Closing the Curious Case of Neural Text Degeneration

Matthew Finlayson[1] John Hewitt[2] Alexander Koller[3]
Swabha Swayamdipta[1] Ashish Sabharwal[4]

[1]USC [2]Stanford [3]Saarland University [4]AI2

January 18, 2024

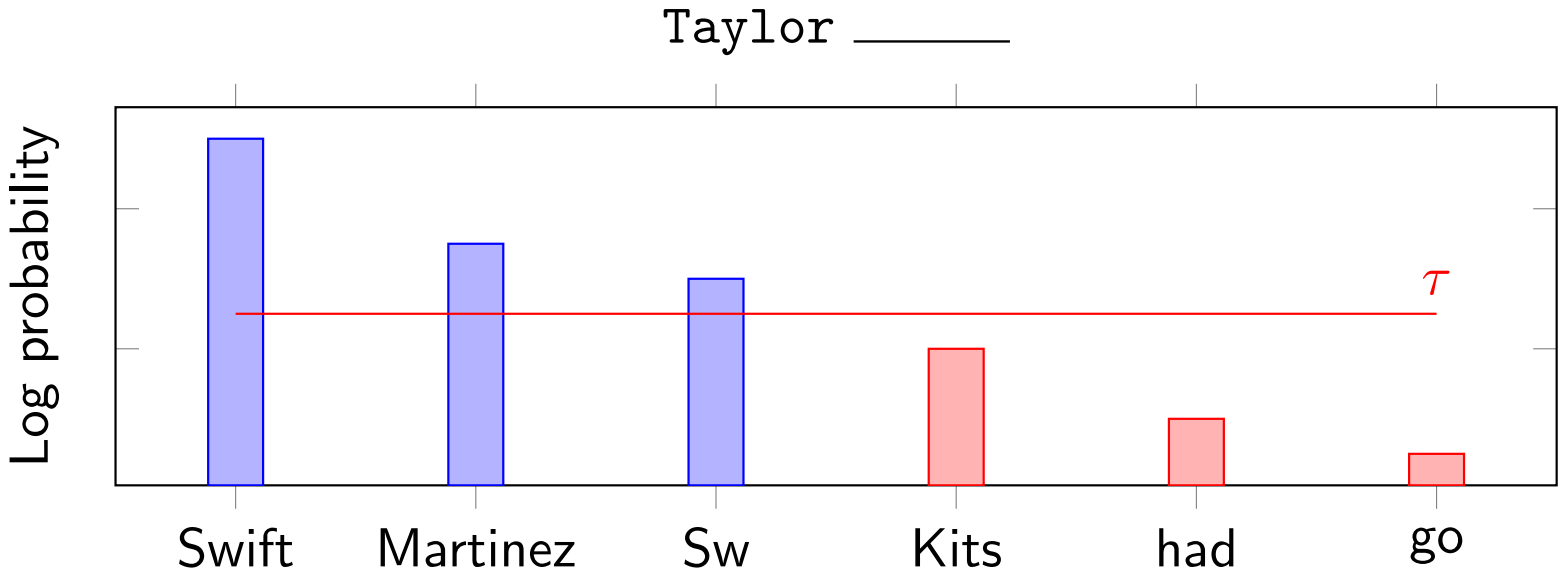
Accepted to ICLR 2024

Goal

- Improve threshold sampling methods (e.g., top-k) by directly addressing the source of errors.

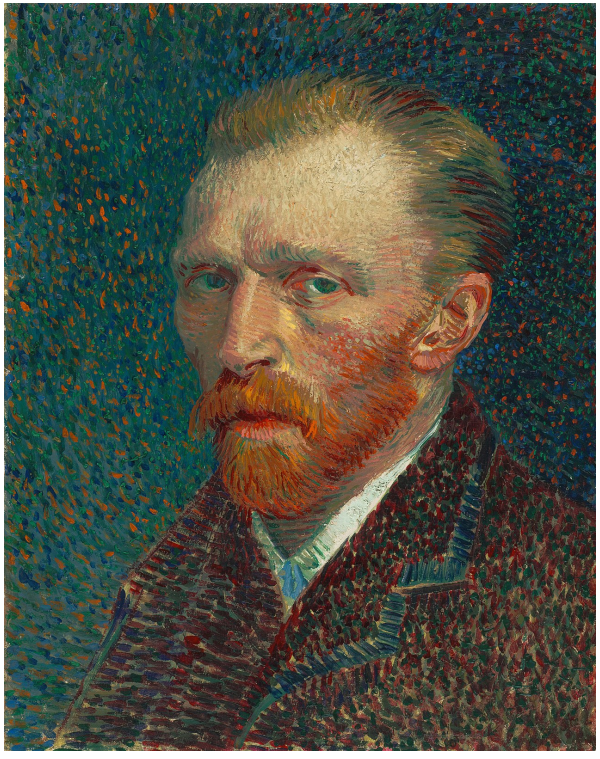
Threshold sampling

E.g., top-k, top-p



Choose a threshold τ and only sample tokens with probability greater than τ.

Van Gogh



Van Gogh

The Art Institute of Chicago Recreates Van Gogh's Famous Bedroom to be Rented on Airbnb

FEBRUARY 9, 2016

KATE SIERZPUTOWSKI

We don't know what Van Gogh's real bedroom looked like

What color is the floor? Is the towel brown? Is the pitcher glass?

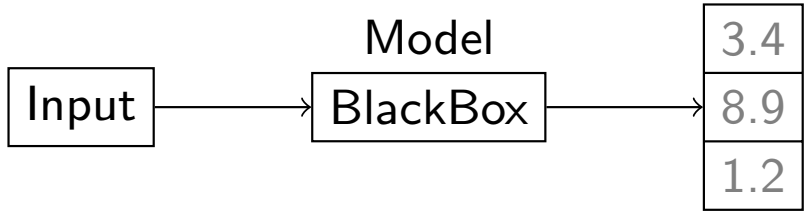
We don't know what Van Gogh's real bedroom looked like

What color is the floor? Is the towel brown? Is the pitcher glass?

The Yellow House, building destroyed June 25, 1944 by Allied bombing in France (WWII)

What does this have to do with LM decoding?

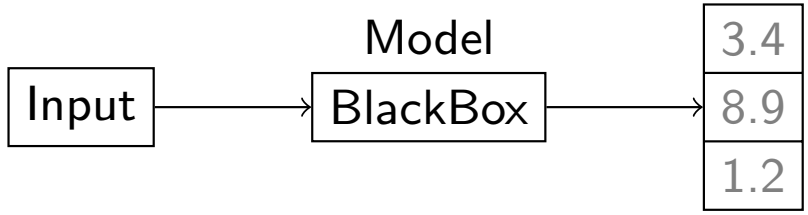
A language model is like Van Gogh: it has a limited palatte.



Contextualized embedding $\boldsymbol{h} \in \mathbb{R}^d$

What does this have to do with LM decoding?

A language model is like Van Gogh: it has a limited palatte.



Contextualized embedding $\boldsymbol{h} \in \mathbb{R}^d$

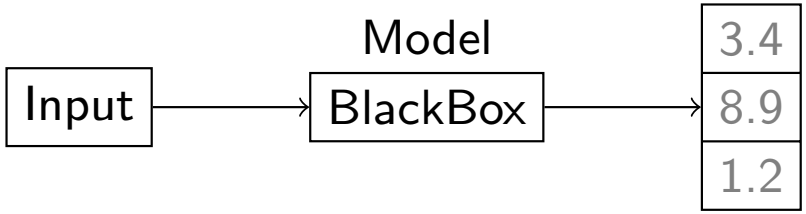
The size d of the embedding means the model paints with d colors.

8.8	4.1	2.3	1.2	6.3	8.5
0.3	2.4	4.1	4.3	0.5	4.2
1.5	2.4	2.4	9.3	3.9	8.4

Softmax matrix $\boldsymbol{W} \in \mathbb{R}^{v \times d}$

What does this have to do with LM decoding?

A language model is like Van Gogh: it has a limited palatte.



