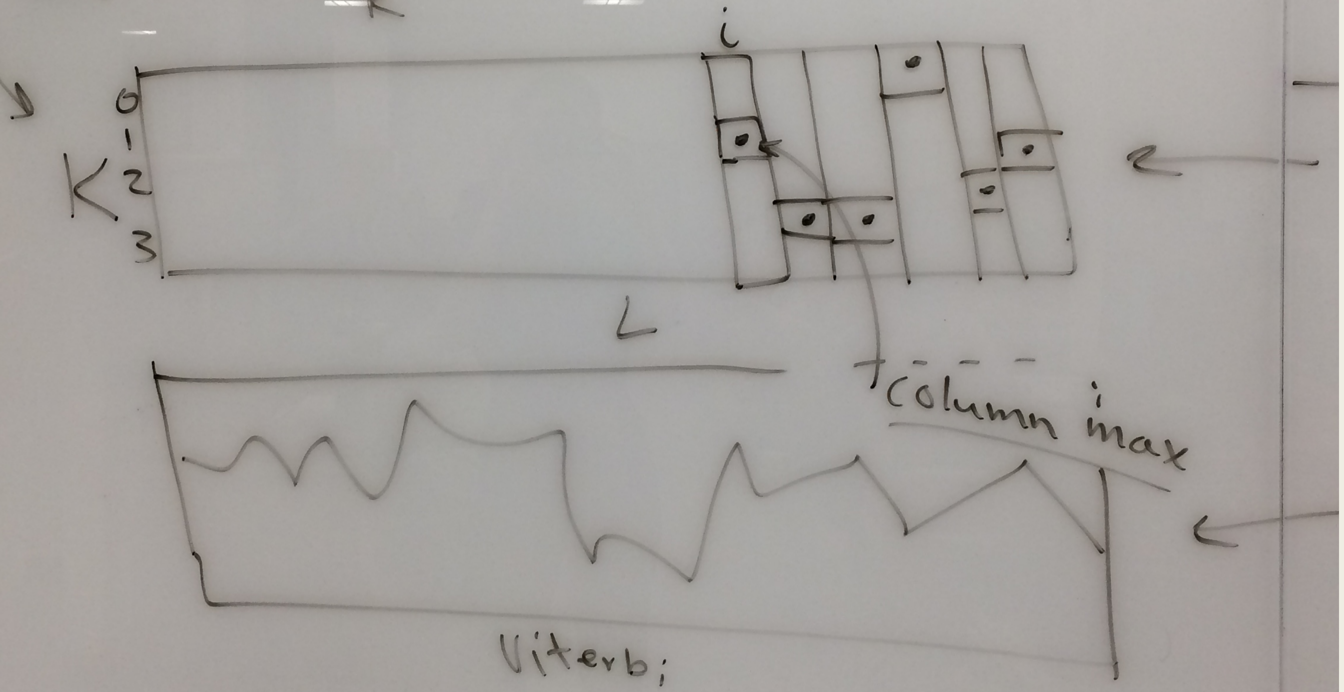
Lab A

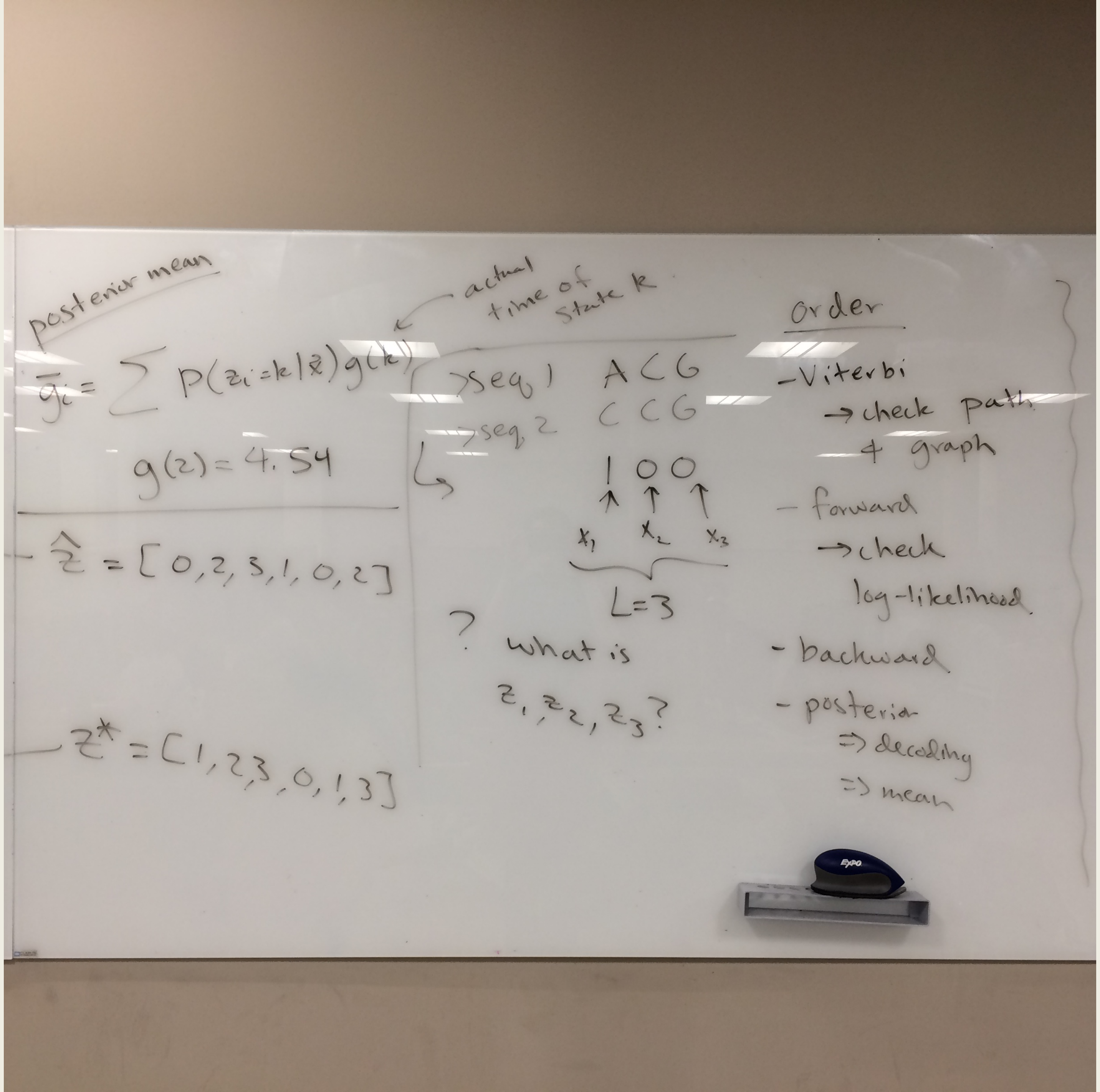

Yesterday
$f_i(k) = P(\underbrace{x_1, \dots x_i}_{\text{obs through } x_i}, \underbrace{z_i = k}_{\text{end in state } k})$
termination
$P(\vec{x}) = \sum_k P(\vec{x}, z_L = k)$
last state
$\boxed{P(\vec{x}) = \sum_k f_k(L)}$
likelihood
$\log P(\vec{x}) \leftarrow$ log-likelihood
posterior prob of
"given data"
$\vec{x}$

