# Sequence Modeling

Berkeley NLP

EECS 183/283a: Natural Language Processing

# Modeling Language

Modeling the <u>expectations</u> we have over utterances we encounter

- How can we find structure in continuous signals like speech?
- What words are more frequent vs. rare?
- What combinations of words are more likely vs. unlikely?
- What sequences of utterances are plausible vs. implausible?
- How likely is it for certain words (or combinations of words) to appear alongside different contexts vs. others?

# Modeling Sequences

- For now: let's assume utterances are sequences of tokens from our vocabulary

- Our vocabulary has a fixed size and consists of discrete wordtypes (we'll get to modeling continuous language signals, like speech, in a few weeks!)

- A sequence is denoted as:

$$\overline{x} = \langle x_1, \ldots, x_n \rangle \qquad x \in \mathcal{V} \qquad x_n = \texttt{EOS}$$

- We can also consider writing out all possible sequences given our vocabulary (though this set is infinitely large): $\mathcal{V}^+$

# Modeling Sequences

- A sequence model imposes a probability distribution over $\mathcal{V}^+$

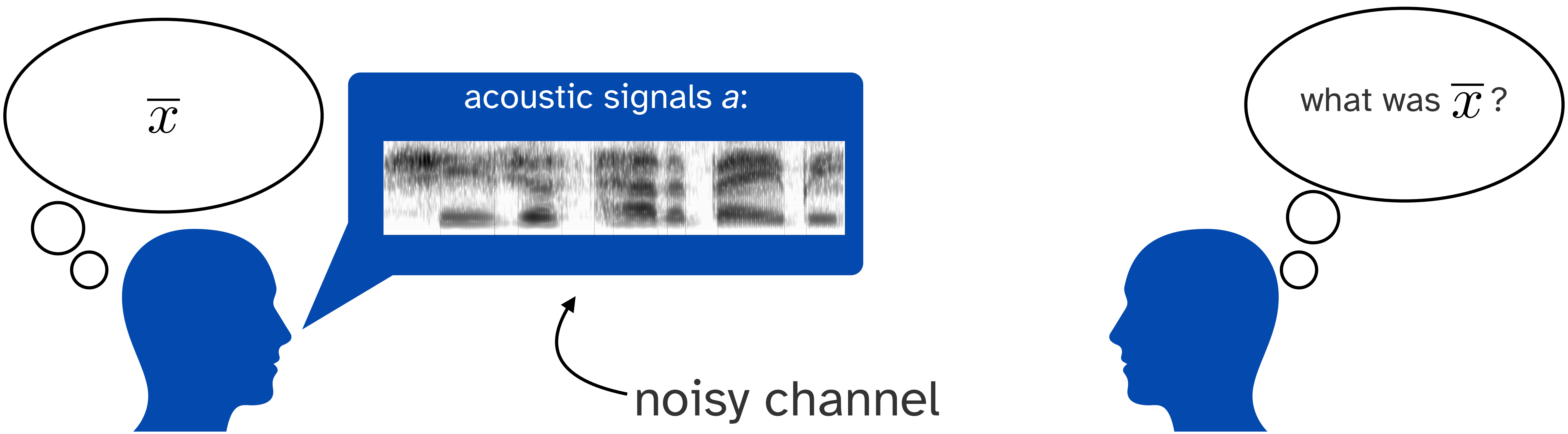
$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

# Modeling Sequences

- A sequence model imposes a probability distribution over $\mathcal{V}^+$

$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

- Why is this useful?