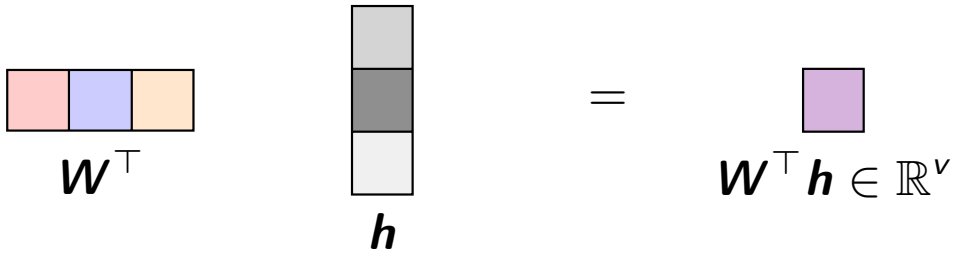
Contextualized embedding $\boldsymbol{h} \in \mathbb{R}^d$

The size d of the embedding means the model paints with d colors.

8.8	4.1	2.3	1.2	6.3	8.5
0.3	2.4	4.1	4.3	0.5	4.2
1.5	2.4	2.4	9.3	3.9	8.4

Softmax matrix $\boldsymbol{W} \in \mathbb{R}^{v \times d}$

The embedding $\boldsymbol{h}$ specifies the combination of rows of $\boldsymbol{W}$ to create the logits $\boldsymbol{W}^\top \boldsymbol{h}$.



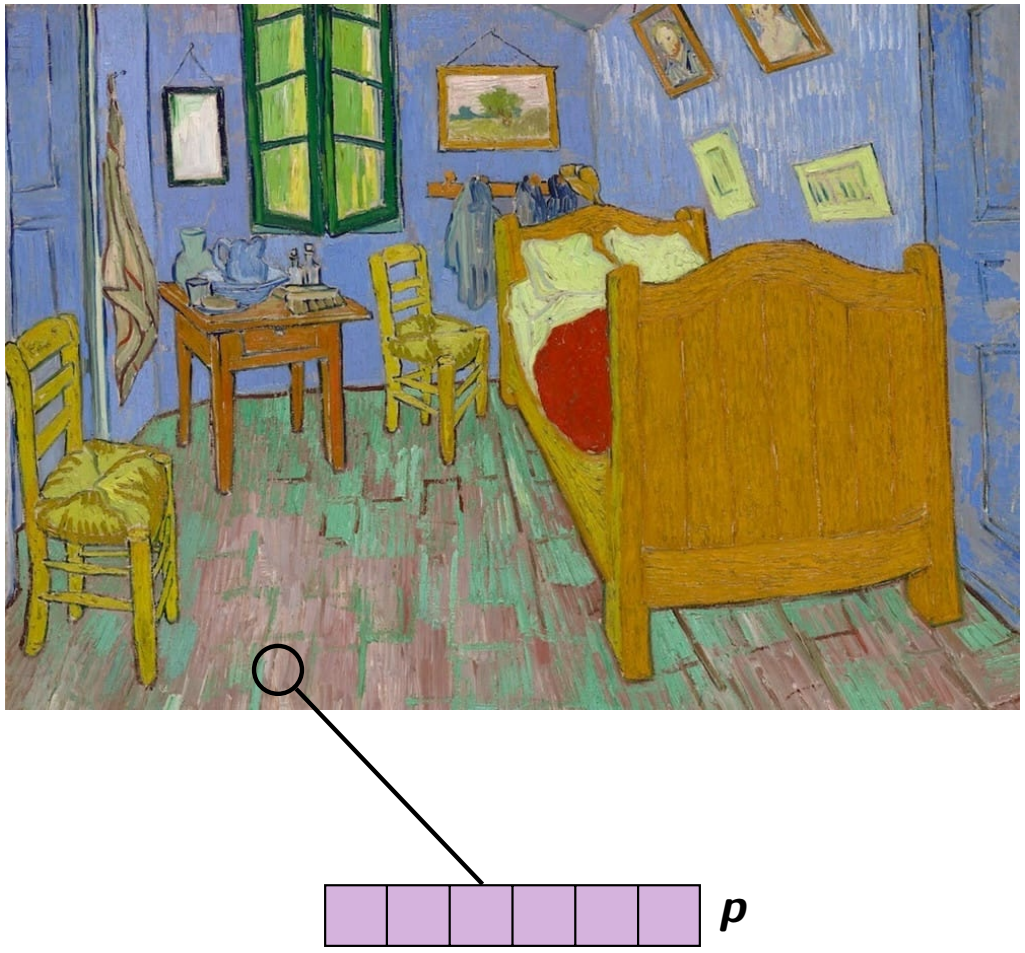
LM logits define a distribution over vocab items

Martinez 0.2
Swift 0.1
had 0.1
Taylor 0.4
Sw 0.1
go 0.1

$$\boldsymbol{p} = \text{softmax}(\boldsymbol{W}^\top \boldsymbol{h}) \in \mathbb{R}^v$$

Models cannot output arbitrary distributions

Van Gogh's palatte could not reproduce all the colors in the room: LMs cannot output arbitrary distributions over tokens.



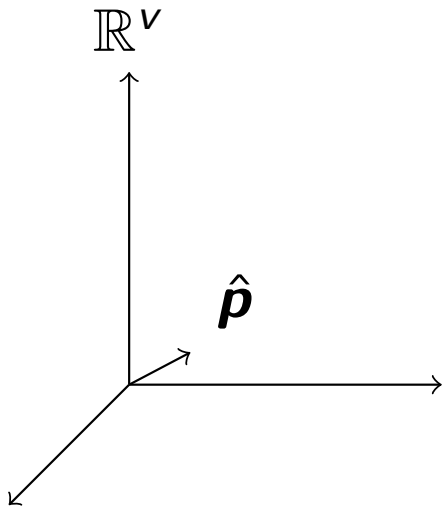
This is known as the "softmax bottleneck".

- The designers reverse engineered (kind of) the “true” colors.
- We want to reverse engineer the “true” distribution.

What are next-token distributions ($\boldsymbol{p}$)?

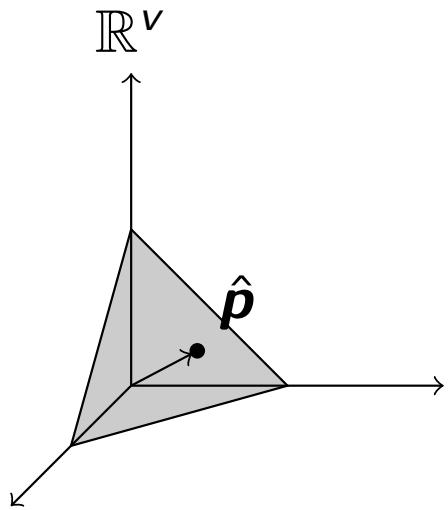
What are next-token distributions ($\boldsymbol{p}$)?

- Distributions over v items are vectors in $\mathbb{R}^v$.



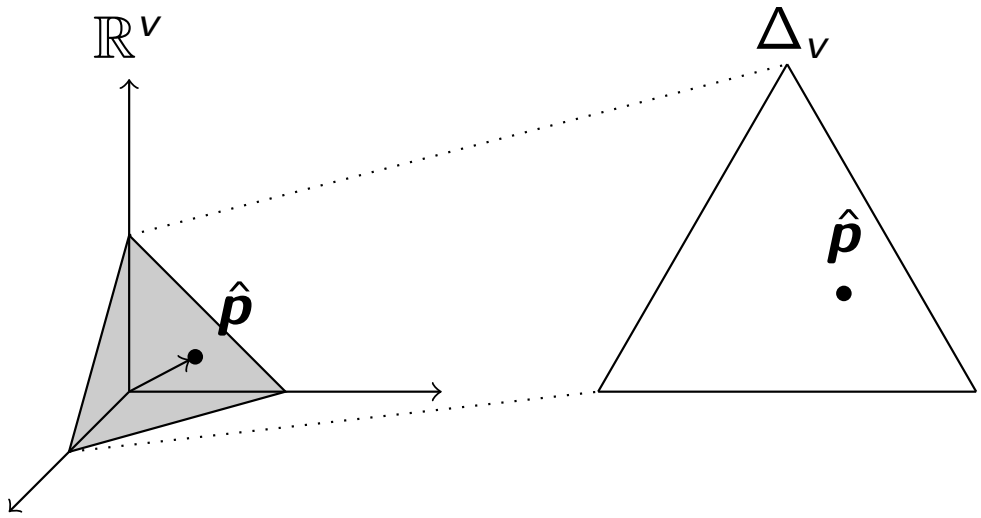
What are next-token distributions ($\boldsymbol{p}$)?

- Distributions over v items are vectors in $\mathbb{R}^v$.
- All distributions must sum to one.



