

Language Models



Dan Klein
UC Berkeley

Neural Language Models



Bigram Models

sent $\rightarrow P(\text{sent})$
 \uparrow
 $P(w_i | w_{i-1}, w_{i-2})$

w'



$$\hat{P}(w|w')$$

first the
 $C(w, w') - d$
 $\sum_v C(v, w')$
* the

$C(\text{the})$



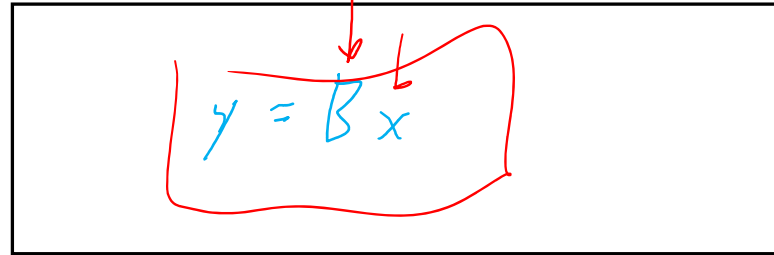
$P(w|w')$



Bigram Models

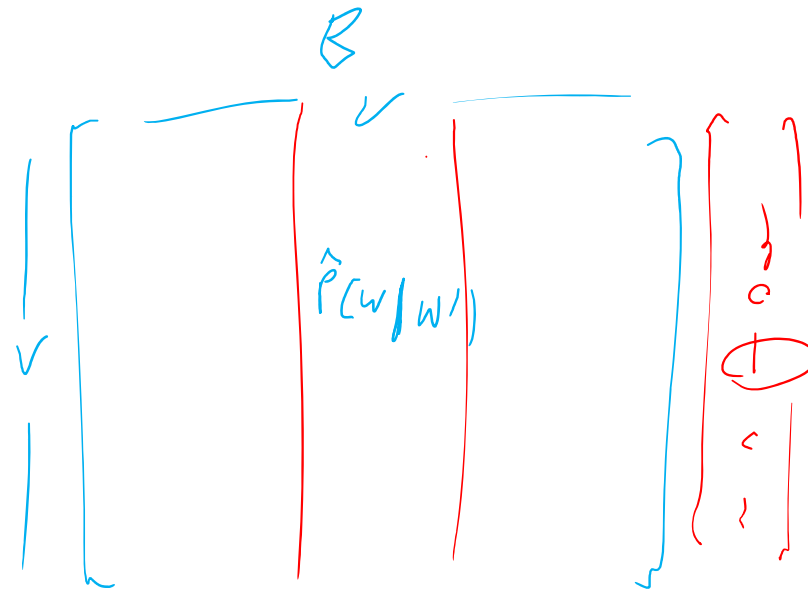
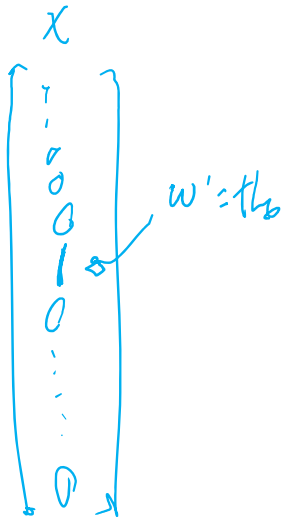
w'