# Modeling Sequences

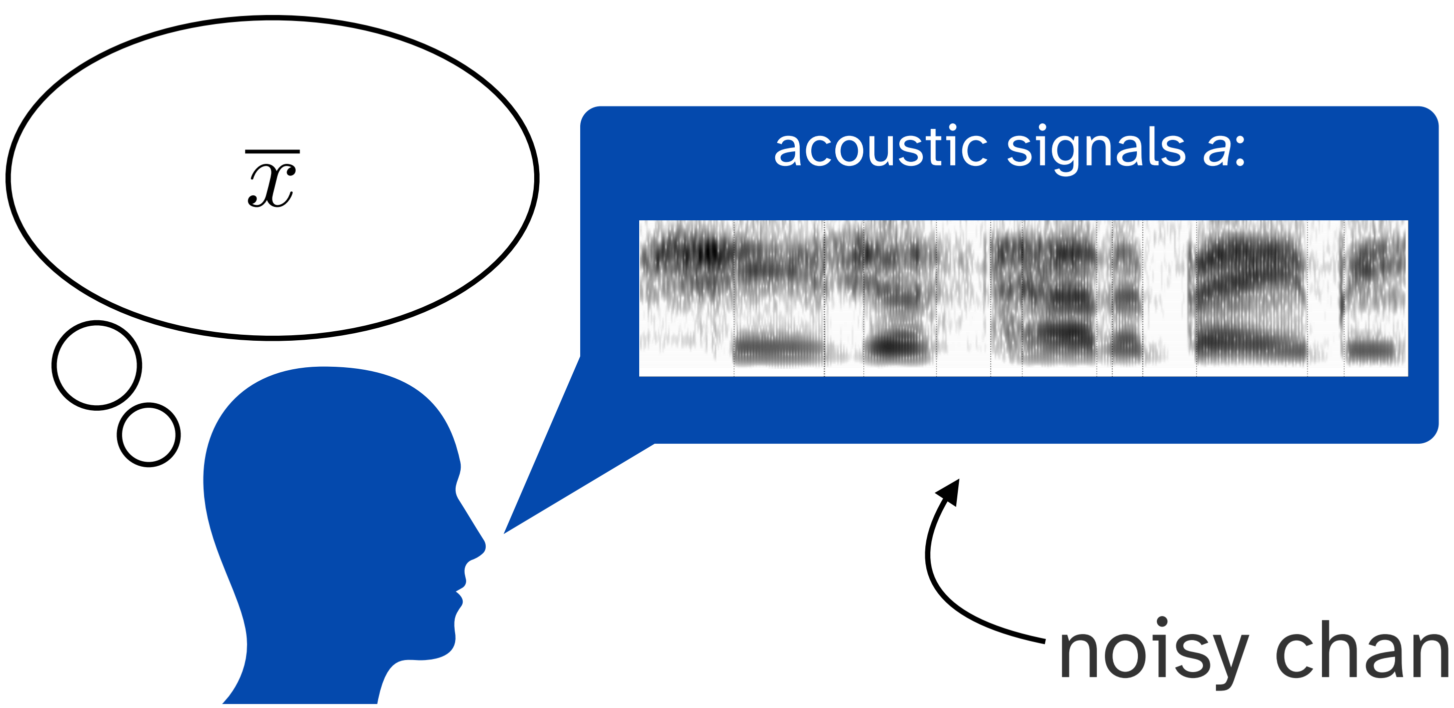
- A sequence model imposes a probability distribution over $\mathcal{V}^+$

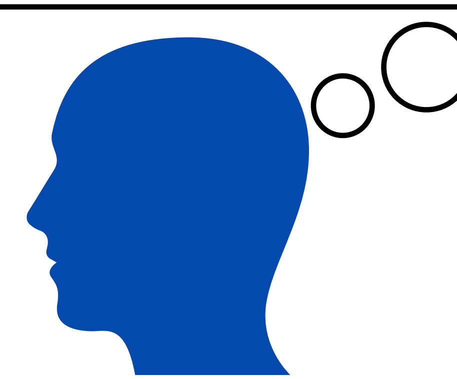
$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

- Why is this useful?

Bayes' rule

$$\begin{aligned}\overline{x}^* &= \arg\max_{\overline{x}} p(\overline{x} \mid a) \\ &= \arg\max_{\overline{x}} \frac{p(a \mid \overline{x})p(\overline{x})}{p(a)} \\ &= \arg\max_{\overline{x}} \underbrace{p(a \mid \overline{x})}_{\text{acoustic model}}\underbrace{p(\overline{x})}_{\text{language model}}\end{aligned}$$