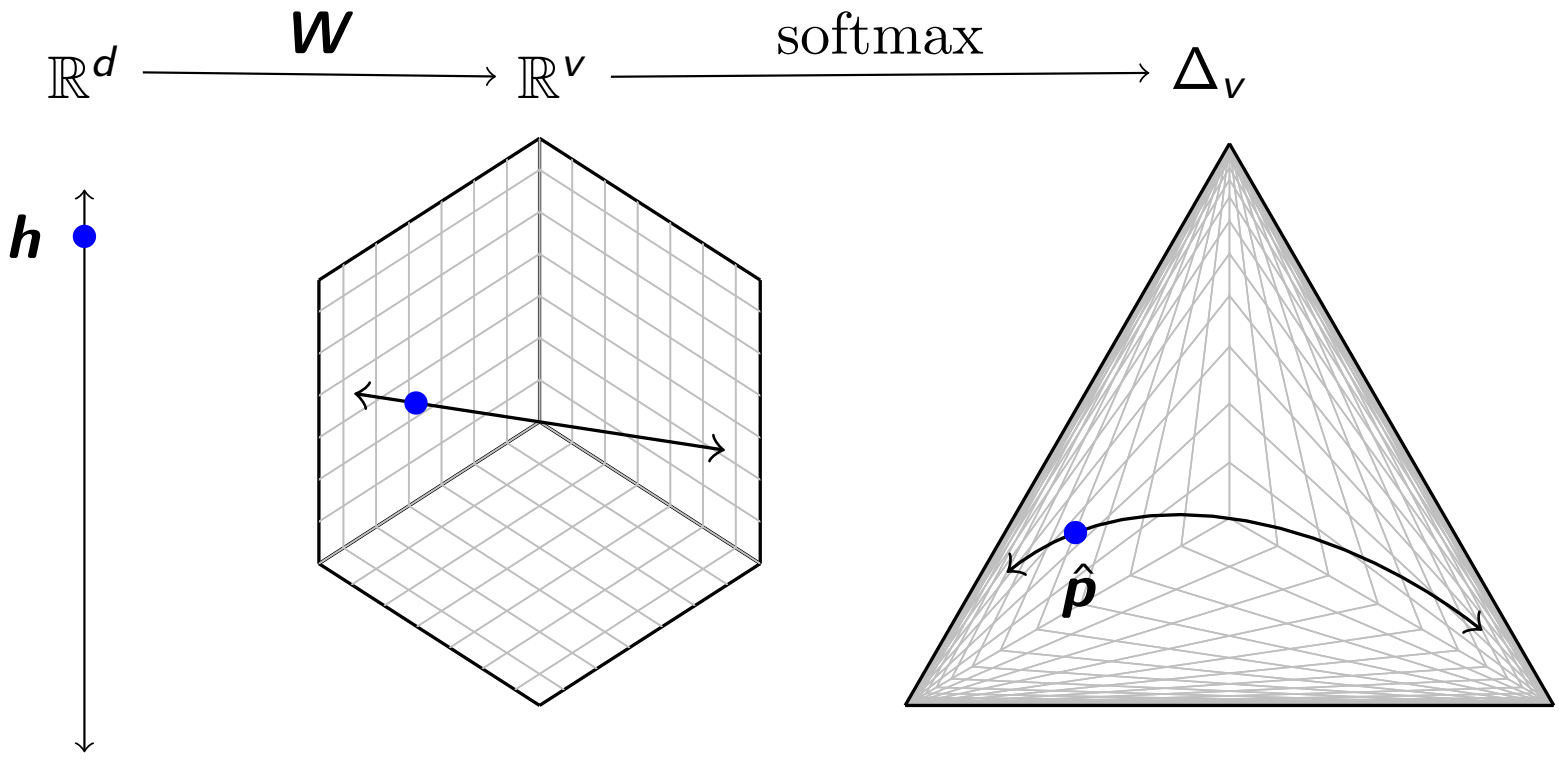
What are next-token distributions ($\boldsymbol{p}$)?

- Distributions over v items are vectors in $\mathbb{R}^v$.
- All distributions must sum to one.
- Distributions over v items are vectors in the v-simplex, or Δ_v.



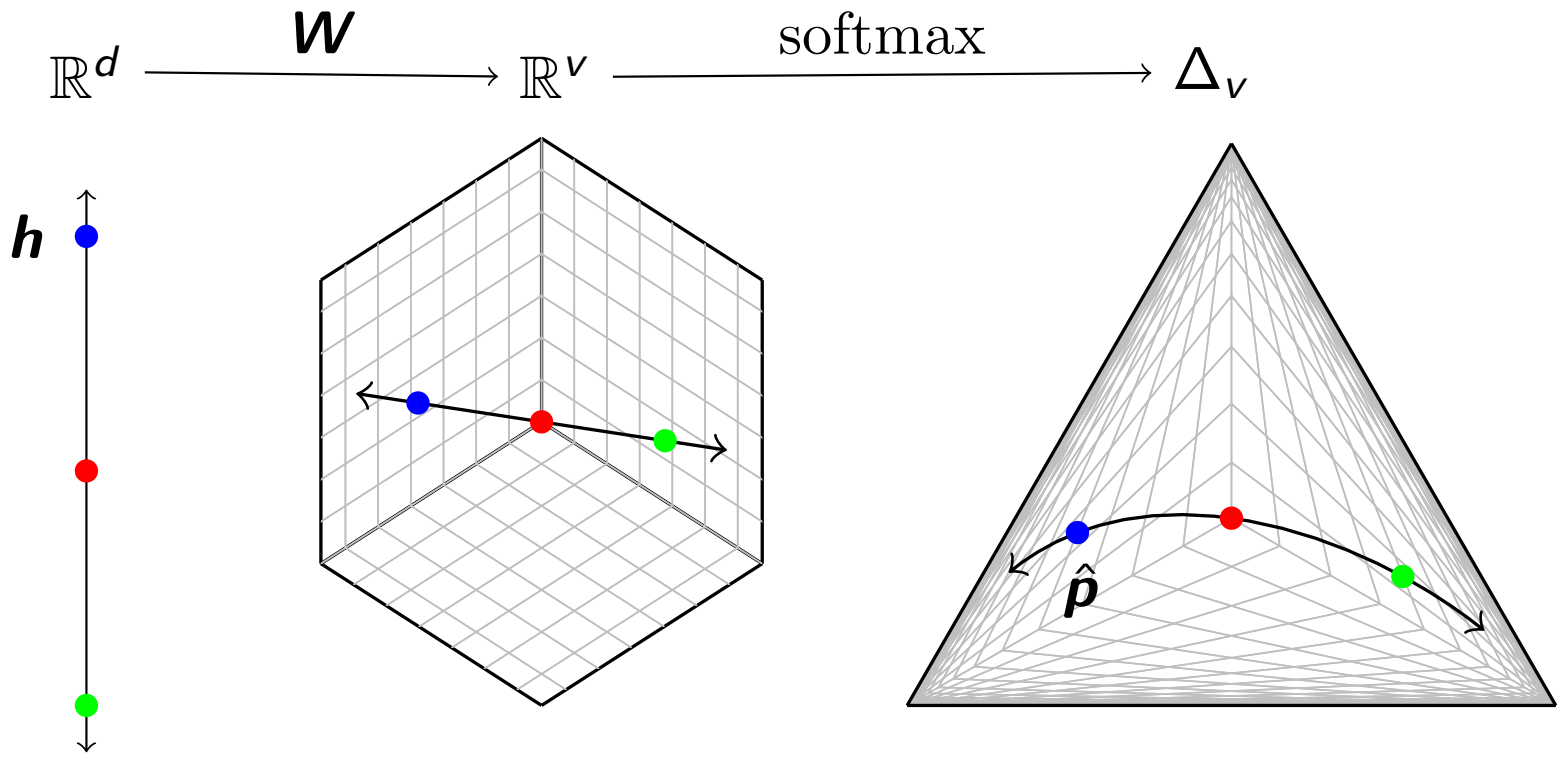
The softmax bottleneck restricts model outputs

Example: $d = 1$, $v = 3$.