x



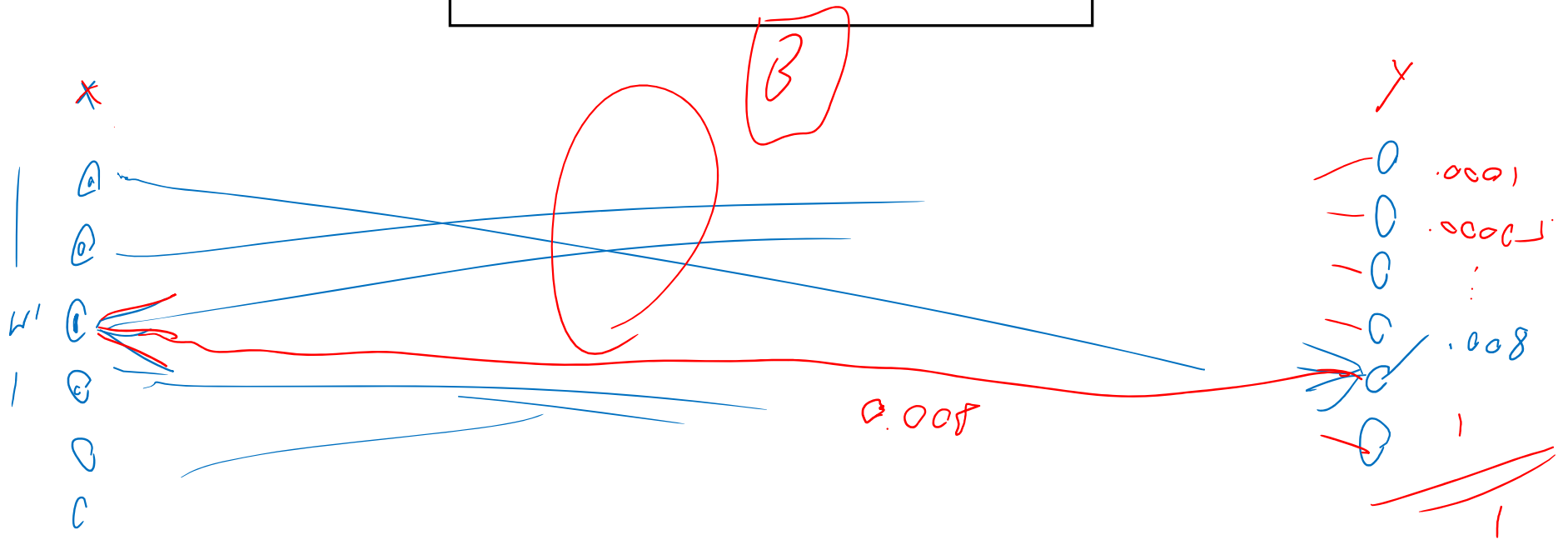
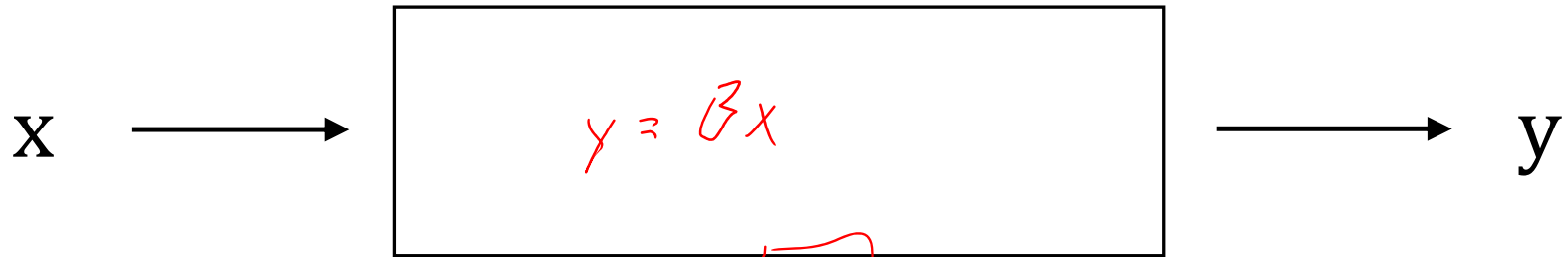
$P(w|w')$

y





Bigram Models





Learning a Model

Model $y = \beta x$

Data $W = w_1, w_2, w_3, w_4, \dots$

$$\max_{\beta} L(W|\beta)$$

$$L(W|\beta) = \prod_i P(w_i | w_{i-1}) = \prod b_{w_i, w_{i-1}}$$

X under flow

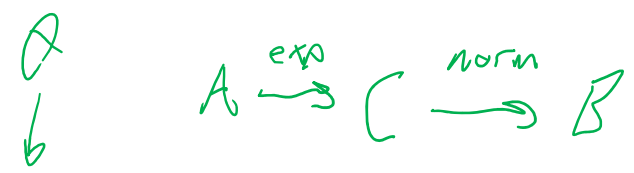
$$\max_{\beta} \mathcal{L}(W|\beta)$$

$$\mathcal{L}(W|\beta) = \sum_i \log P(w_i | w_{i-1}) = \sum_i \log b_{w_i, w_{i-1}}$$

