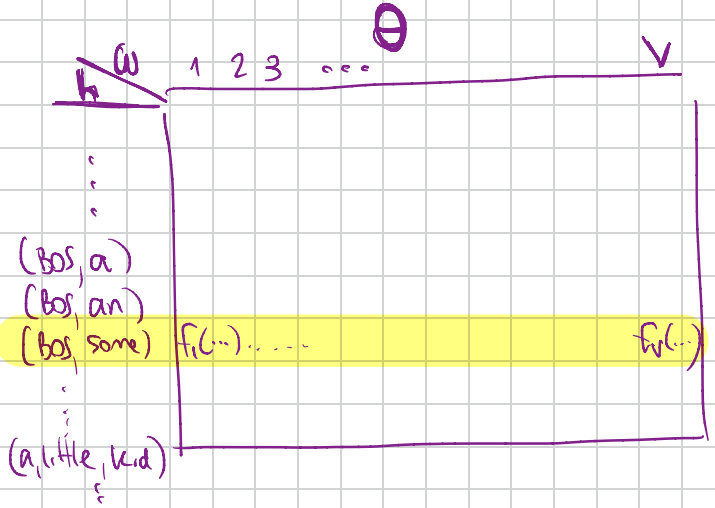
① Categorical Distribution

$$W \mid H=h \sim \text{Categorical}\Big(\underbrace{f(h;\theta)}_{(f_1(h;\theta), f_2(h;\theta), \ldots, f_V(h;\theta))}\Big)$$

$W \rightarrow \{1, \ldots, V\}$

TABULAR REPRESENTATION



$f \rightarrow$ table lookup operation

$$f\Big(\overbrace{(\text{BOS}, \text{some})}^{h}; \theta\Big) = \Big(\theta_1^{(h)}, \theta_2^{(h)}, \theta_3^{(h)}, \ldots, \theta_V^{(h)}\Big)$$

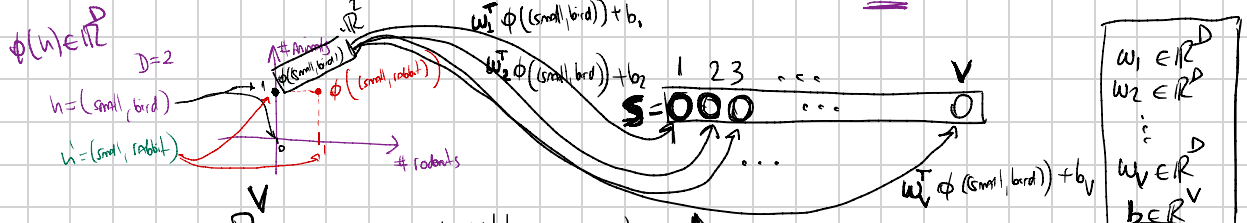
② Logistic Representation Categorical CPDs

$$W \mid H = h \sim \text{Categorical}\Big(\underbrace{\mathbf{f}(h;\theta)}_{(f_1(h;\theta), f_2(h;\theta), \ldots, f_V(h;\theta))}\Big)$$

$W \mid H=h \rightarrow \{1, \ldots, V\}$

Instead of storing the cond. prob. masses for every (h, w) condition-outcome pair, we predict the cond. prob. masses using A log-linear model

- Feature function: map an instance of h to a fixed dimension Real coordinate space



$s \in \mathbb{R}^V \rightarrow$ probability simplex Δ_{V-1}

$$[\text{softmax}(s)]_i = \frac{\exp(s_i)}{\sum_{j=1}^{V} \exp(s_j)}$$

1. The model size does not depend on how many different instances of b exist

$$w_1 \in \mathbb{R}^D \quad \cdots \quad w_V \in \mathbb{R}^D \Rightarrow V \times D$$

$$b_1 \in \mathbb{R} \quad \cdots \quad b_V \in \mathbb{R} \Rightarrow V$$

$$V(D+1)$$

2. Histories / Conditioning context are no longer treated as unrelated to one another

	#cool	#dog	#cat	#dry	#(positive)	#(animals)	#(positive animal)
cool dog	1	1	0	0	1	1	1
cool cat	1	0	1	0	1	1	1
cool dry	1	0	0	1	1	0	0

③ ESTIMATION

$$\mathcal{D} = \{ (h_n, w_n)_{n=1}^{N} \}$$

$$W|H=h \sim \text{Categorical}\left(f(h;\theta) \right)$$

$$s = W\phi(h) + b$$

$$f(h;\theta) = \text{softmax}(s)$$

$$\phi(h) \in \mathbb{R}^{D}$$

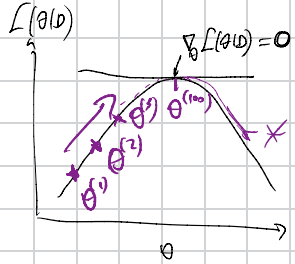
$$W \in \mathbb{R}^{V \times D}$$

$$b \in \mathbb{R}^{V}$$

$$\mathcal{L}(\theta|\mathcal{D}) = \sum_{n=1}^{N} \log p(w_n|h_n;\theta)$$

$$= \sum_{n=1}^{N} \log \left[f(h_n;\theta) \right]_{w_n}$$

$$= \sum_{n=1}^{N} \log \left[\text{softmax}\left(W\phi(h_n) + b \right) \right]_{w_n}$$



$$\theta^{(t+1)} = \theta^{(t)} + \gamma_{t+1} \nabla_{\theta^{(t)}} \mathcal{L}(\theta^{(t)}|\mathcal{D})$$