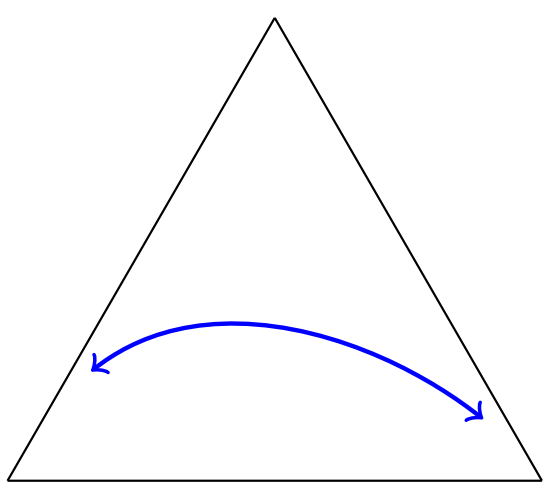
The softmax bottleneck restricts model outputs

Example: $d = 1$, $v = 3$.



Any choice of embedding $\boldsymbol{h}$ will result in a distribution $\boldsymbol{p}$ within a low-dimensional subspace (curved line on right).

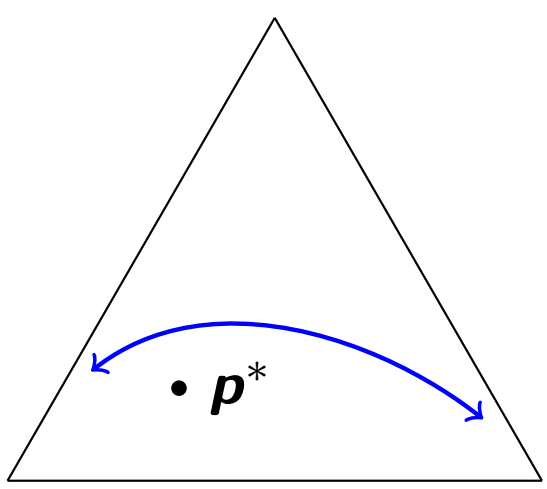
Directly addressing the softmax bottleneck



$\longleftrightarrow$ Possible model outputs

- Model outputs are constrained.

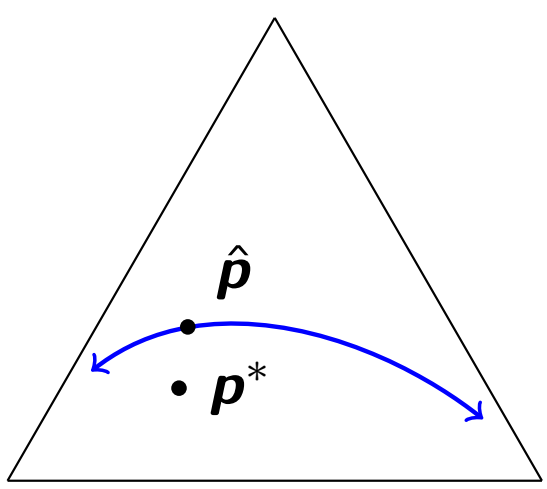
Directly addressing the softmax bottleneck



$\longleftrightarrow$ Possible model outputs

- Model outputs are constrained.
- Model cannout output true distribution $\boldsymbol{p}^*$.

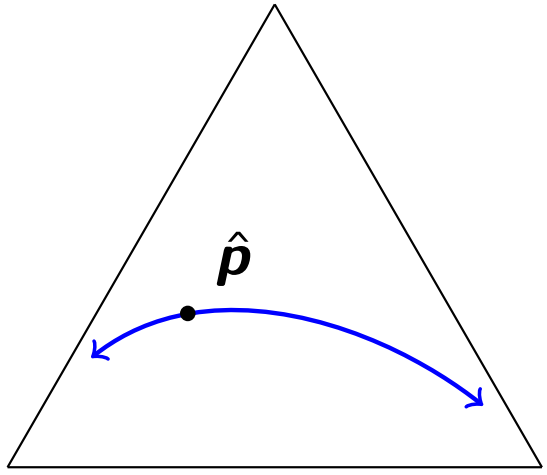
Directly addressing the softmax bottleneck



$\longleftrightarrow$ Possible model outputs

- Model outputs are constrained.
- Model cannout output true distribution $\boldsymbol{p}^*$.
- The model outputs $\hat{\boldsymbol{p}}$ that minimizes cross-entropy with $\boldsymbol{p}^*$.

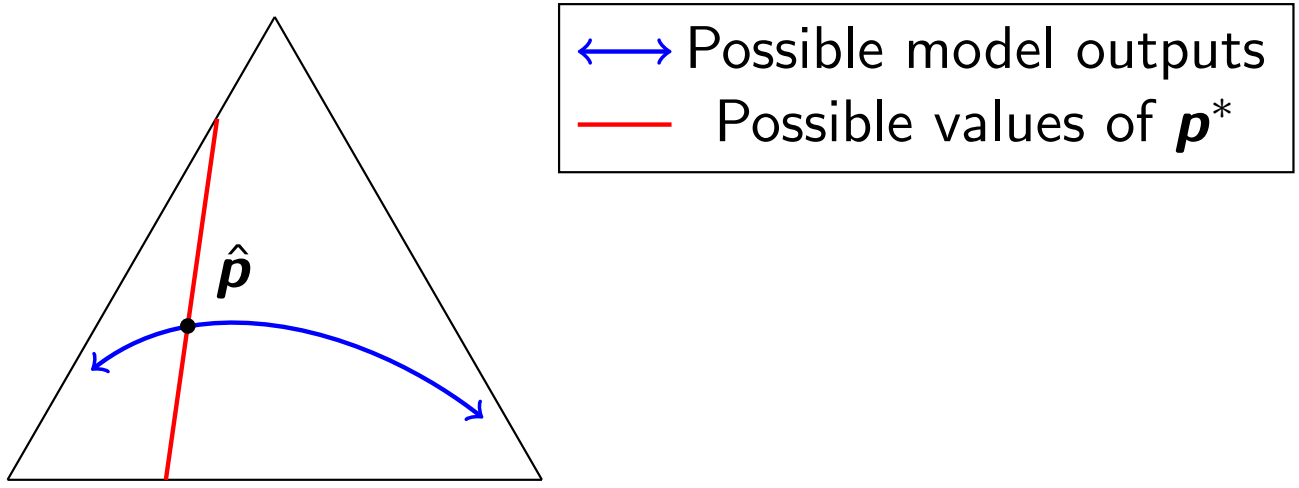
Directly addressing the softmax bottleneck



$\longleftrightarrow$ Possible model outputs

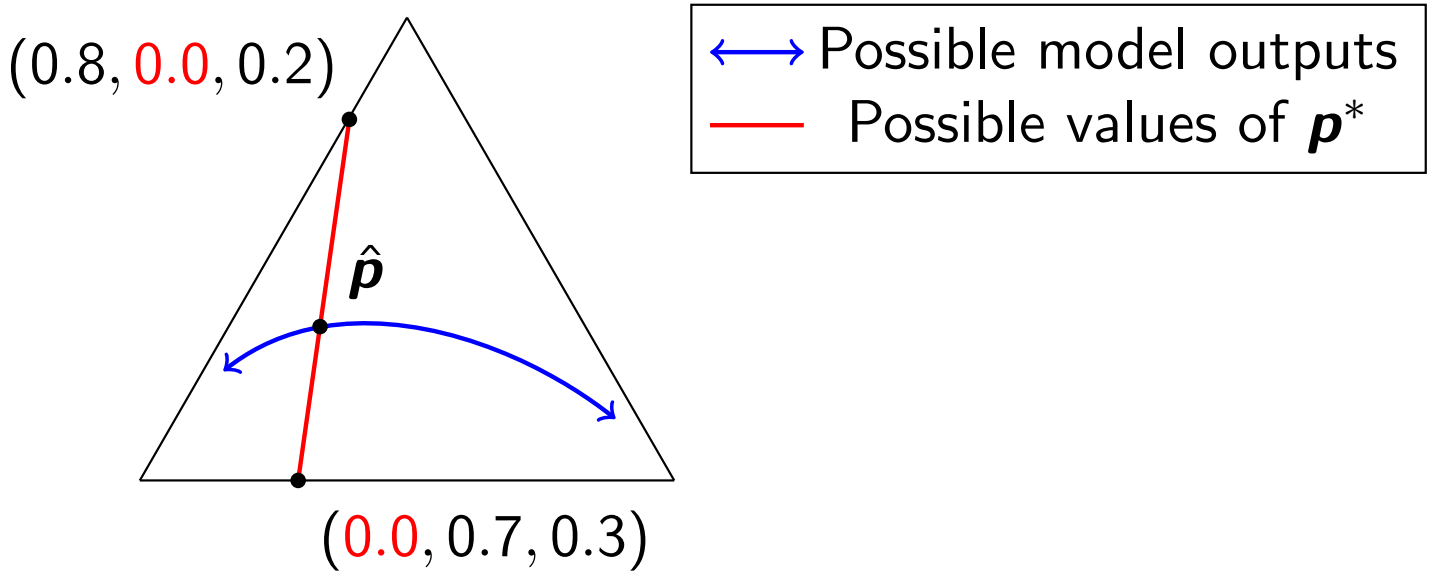
- Reverse: given $\hat{\boldsymbol{p}}$, what is $\boldsymbol{p}^*$?

