Language Models

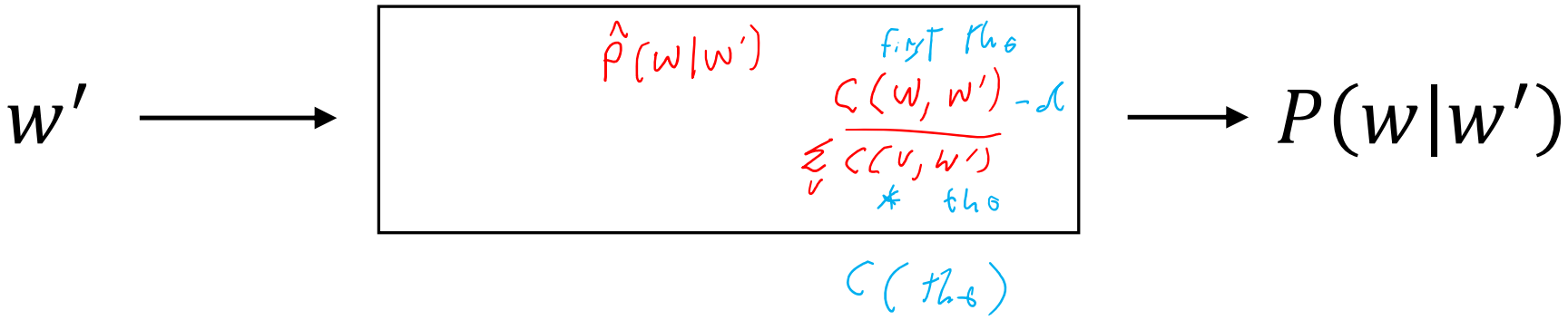
Dan Klein
UC Berkeley

Neural Language Models

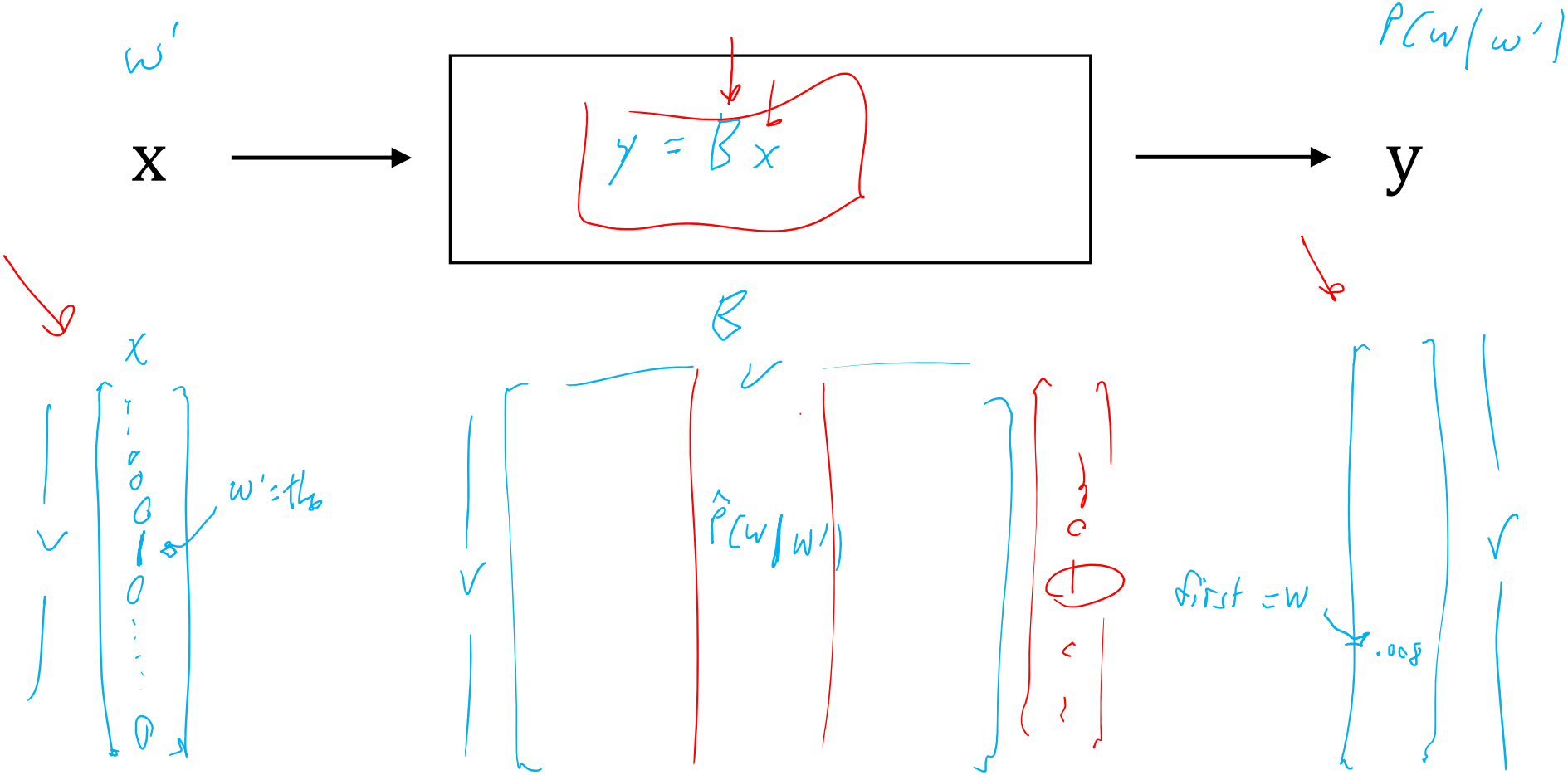
Bigram Models

sent → P(sent)