$z_i = k$

$$P(z_i = k \mid \vec{x}) = \frac{f_k(i)\, b_k(i)}{P(\vec{x})}$$

$$\hat{z}_i = \underset{k}{\operatorname{argmax}}\, P(z_i = k \mid \vec{x})$$



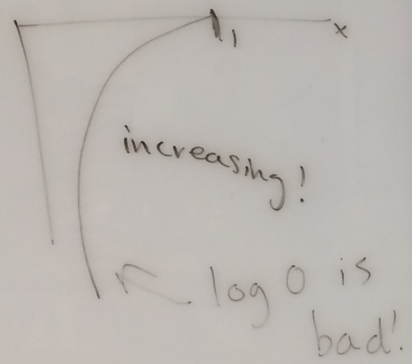

posterior mean
actual time of state k
$\bar{g}_i = \sum p(z_i = k | \vec{x}) g(k)$
$g(2) = 4.54$
$\hat{z} = [0, 2, 3, 1, 0, 2]$
$z^* = [1, 2, 3, 0, 1, 3]$
>seq 1 ACG
>seq 2 CCG
1 0 0
x_1 x_2 x_3
L=3
? what is
z_1, z_2, z_3?
Order
-Viterbi
→ check path
& graph
- forward
→ check
log-likelihood
- backward
- posterior
⇒ decoding
⇒ mean

log-space

$$\log\Big(V_k(i) = e_k(x_i) \cdot \overbrace{\max_l \{ V_l(i-1) \cdot a_{lk} \}}^{\text{all be zero!}} \Big)$$

$$\tilde{V}_k(i) = \log(e_k(x_i)) + \max_l \{ \tilde{V}_l(i-1) + \log(a_{lk}) \}$$