Directly addressing the softmax bottleneck



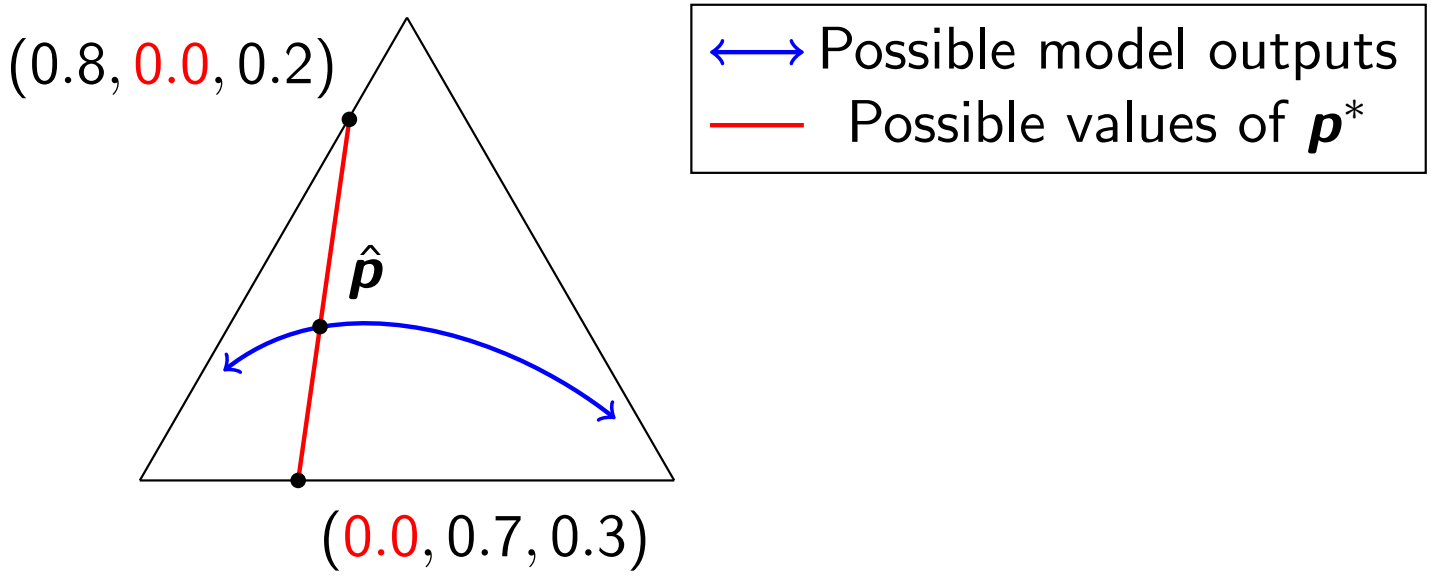
- Reverse: given $\hat{\boldsymbol{p}}$, what is $\boldsymbol{p}^*$?
- $\boldsymbol{p}^*$ could be any distribution $\hat{\boldsymbol{p}}$ minimizes cross-entropy with!

Directly addressing the softmax bottleneck



- Reverse: given $\hat{\boldsymbol{p}}$, what is $\boldsymbol{p}^*$?
- $\boldsymbol{p}^*$ could be any distribution $\hat{\boldsymbol{p}}$ minimizes cross-entropy with!
- Don't sample tokens that could have 0 true probability.

Directly addressing the softmax bottleneck

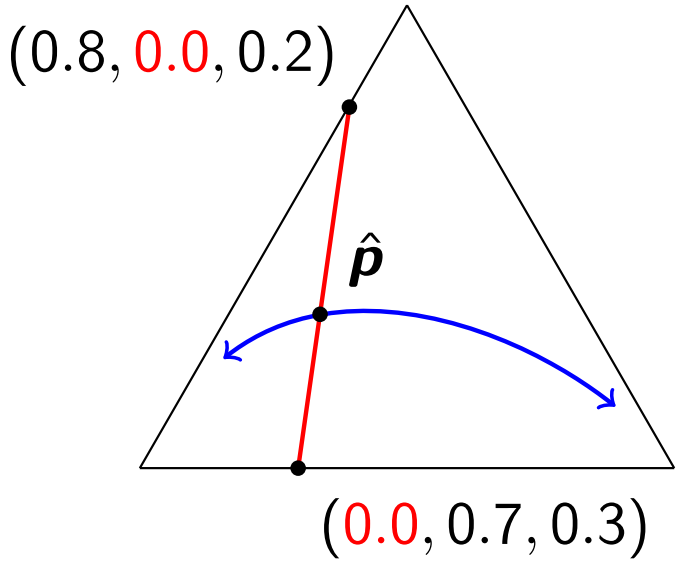


- Reverse: given $\hat{\boldsymbol{p}}$, what is $\boldsymbol{p}^*$?
- $\boldsymbol{p}^*$ could be any distribution $\hat{\boldsymbol{p}}$ minimizes cross-entropy with!
- Don't sample tokens that could have 0 true probability.
- The result: a more specific sampling rule!

Only sample tokens i if there is no solution $\boldsymbol{p} \in \Delta_v$ to

$$p_i = 0, \quad \boldsymbol{W}^T \boldsymbol{p} = \boldsymbol{W}^T \hat{\boldsymbol{p}}.$$

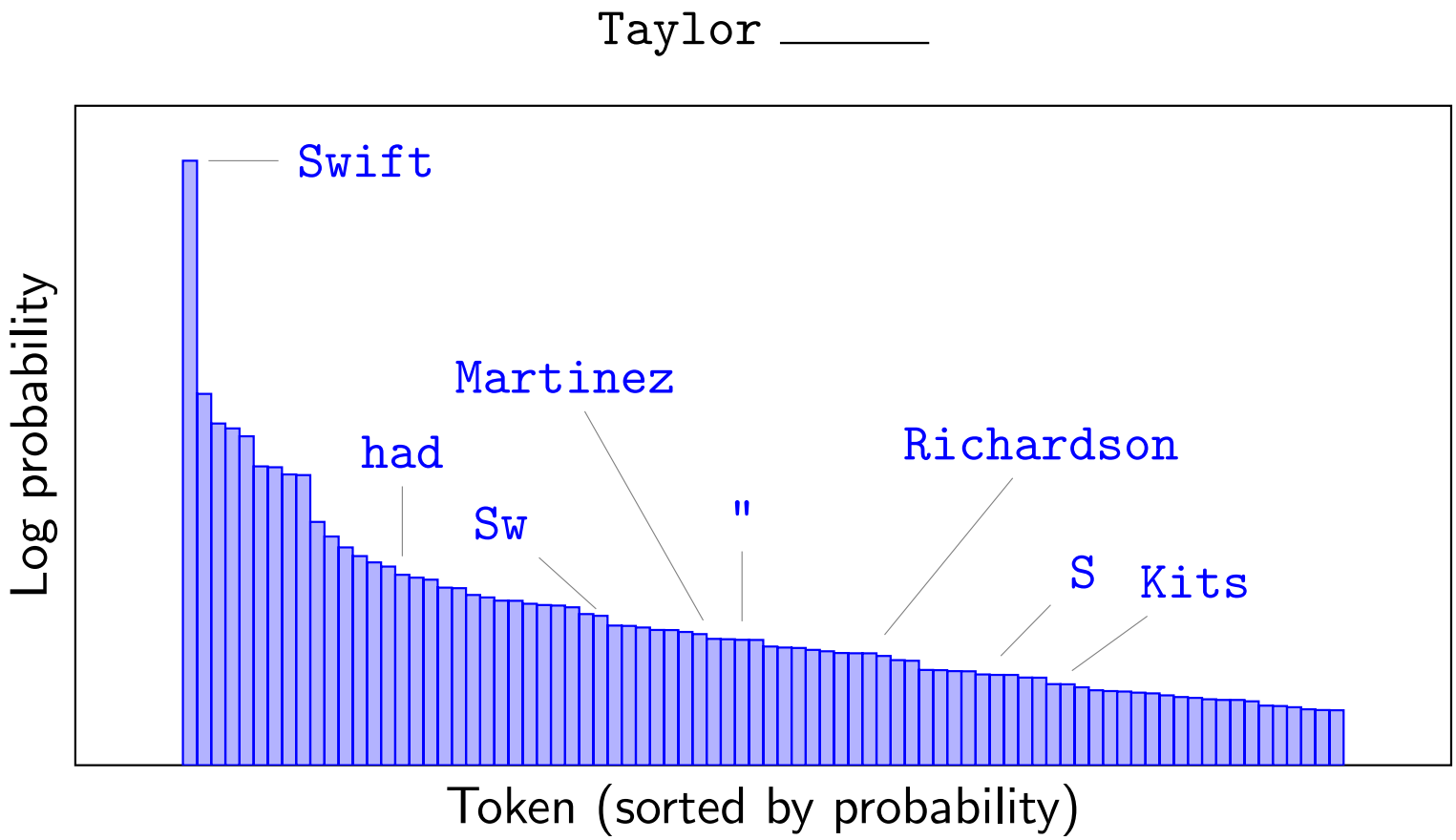
Directly addressing the softmax bottleneck