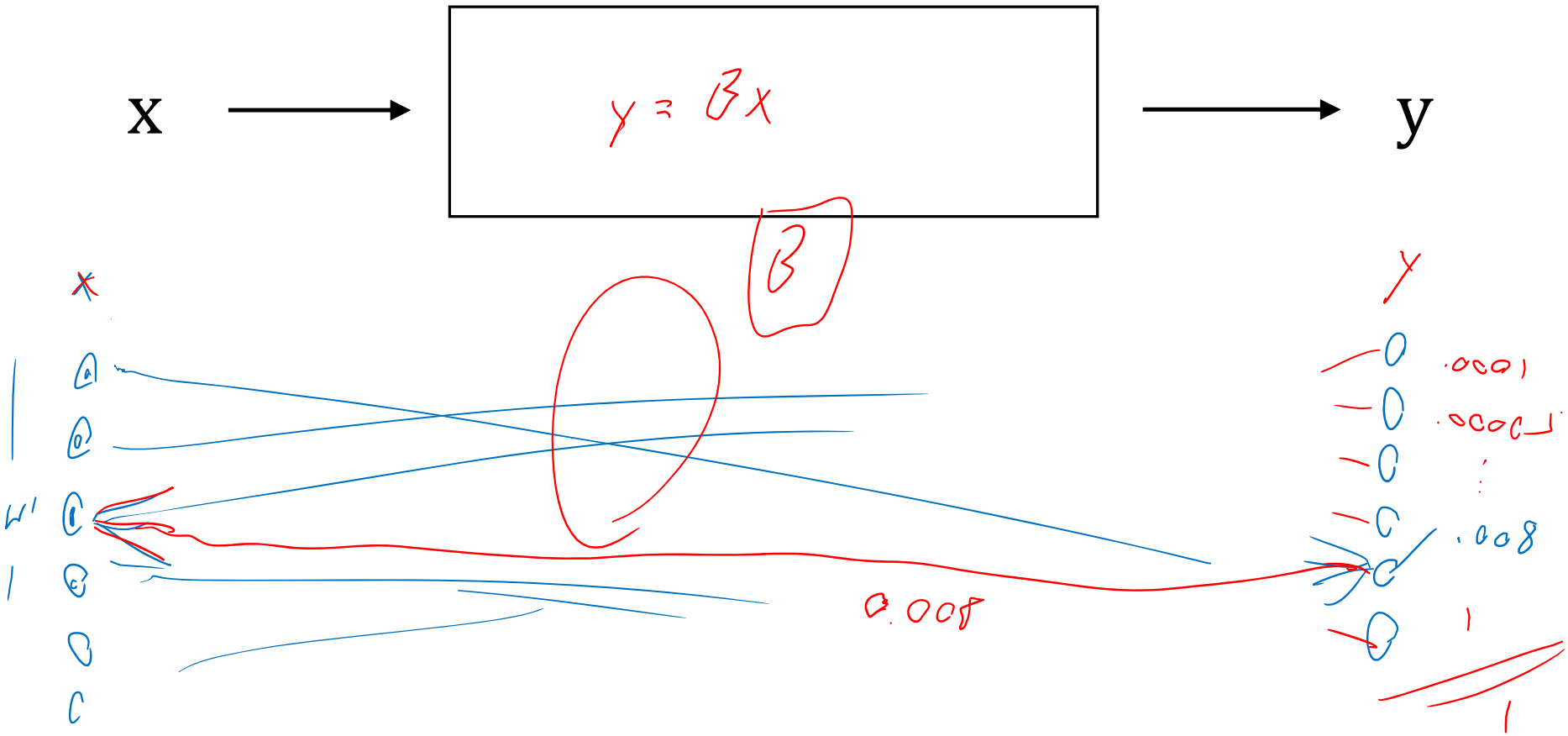
$\prod_i P(w_i \mid w_{i-1}, w_{i-2})$



Bigram Models



Bigram Models



Learning a Model

Model $y = Bx$

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

$\max_B L(W|B)$