# Modeling Sequences

- A sequence model imposes a probability distribution over $\mathcal{V}^+$

$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

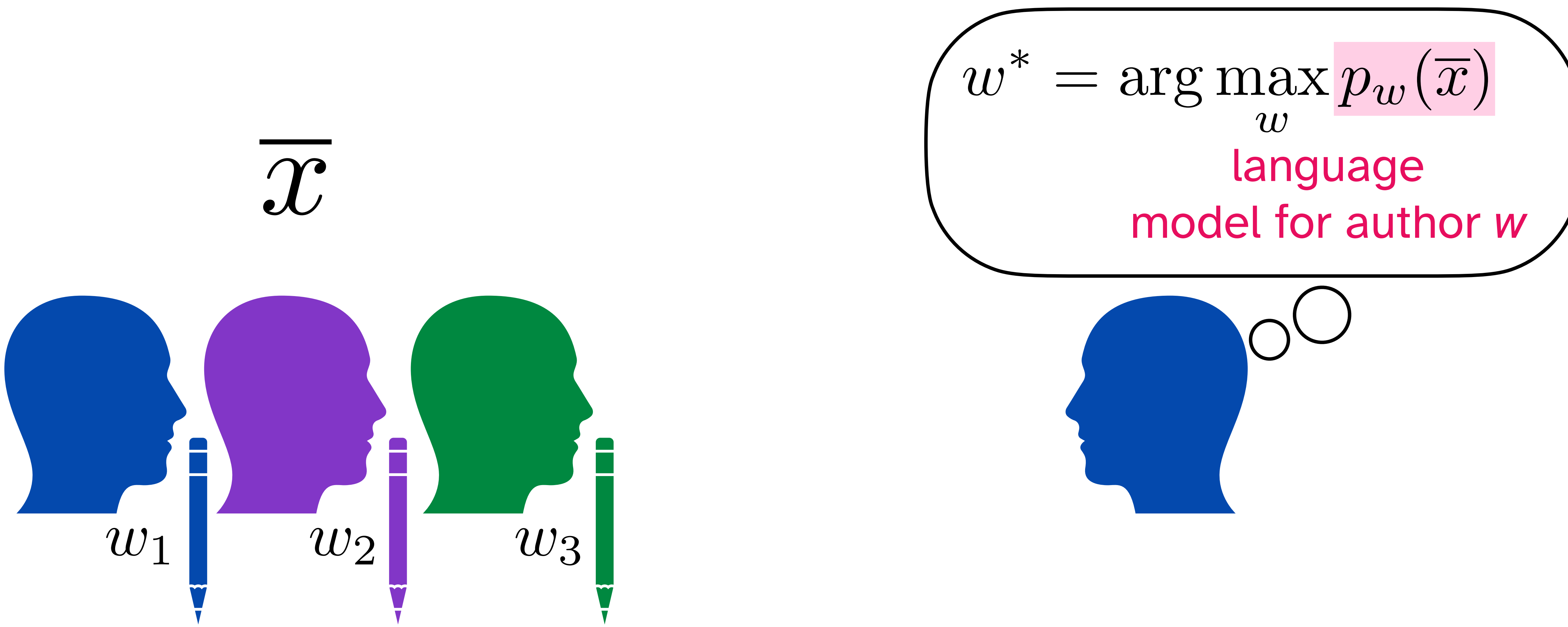
- Why is this useful?

# Modeling Sequences

- A sequence model imposes a probability distribution over $\mathcal{V}^+$

$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

- Why is this useful?