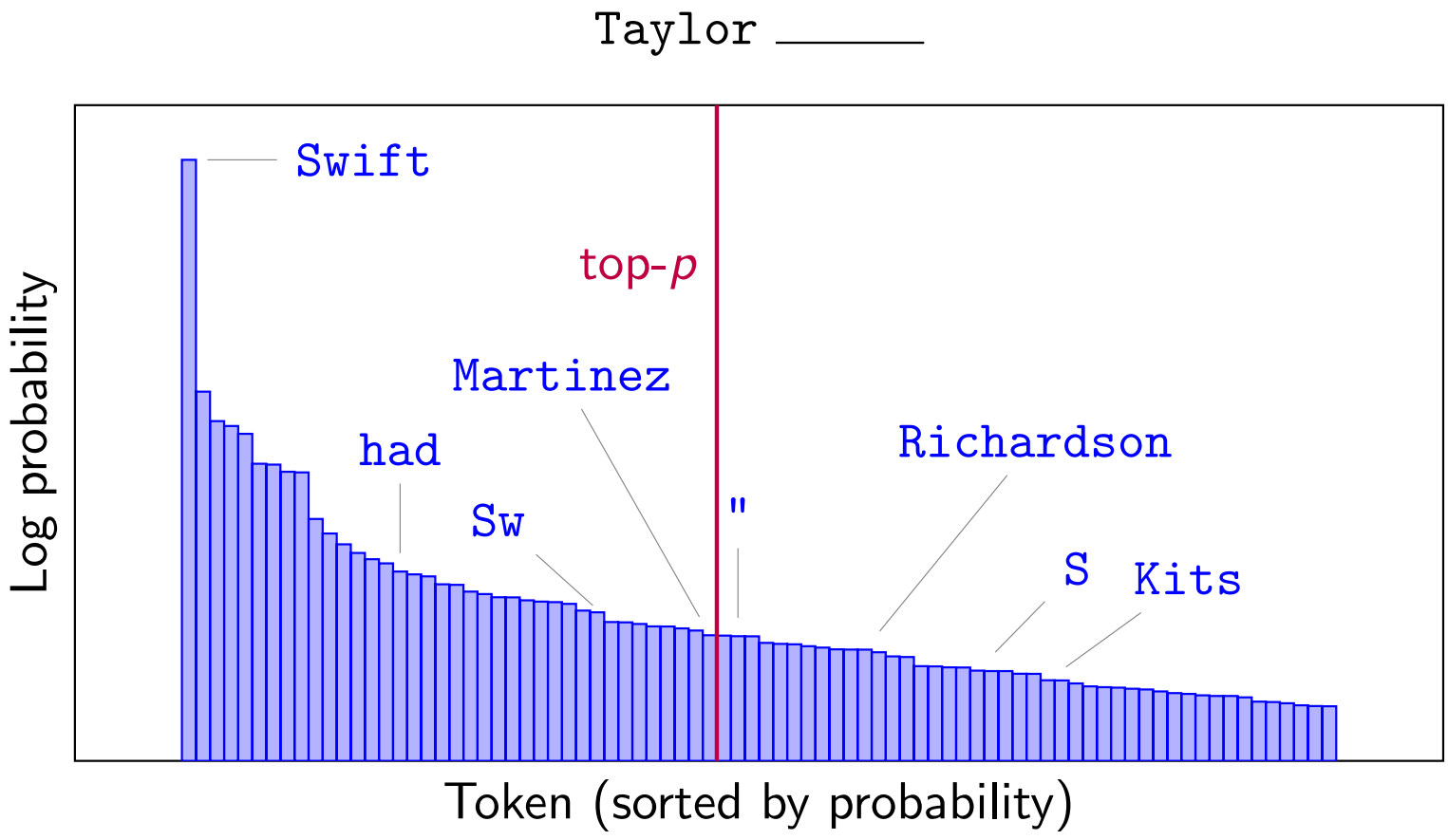
$\longleftrightarrow$ Possible model outputs