$$\log(p + \underbrace{q + r})$$



log

=

=

=

$$\log\left(f_k(i) = e_k(x_i) \sum_{l} f_l(i-1)\, a_{lk}\right)$$

$\tilde{p} = \log p$

$\tilde{q} = \log q$

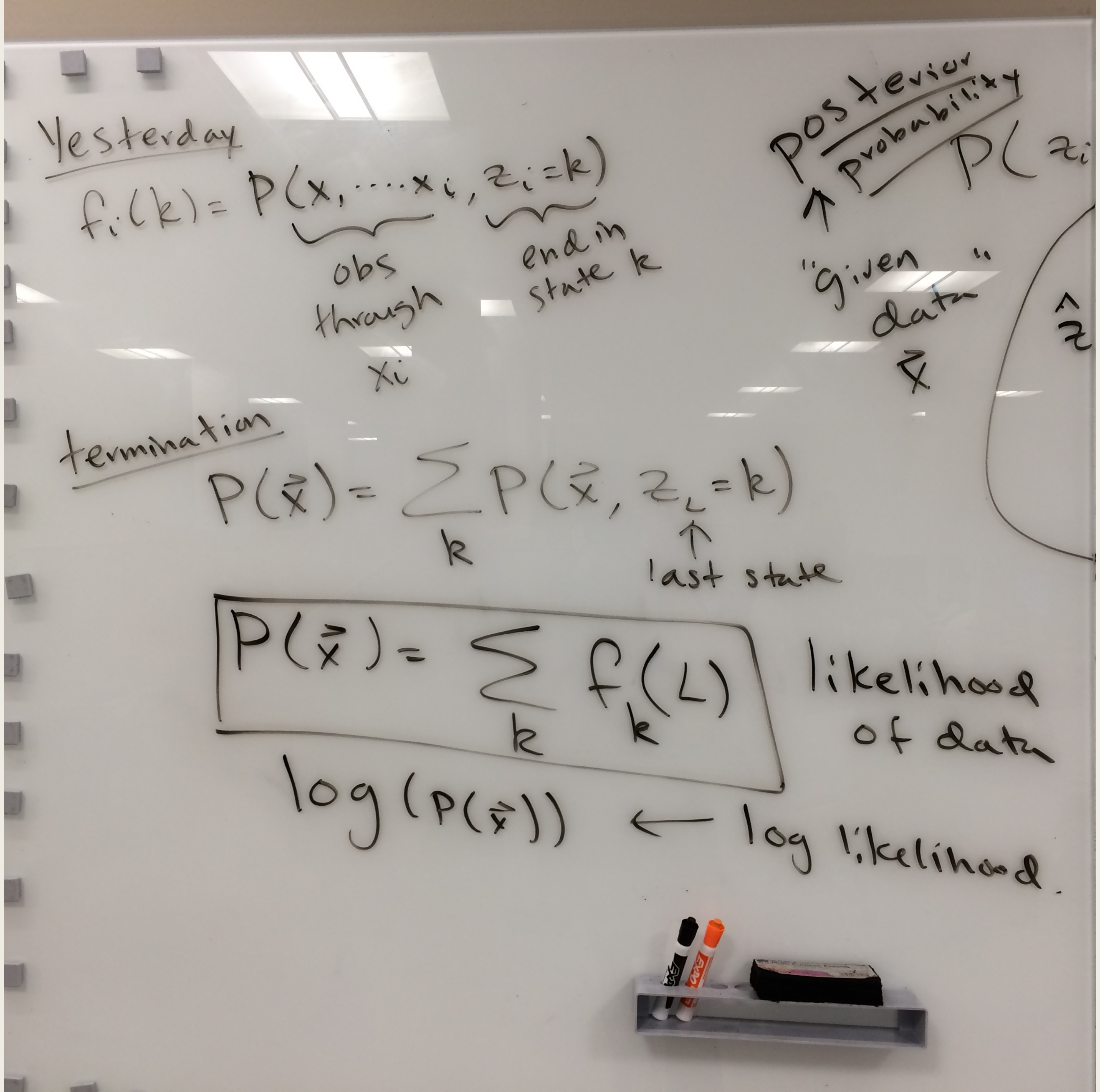
$$\log(p+q)$$

$$= \log\left(e^{\tilde{p}} + e^{\tilde{q}}\right)$$

$$= \log\left(e^{\tilde{p}} \cdot \left(1 + e^{\tilde{q}-\tilde{p}}\right)\right)$$

$$= \boxed{\tilde{p} + \log\left(1 + e^{\tilde{q}-\tilde{p}}\right)}$$

Lab B


Yesterday
$f_i(k) = P(x_1 \cdots x_i, z_i = k)$
obs through x_i
end in state k
termination
$P(\vec{x}) = \sum_k P(\vec{x}, z_L = k)$
last state
$P(\vec{x}) = \sum_k f_k(L)$
likelihood of data
$\log(P(\vec{x}))$ ← log likelihood
posterior probability
$P(z_i$
"given data"
$\vec{x}$
$\hat{z}$

posterior probability

$$P(z_i = k \mid \vec{x}) = \frac{f_k(i)\, b_k(i)}{P(\vec{x})}$$

"given data" $\vec{x}$

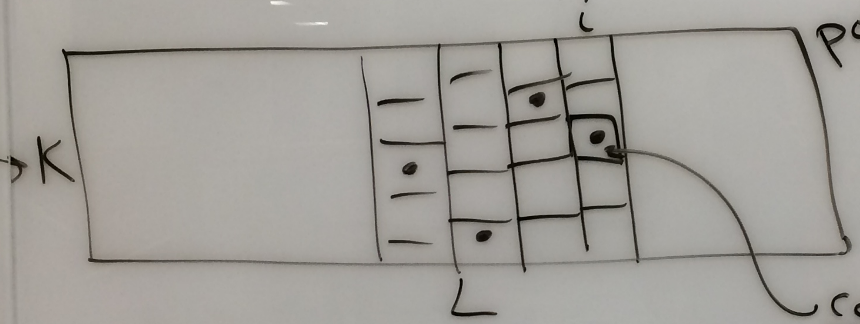
$$\hat{z}_i = \underset{k}{\arg\max}\; P(z_i = k \mid \vec{x})$$

k)

