Estimating *n*-gram probabilities

- The maximum likelihood estimate of an *n*-gram model:

$$P(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-n)}, \ldots, \mathbf{x}^{(t-1)}) \overset{MLE}{=} \frac{count\left(\mathbf{x}^{(t-n)}, \ldots, \mathbf{x}^{(t)}\right)}{\sum_{j=1}^{N} count\left(\mathbf{x}^{(t-n)}, \ldots, \mathbf{x}_j^{(t)}\right)}$$

Example from Jurafsky & Martin

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

Here are the calculations for some of the bigram probabilities from this corpus

$P(\texttt{I}|\texttt{<s>}) = \frac{2}{3} = .67 \qquad P(\texttt{Sam}|\texttt{<s>}) = \frac{1}{3} = .33 \qquad P(\texttt{am}|\texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>}|\texttt{Sam}) = \frac{1}{2} = 0.5 \qquad P(\texttt{Sam}|\texttt{am}) = \frac{1}{2} = .5 \qquad P(\texttt{do}|\texttt{I}) = \frac{1}{3} = .33$

Choosing the right *n*

- Choosing *n* is a trade off between:
  - Modeling capacity (increases with *n*)

Examples from Eisenstein (2018)

Gorillas always like to groom **their** friends.

The **computer** that's on the 3rd floor of our office building **crashed**.

  - Tractable estimation (harder with larger *n* … sparsity)

Example from Jurafsky & Martin

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Figure 3.1 Bigram counts for eight of the words (out of $V = 1446$) in the Berkeley Restaurant Project corpus of 9332 sentences. Zero counts are in gray.