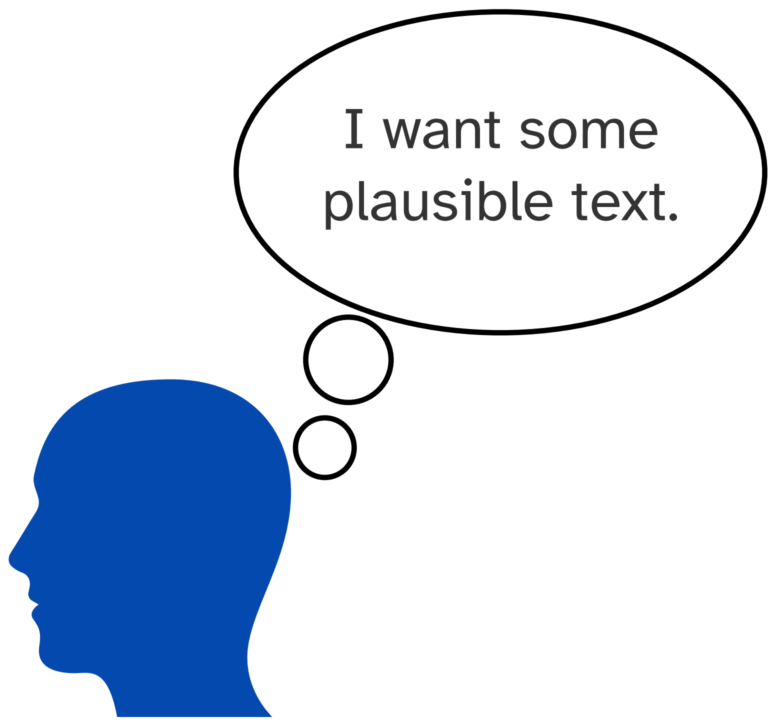


# Modeling Sequences

- A sequence model imposes a probability distribution over $\mathcal{V}^+$

$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

- Why is this useful?

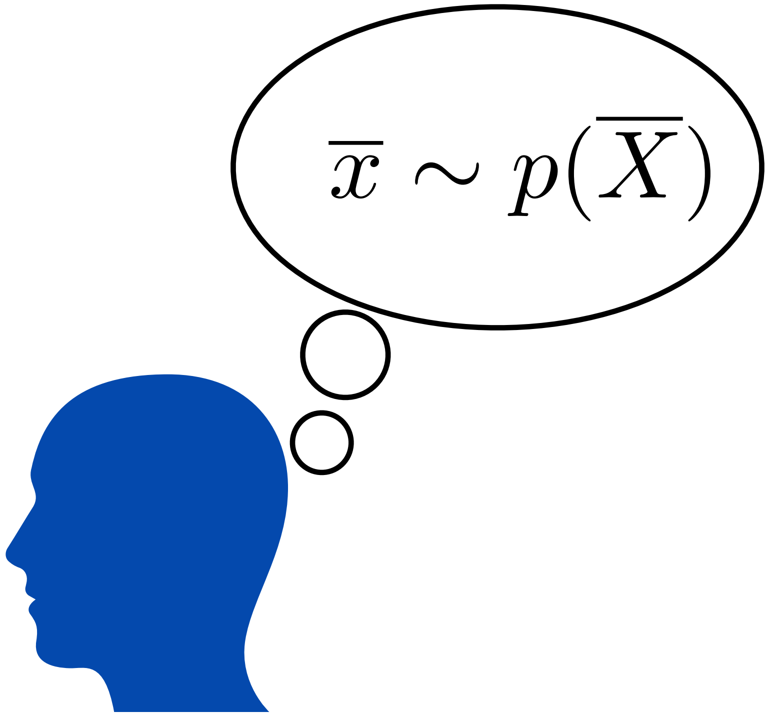


# Modeling Sequences

- A sequence model imposes a probability distribution over $\mathcal{V}^+$

$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

- Why is this useful?



# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l}\texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)}\end{array}$$

$$\mathcal{V} = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{u}\}$$

$$\mathcal{V}^{+} = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \ldots, \mathrm{bb}, \mathrm{bg}, \ldots, \mathrm{bug}, \mathrm{bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l}\texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)}\end{array}$$

$$\mathcal{V} = \{\mathrm{b, g, h, n, p, s, u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b, g, h}, \ldots, \mathrm{bb, bg}, \ldots, \mathrm{bug, bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Learning problem: we want a to estimate the probability distribution $p(\overline{X}) \in \Delta^{\mathcal{V}^+}$ that generated our observations $\mathcal{D}$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l} \texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)} \end{array}$$

$$\mathcal{V} = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \ldots, \mathrm{bb}, \mathrm{bg}, \ldots, \mathrm{bug}, \mathrm{bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Learning problem: we want a to estimate the probability distribution $p(\overline{X}) \in \Delta^{\mathcal{V}^+}$ that generated our observations $\mathcal{D}$