X not probs

$$= \sum_i \log \frac{c_{w_i, w_{i-1}}}{\sum_u c_u w_{i-1}}$$

X ≥ 0

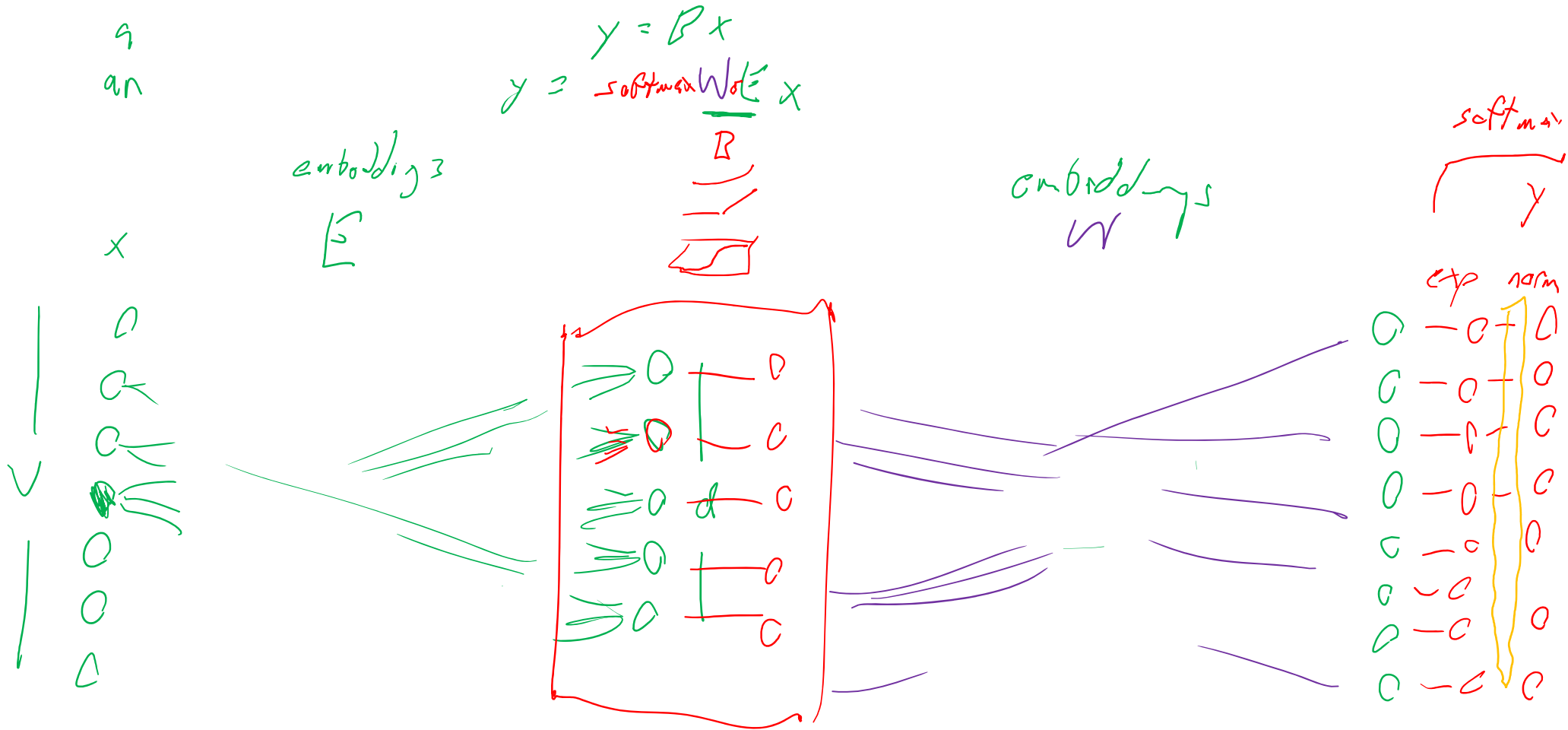


$$\max_{\beta} \mathcal{L}(W|A)$$

$$= \sum_i \log \frac{e^{a_{w_i, w_{i-1}}}}{\sum_u e^{a_u w_{i-1}}} \quad a = \theta$$