— Possible values of $\boldsymbol{p}^*$

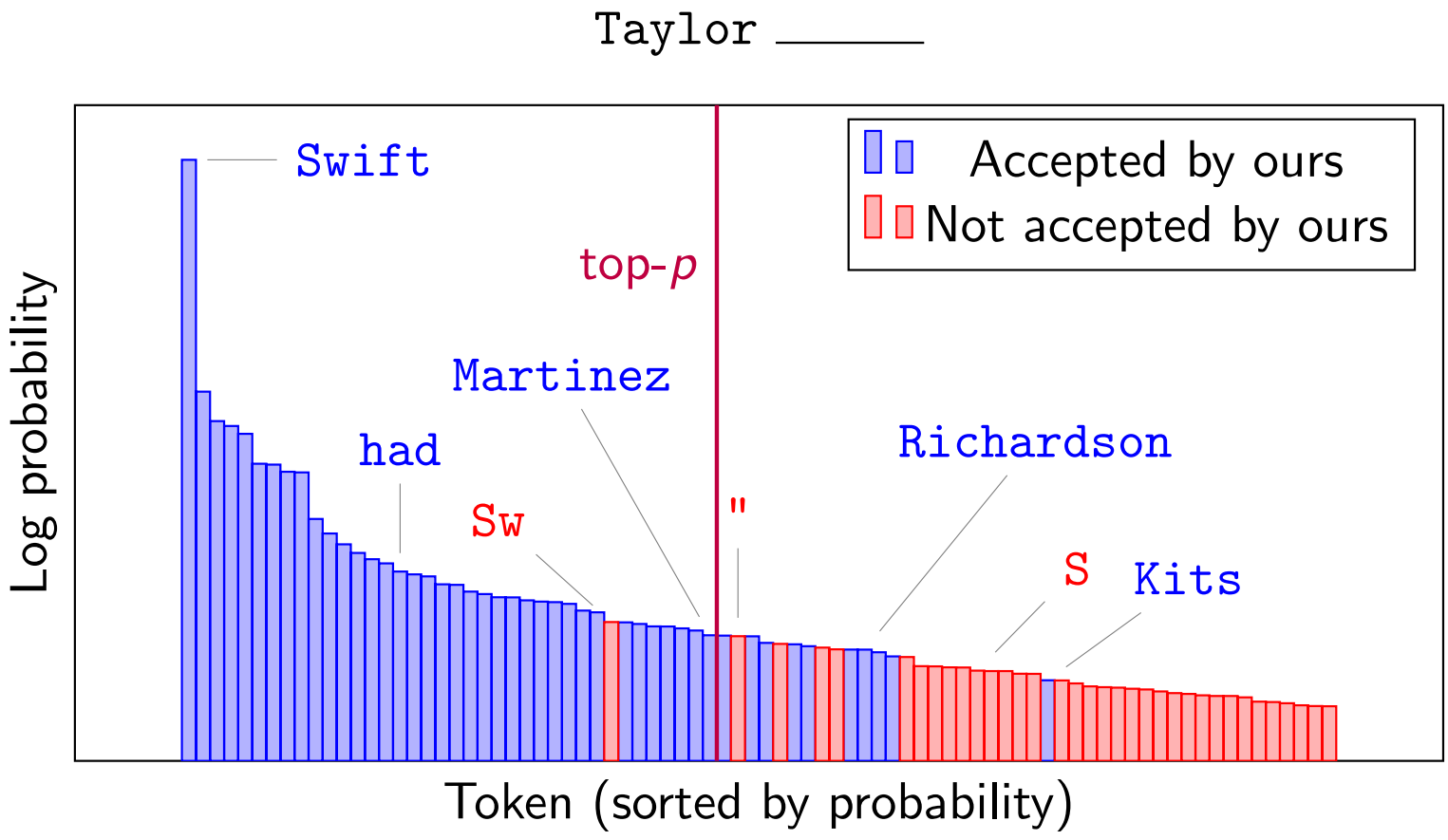
Anecdotal evidence



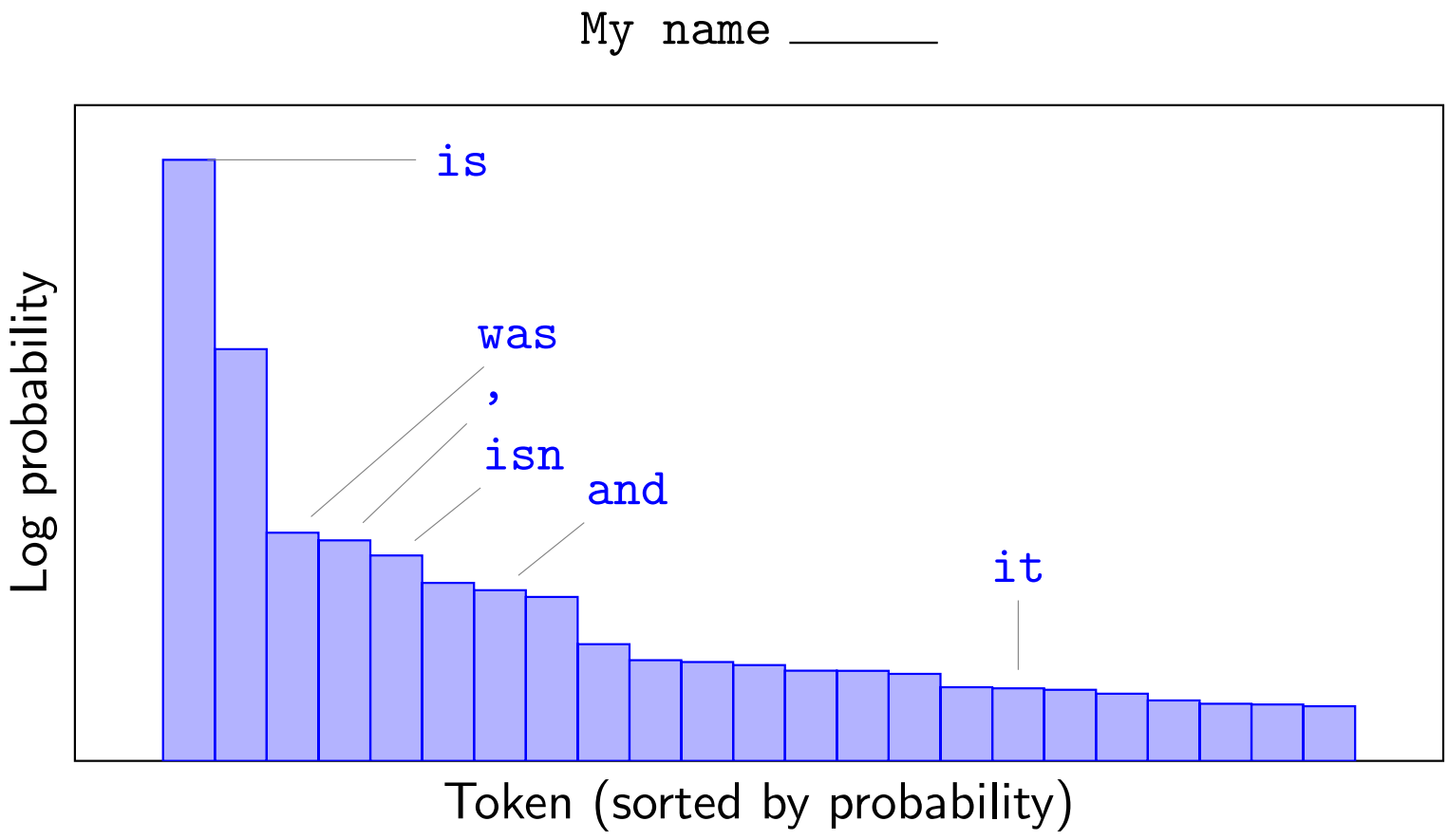
Anecdotal evidence



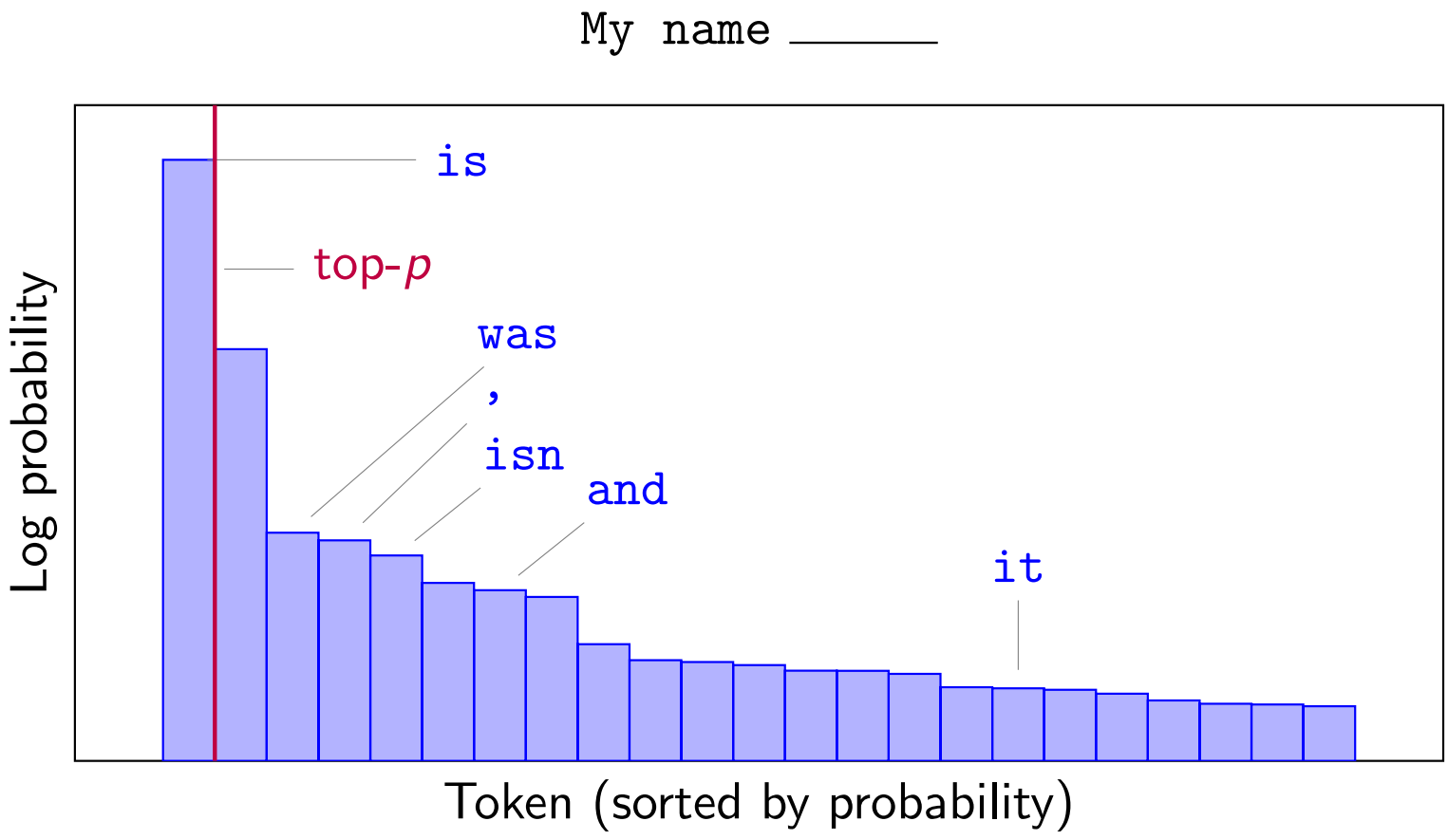
Anecdotal evidence



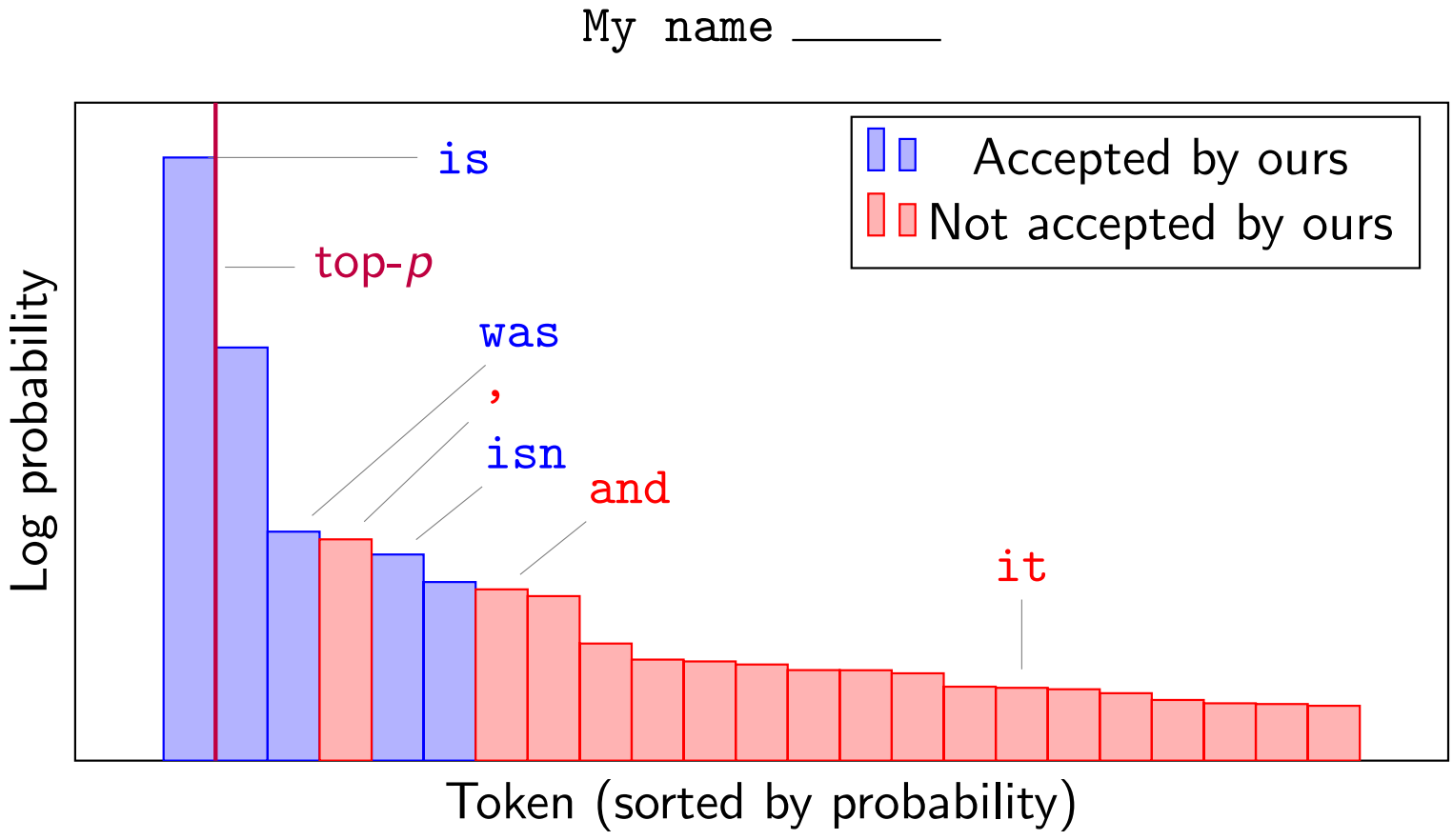
Anecdotal evidence



Anecdotal evidence

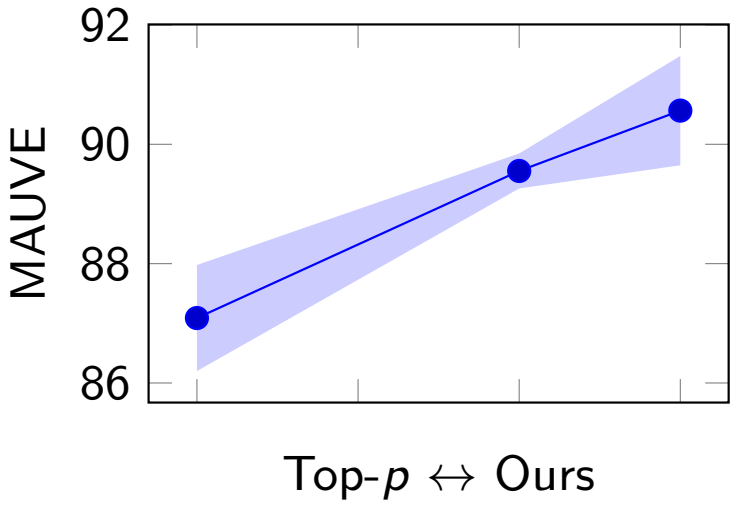


Anecdotal evidence



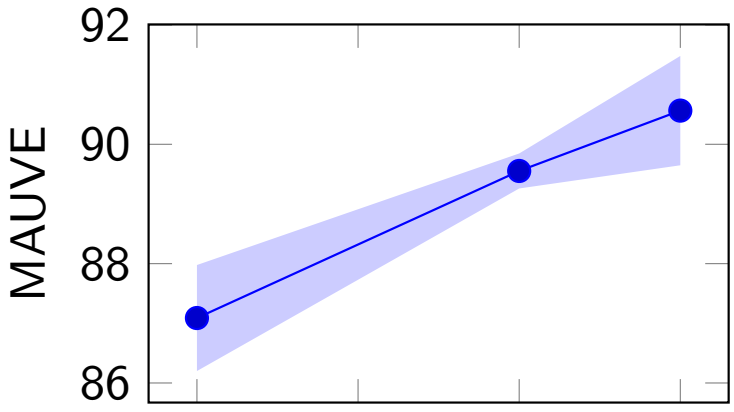
Basis-aware threshold (BAT) sampling

- More BAT-like, higher MAUVE (similarity to human text).

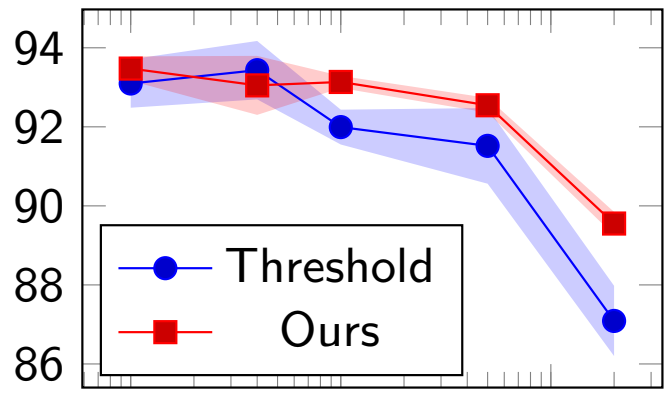


Basis-aware threshold (BAT) sampling

- More BAT-like, higher MAUVE (similarity to human text).
- Low-entropy (closer to greedy) BAT $>$ threshold sampling.



Top-p $\leftrightarrow$ Ours



$\rightarrow$ closer to greedy

Basis-aware threshold sampling

The smoking gun

Our method outperforms greedy-like decoding across model sizes

Size Method	Small	Medium	Large	XL
Threshold	$85.0_{1.4}$	$90.4_{0.1}$	$86.0_{0.5}$	$87.1_{1.2}$
Ours	$\mathbf{87.8_{1.0}}$	$\mathbf{92.2_{0.6}}$	$\mathbf{88.4_{0.5}}$	$\mathbf{89.6_{0.4}}$

Recap

- Threshold sampling (e.g., top-p) is a coarse heuristic to avoid model errors.
- Understanding a source of model errors (the softmax bottleneck) allows us to better recover the true distribution.
- We can leverage this method to sample more precisely.
- BAT outperforms threshold sampling in greedy-like settings.
- Lots of room for improvement :)

Thank you!