- One simple option: just count based on document occurrence

$$p(\overline{x}) = \frac{C(\overline{x})}{|\mathcal{D}|}$$

| $\overline{x}$ | $p(\overline{x}) = \frac{C(\overline{x})}{\vert\mathcal{D}\vert}$ |
|---|---|
| hug | `10/36 = 0.28` |
| pug | `5/36 = 0.14` |
| pun | `12/36 = 0.33` |
| bun | `4/36 = 0.11` |
| hugs | `5/36 = 0.14` |
| $\overline{x} \in \mathcal{V}^+ \setminus \mathcal{D}$ | `0/36 = 0.00` |

$p(\overline{X})$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l}\texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)}\end{array}$$

$$\mathcal{V} = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \ldots, \mathrm{bb}, \mathrm{bg}, \ldots, \mathrm{bug}, \mathrm{bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Let's sample from $p(\overline{X})$!

| $\overline{x}$ | $p(\overline{x}) = \frac{C(\overline{x})}{\lvert\mathcal{D}\rvert}$ |
|---|---|
| hug | `10/36 = 0.28` |
| pug | `5/36 = 0.14` |
| pun | `12/36 = 0.33` |
| bun | `4/36 = 0.11` |
| hugs | `5/36 = 0.14` |
| $\overline{x} \in \mathcal{V}^+ \setminus \mathcal{D}$ | `0/36 = 0.00` |

$p(\overline{X})$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l}\texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)}\end{array}$$

$$\mathcal{V} = \{\mathrm{b, g, h, n, p, s, u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b, g, h}, \ldots, \mathrm{bb, bg}, \ldots, \mathrm{bug, bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Let's sample from $p(\overline{X})$!

  0.98 → *hugs*

| $\overline{x}$ | $p(\overline{x}) = \frac{C(\overline{x})}{\vert\mathcal{D}\vert}$ |
|---|---|
| hug | 10/36 = 0.28 |
| pug | 5/36 = 0.14 |
| pun | 12/36 = 0.33 |
| bun | 4/36 = 0.11 |
| hugs | 5/36 = 0.14 |
| $\overline{x} \in \mathcal{V}^+ \setminus \mathcal{D}$ | 0/36 = 0.00 |

$p(\overline{X})$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l}\texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)}\end{array}$$

$$\mathcal{V} = \{\mathrm{b, g, h, n, p, s, u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b, g, h}, \ldots, \mathrm{bb, bg}, \ldots, \mathrm{bug, bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Let's sample from $p(\overline{X})$!

  0.98 → *hugs*

  0.84 → *bun*

| $\overline{x}$ | $p(\overline{x}) = \frac{C(\overline{x})}{\lvert\mathcal{D}\rvert}$ |
|---|---|
| hug | 10/36 = 0.28 |
| pug | 5/36 = 0.14 |
| pun | 12/36 = 0.33 |
| bun | 4/36 = 0.11 |
| hugs | 5/36 = 0.14 |
| $\overline{x} \in \mathcal{V}^+ \setminus \mathcal{D}$ | 0/36 = 0.00 |

$p(\overline{X})$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l} \texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)} \end{array}$$

$$\mathcal{V} = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \ldots, \mathrm{bb}, \mathrm{bg}, \ldots, \mathrm{bug}, \mathrm{bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Let's sample from $p(\overline{X})$!

  0.98 → *hugs*

  0.84 → *bun*

  0.55 → *pun*

| $\overline{x}$ | $p(\overline{x}) = \frac{C(\overline{x})}{\lvert\mathcal{D}\rvert}$ |
|---|---|
| hug | 10/36 = 0.28 |
| pug | 5/36 = 0.14 |
| pun | 12/36 = 0.33 |
| bun | 4/36 = 0.11 |
| hugs | 5/36 = 0.14 |
| $\overline{x} \in \mathcal{V}^+ \setminus \mathcal{D}$ | 0/36 = 0.00 |

$p(\overline{X})$

# Example: Count-Based Language Model

Documents + frequencies:

$$\mathcal{D} = \begin{array}{l}\texttt{('hug', 10), ('pug', 5),} \\ \texttt{('pun', 12), ('bun', 4),} \\ \texttt{('hugs', 5)}\end{array}$$

$$\mathcal{V} = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{u}\}$$

$$\mathcal{V}^+ = \{\mathrm{b}, \mathrm{g}, \mathrm{h}, \ldots, \mathrm{bb}, \mathrm{bg}, \ldots, \mathrm{bug}, \mathrm{bun}, \ldots, \mathrm{sssssss}, \ldots\}$$

- Let's sample from $p(\overline{X})$!