$L(W|B) = \prod_i P(w_i|w_{i-1}) = \prod_i b_{w_i, w_{i-1}}$ ✗ underflow

$\max_B LL(W|B)$

$LL(W|B) = \sum_i \log P(w_i|w_{i-1}) = \sum_i \log b_{w_i, w_{i-1}}$ ✗ not probs

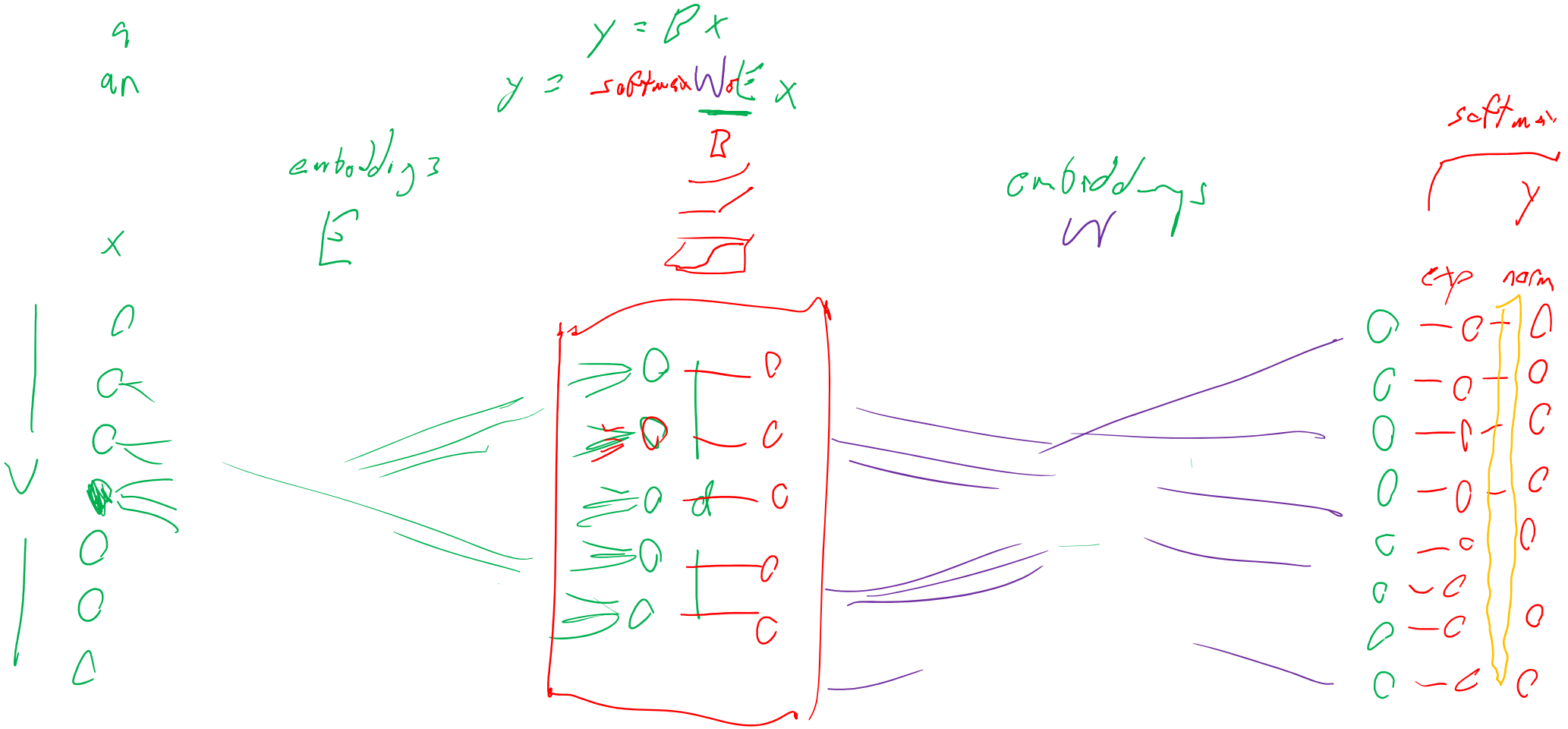
$= \sum_i \log \frac{c_{w_i, w_{i-1}}}{\sum_v c_{v, w_{i-1}}}$ ✗ ≥ 0