Embeddings and Generalization





Some Efficiency Issues

$$y = \sum_{i=1}^n W_i x_i$$

the

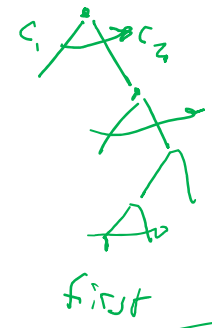
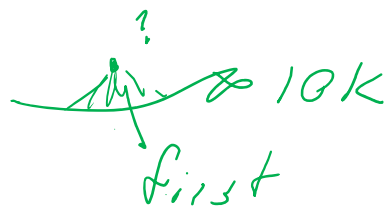
$$\max_{\theta} LL(D|\theta) =$$

~~$\log P(w|w, \dots)$~~

$$P(w | \dots)$$

$$P(\text{other} | \dots)$$

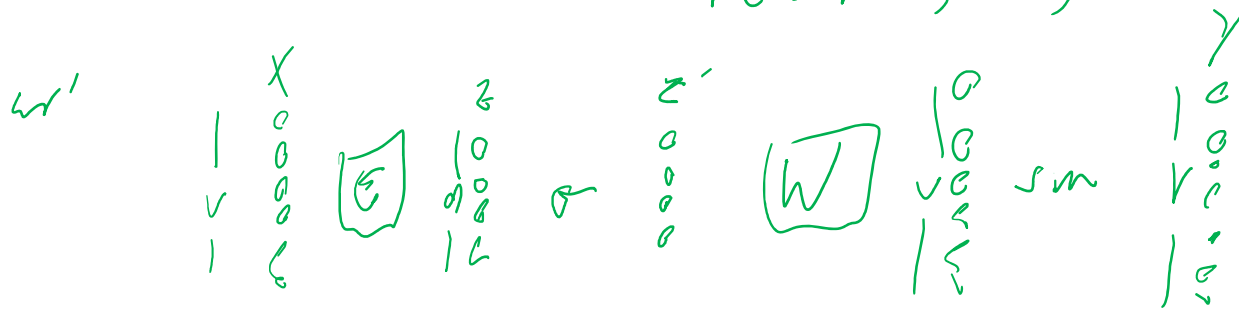
first





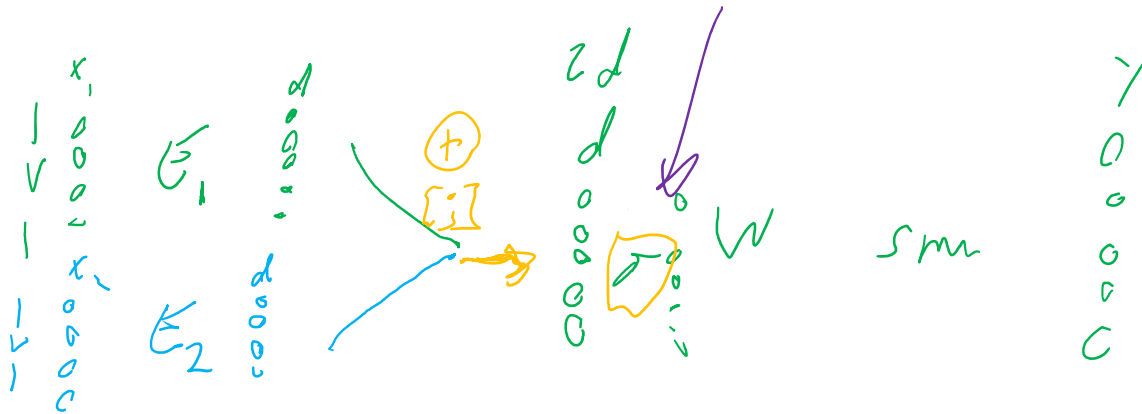
Neural N-Gram Models

$$P(w/w', w'')$$



$$P(w/w')$$

w', w''



$$P(? | the) = \text{first} \\ \text{sum} \\ \text{second}$$

$$P(? | may) = \text{be} \\ \text{do}$$

$$P(? | may the)$$