  0.98 → *hugs*

  0.84 → *bun*

  0.55 → *pun*

- This directly generates our observation data! But nobody ever does this
- Why not?

| $\overline{x}$ | $p(\overline{x}) = \frac{C(\overline{x})}{\lvert\mathcal{D}\rvert}$ |
|---|---|
| hug | 10/36 = 0.28 |
| pug | 5/36 = 0.14 |
| pun | 12/36 = 0.33 |
| bun | 4/36 = 0.11 |
| hugs | 5/36 = 0.14 |
| $\overline{x} \in \mathcal{V}^+ \setminus \mathcal{D}$ | 0/36 = 0.00 |

$p(\overline{X})$

# Language Modeling is Hard

$$p(\overline{X}) \in \Delta^{\mathcal{V}^+} \qquad \overline{x} \in \mathcal{V}^+ \qquad p(\overline{x})$$

- How might we go about assigning a probability to **any possible sequence,** even ones we've never seen before?
- One intuition: sentences have internal consistencies!

*você fala inglês?*
*do you speak English?*

| você | fala |
|---|---|
| 2nd person singular pronoun, formal | 3rd person singular present indicative |

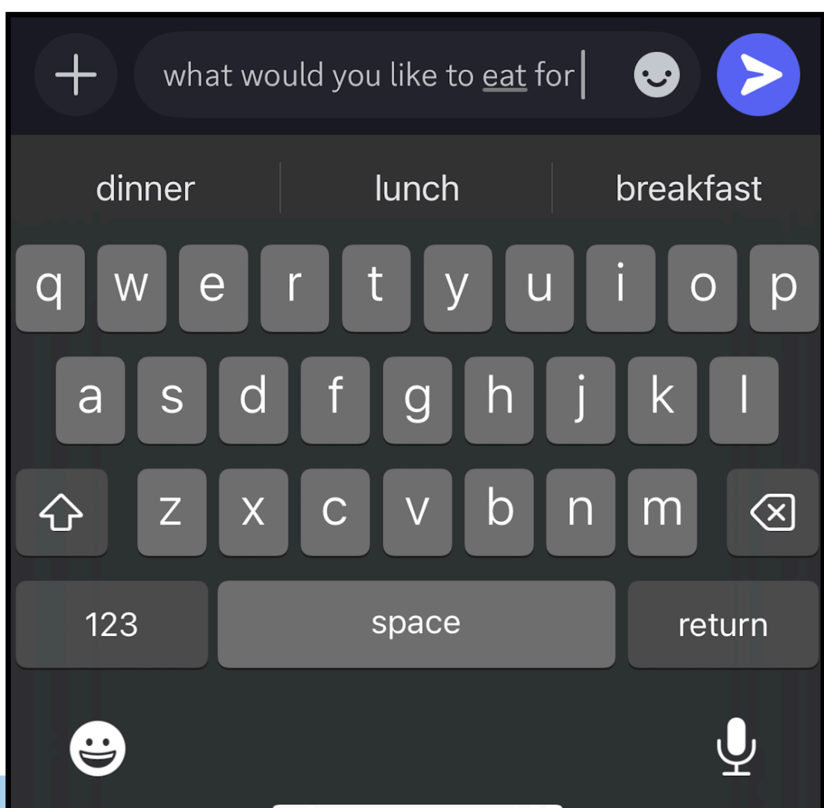
| 3rd person singular pronoun, masculine | accusative case of "dog" | 3rd person singular present, masculine |
|---|---|---|
| అతను | కుక్కను | చూస్తాడు |
| atanu | kukkanu | cūstāḍu |
| he | the dog | sees |