state

likelihood of data

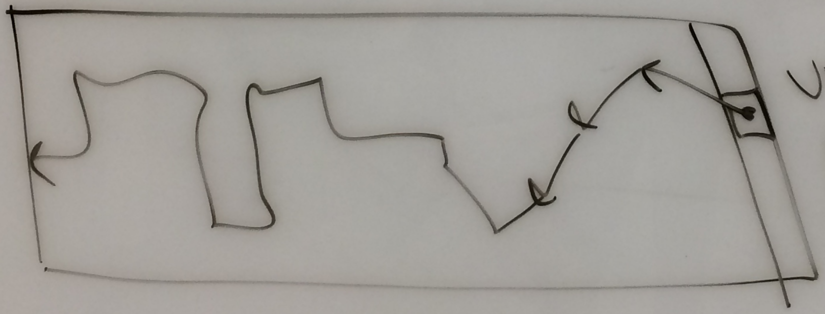
log likelihood



posterior decoding

$$\boxed{\hat{z}} = \underbrace{[0, 2, 3, 1, 0, 2]}_{L}$$

column max



Viterbi

$$\boxed{z^*} = \underbrace{[1, 2, 0, 3, 2, 1]}_{L}$$

posterior mean

posterior mean

$$\bar{g}_i =$$

posterior mean

$g(k) = $ coal. time of state k

$g(2) = 4.54$

posterior mean

$$\bar{g}_i = \sum_k P(z_i = k \mid \vec{x})\, g(k)$$

$$\underbrace{[0,2,3,1,0,2]}_{L}$$

$$\underbrace{[1,2,0,3,2,1]}_{L}$$

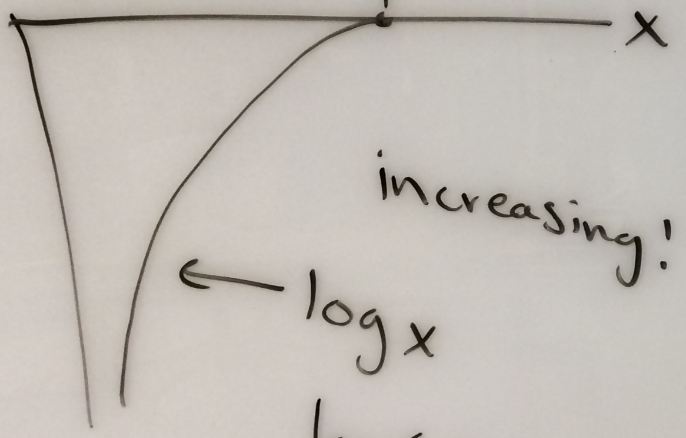
Order

- Viterbi
  - → check path + graph
- forward
  - → check log-likelihood
- backward
- posterior
  - ⇒ decoding
  - ⇒ mean

log-space

$$\log\left(V_k(i) = e_k(x_i) \cdot \overbrace{\max_\ell \left\{ V_\ell(i-1)\, a_{\ell k} \right\}}^{\text{becomes } 0!}\right)$$

$$\tilde{V}_k(i) = \log(e_k(x_i)) + \max_\ell \left\{ \tilde{V}_\ell(i-1) + \log(a_{\ell k}) \right\}$$

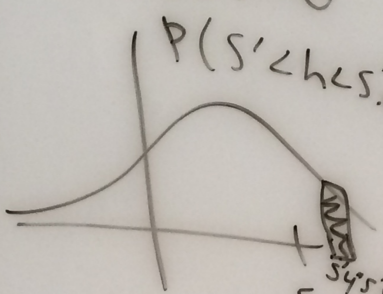


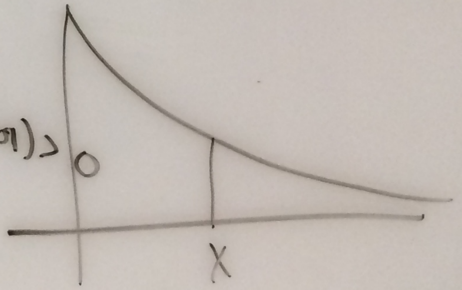
log(0) is bad!

$$\log(p+(q+r))$$

P(5')=0

P(5'<h<5'.001)>0





$$\log\left(f_k(i) = e_k(x_i) \cdot \sum_{l} f_l(i-1)\, a_{lk}\right)$$

become 0!

$$\log(p+q) \qquad \boxed{\tilde{p} = \log p,\ \tilde{q} = \log q}$$

$$= \log\left(e^{\tilde{p}} + e^{\tilde{q}}\right) = \log\left(e^{\tilde{p}} \cdot \left(1 + e^{\tilde{q}-\tilde{p}}\right)\right)$$

$$= \boxed{\tilde{p} + \log\left(1 + e^{\tilde{q}-\tilde{p}}\right)}$$