$A \xrightarrow{\exp} C \xrightarrow{\text{norm}} B$

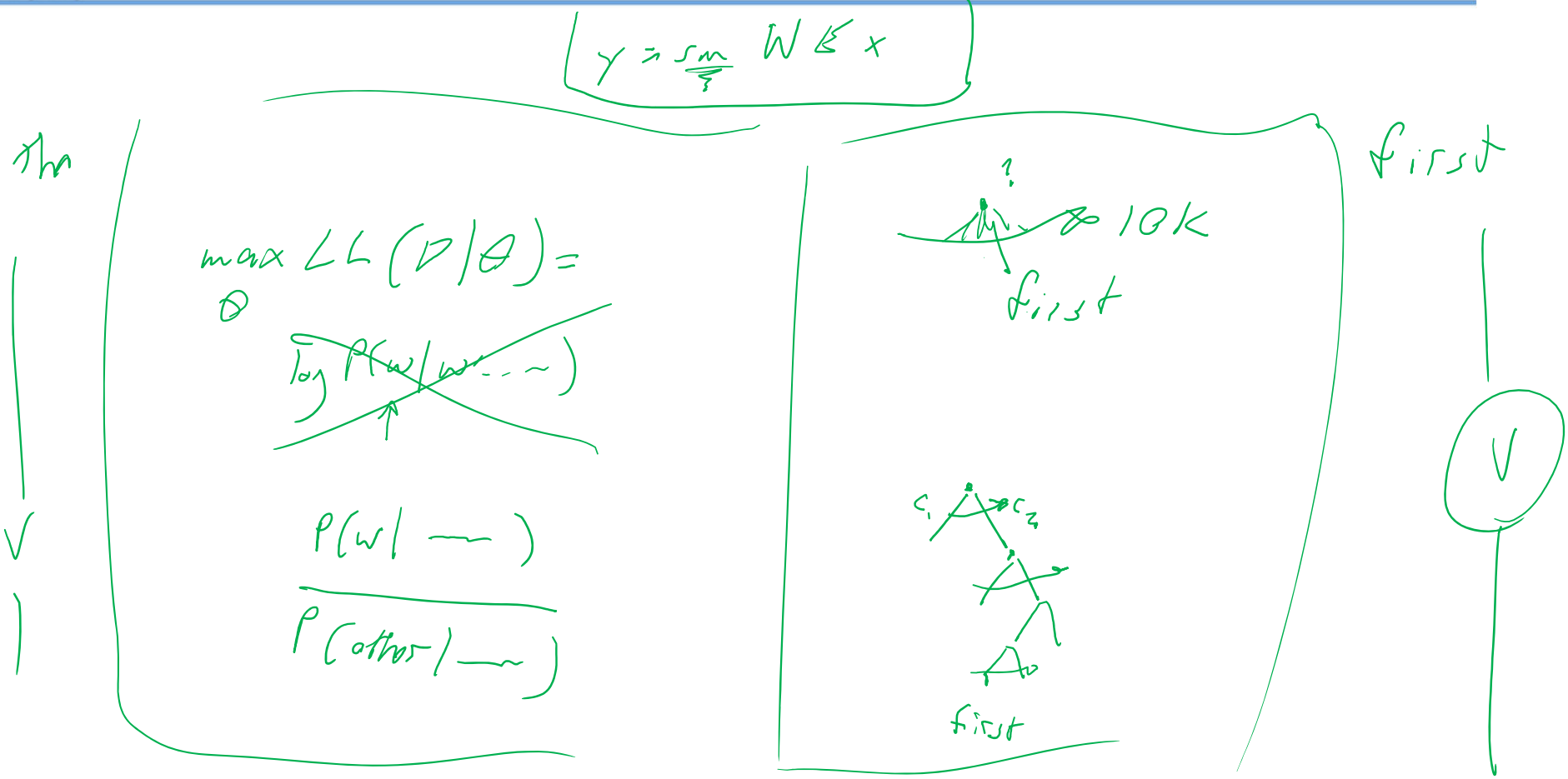
$\max_A LL(W|A)$

$= \sum_i \log \frac{e^{a_{w_i, w_{i-1}}}}{\sum_v e^{a_{v, w_{i-1}}}}$ $q = 0$

Embeddings and Generalization



Some Efficiency Issues



Neural N-Gram Models