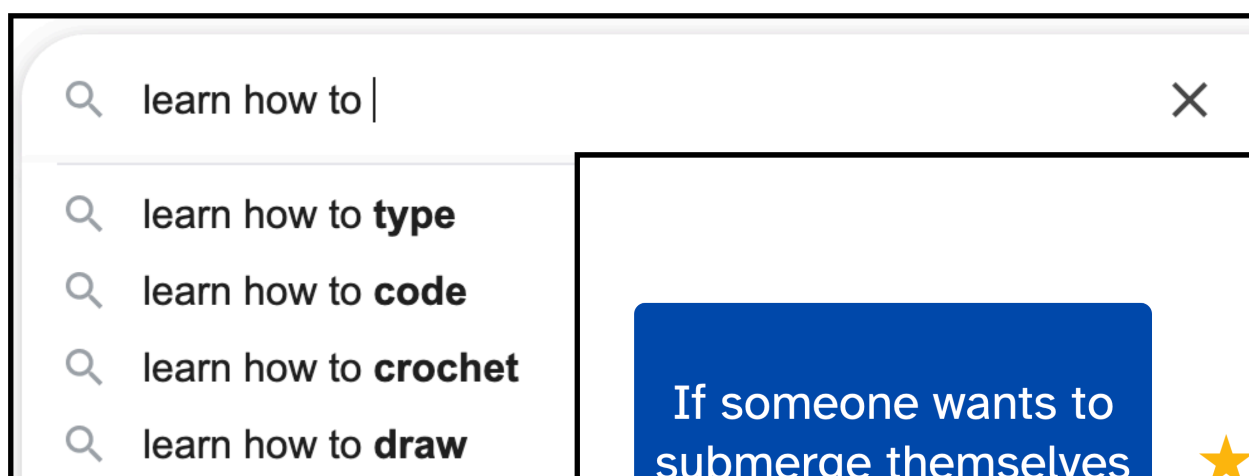
# One Approximation

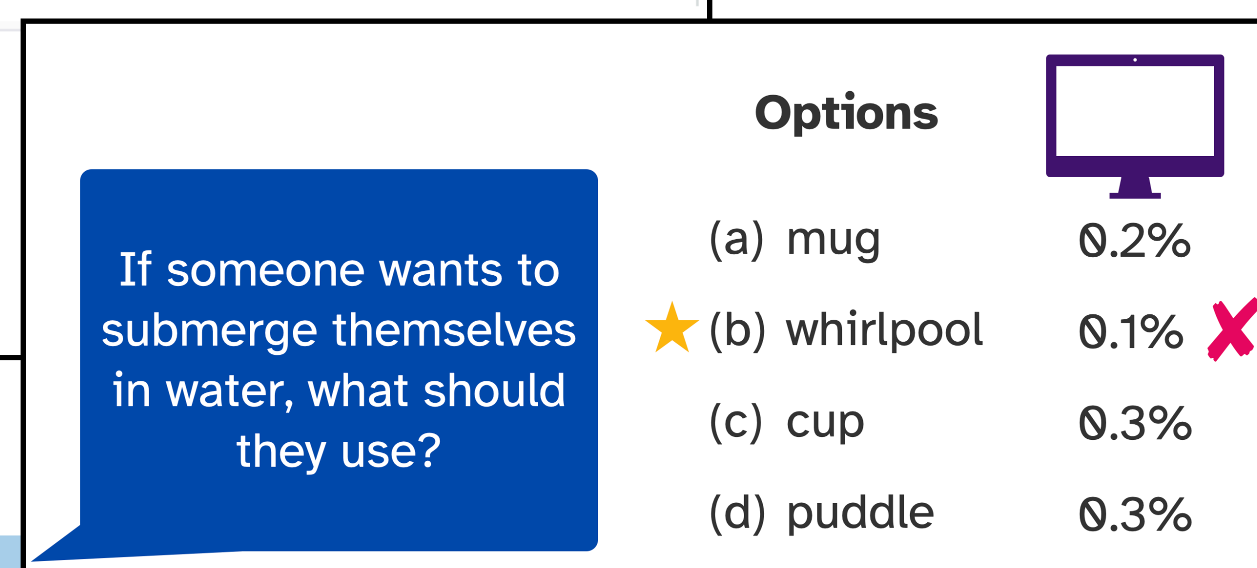
**Autoregressive language modeling:**

- The probability of a sequence is a product of local token probabilities
- The probability of a token depends on the ones that came before it

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$



# Another Approximation

**Masked language modeling:**

- The probability of a sequence is a product of local token probabilities
- The probability of a token depends on the ones that came before *and after* it

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1} \ldots x_n)$$

The name *anteater* refers to the species' [redacted], which consists mainly of ants and termites.

This approximation will come up later in the class! For now, we'll focus on autoregressive language modeling.

(There are other approximations too!)

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

**Chain rule decomposition**

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$

$p(X_1)$

| $x$ | $p(x)$ |
|---|---|
| the | 0.04 |
| in | 0.02 |
| and | 0.02 |
| every | 0.02 |
| … | |

$\overline{x} = \langle$ *the*

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$
- Sample the second word $x_2 \sim p(X_2 \mid x_1) \in \Delta^{\mathcal{V}}$

$p(X_2 \mid x_1 = \text{the})$

| $x$ | $p(x)$ |
|---|---|
| -ir | 0.03 |
| same | 0.02 |
| impact | 0.02 |
| line | 0.02 |
| … | |

$\overline{x} = \langle$ *the, same*

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$
- Sample the second word
- Sample the third word

$p(X_3 \mid \langle \text{the}, \text{same} \rangle)$

| $x$ | $p(x)$ |
|---|---|
| as | 0.08 |
| way | 0.04 |
| thing | 0.04 |
| time | 0.04 |
| ... | |

$\overline{x} = \langle$ *the, same, thing*

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$ $p(X_i \mid \langle x_1, \ldots, x_{i-1} \rangle)$
- Sample the second word
- Sample the third word
- Keep sampling until we hit `EOS`

$\overline{x} = \langle$ *the, same, thing, that*

| $x$ | $p(x)$ |
|---|---|
| . | 0.11 |
| as | 0.08 |
| that | 0.07 |
| , | 0.06 |
| ... | |

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$ $p(X_i \mid \langle x_1, \ldots, x_{i-1} \rangle)$
- Sample the second word
- Sample the third word
- Keep sampling until we hit EOS

$\overline{x} = \langle$ *the, same, thing, that, the*

| $x$ | $p(x)$ |
|---|---|
| happened | 0.11 |
| happens | 0.08 |
| makes | 0.07 |
| the | 0.06 |
| … | |

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1)\ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$ $p(X_i \mid \langle x_1, \ldots, x_{i-1} \rangle)$
- Sample the second word
- Sample the third word
- Keep sampling until we hit $\mathrm{EOS}$

$\overline{x} = \langle$ *the, same, thing, that, the, -y*

| $x$ | $p(x)$ |
|---|---|
| -y | 0.03 |
| -ir | 0.02 |
| person | 0.02 |
| line | 0.02 |
| ... | |

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}} p(X_i \mid \langle x_1, \ldots, x_{i-1} \rangle)$
- Sample the second word
- Sample the third word
- Keep sampling until we hit $\texttt{EOS}$

$\overline{x} = \langle$ *the, same, thing, that, the, -y, had*

| $x$ | $p(x)$ |
|---|---|
| are | 0.12 |
| were | 0.05 |
| have | 0.04 |
| had | 0.02 |
| ... | |

# Sampling from an Autoregressive Language Model

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

**Let's sample a sequence from this approximation:**

- Sample the first word $x_1 \sim p(X_1) \in \Delta^{\mathcal{V}}$ $p(X_i \mid \langle x_1, \ldots, x_{i-1} \rangle)$
- Sample the second word
- Sample the third word
- Keep sampling until we hit `EOS`

| $x$ | $p(x)$ |
|---|---|
| done | `0.08` |
| said | `0.04` |
| EOS | `0.04` |
| to | `0.02` |
| … | |

$\overline{x} = \langle$ *the, same, thing, that, the, -y, had, EOS* $\rangle$

*the same thing that they had*

# Autoregressive Language Models

$$p(\overline{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$= p(x_1)\,p(x_2 \mid x_1) \cdots p(x_n \mid x_1, \ldots, x_{n-1})$$

probability that the first word is $x_1$

probability that the second word is $x_2$, given that the first word is $x_1$

probability that the sentence ends after the sequence $\langle x_1, \ldots, x_{n-1} \rangle$

**Core modeling challenge:**

How do we compute these conditional probabilities?