# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

# Exponential Softmax Backpropagation:

# The Model Marginals

# Back-Propagation Through Exponential Softmax

$$
\begin{aligned}
s^N[N, Y] &= f^N_\Phi(x) \\
s^E[E, Y, Y] &= f^E_\Phi(x)
\end{aligned}
$$

$$
s(\hat{\mathcal{Y}}) = \sum_n s^N[n, \hat{\mathcal{Y}}[n]] + \sum_{\langle n, m \rangle \in \text{Edges}} s^E[\langle n, m \rangle, \hat{\mathcal{Y}}[n], \hat{\mathcal{Y}}[m]]
$$

$$
P_s(\hat{\mathcal{Y}}) = \underset{\hat{\mathcal{Y}}}{\text{softmax}} \; s(\hat{\mathcal{Y}}) \quad \text{all possible } \hat{\mathcal{Y}}
$$

$$
\mathcal{L} = -\ln P_s(\mathcal{Y}) \quad \text{gold label } \mathcal{Y}
$$

We want the gradient tensors $s^N.\text{grad}[N, Y]$ and $s^E.\text{grad}[E, Y, Y]$.

# Model Marginals Theorem

Theorem:

$$s^N.\text{grad}[n, y] = P_{\hat{\mathcal{Y}} \sim P_s}(\ \hat{\mathcal{Y}}[n] = y\ ) - \mathbf{1}[\ \mathcal{Y}[n] = y\ ]$$

$$s^E.\text{grad}[\langle n, m \rangle, y, y'] = P_{\hat{\mathcal{Y}} \sim P_s}(\ \hat{\mathcal{Y}}[n] = y \wedge \hat{\mathcal{Y}}[m] = y'\ ) - \mathbf{1}[\ \mathcal{Y}[n] = y \wedge \mathcal{Y}[m] = y'\ ]$$

We need to compute (or approximate) the model marginals.

# Proof of Model Marginals Theorem

We consider the case of node marginals, the case of edge marginals is similar.

$$
\begin{aligned}
s^N.\mathrm{grad}[n,y] &= \partial \mathcal{L}(\Phi, x, \mathcal{Y}) \;/\; \partial s^N[n,y] \\
&= \partial \left( -\ln \frac{1}{Z} \exp(s(\mathcal{Y})) \right) \;/\; \partial s^N[n,y] \\
&= \partial(\ln Z - s(\mathcal{Y})) \;/\; \partial s^N[n,y] \\
&= \left( \frac{1}{Z} \sum_{\hat{\mathcal{Y}}} e^{s(\hat{\mathcal{Y}})} \left( \partial s(\hat{\mathcal{Y}}) / \partial s^N[n,y] \right) \right) - \left( \partial s(\mathcal{Y}) / \partial s^N[n,y] \right)
\end{aligned}
$$

# Proof of Model Marginals Theorem

$$
\begin{aligned}
s^N.\text{grad}[n, y] &= \left( \frac{1}{Z} \sum_{\hat{\mathcal{Y}}} e^{s(\hat{\mathcal{Y}})} \left( \partial s(\hat{\mathcal{Y}}) / \partial s^N[n, y] \right) \right) - \left( \partial s(\mathcal{Y}) / \partial s^N[n, y] \right) \\
&= \left( \sum_{\hat{\mathcal{Y}}} P_s(\hat{\mathcal{Y}}) \left( \partial s(\hat{\mathcal{Y}}) / \partial s^N[n, y] \right) \right) - \left( \partial s(\mathcal{Y}) / \partial s^N[n, y] \right) \\
s(\hat{\mathcal{Y}}) &= \sum_n s^N[n, \hat{\mathcal{Y}}[n]] + \sum_{\langle n, m \rangle \in \text{Edges}} s^E[\langle n, m \rangle, \hat{\mathcal{Y}}[n], \hat{\mathcal{Y}}[m]] \\
\frac{\partial s(\hat{\mathcal{Y}})}{\partial s^N[n, y]} &= \mathbf{1}[\hat{\mathcal{Y}}[n] = y]
\end{aligned}
$$

# Proof of Model Marginals Theorem

$$
\begin{aligned}
s^N.\text{grad}[n, y] \;=\; & \left( \frac{1}{Z} \sum_{\hat{y}} e^{s(\hat{\mathcal{Y}})} \left( \partial s(\hat{\mathcal{Y}}) / \partial s^N[n, y] \right) \right) - \left( \partial s(\mathcal{Y}) / \partial s^N[n, y] \right) \\
& \left( \sum_{\hat{y}} P_s(\hat{\mathcal{Y}}) \left( \partial s(\hat{\mathcal{Y}}) / \partial s^N[n, y] \right) \right) - \left( \partial s(\mathcal{Y}) / \partial s^N[n, y] \right) \\
=\; & E_{\hat{\mathcal{Y}} \sim P_s} \mathbf{1}[\hat{\mathcal{Y}}[n] = y] - \mathbf{1}[\mathcal{Y}[n] = y] \\
=\; & P_{\hat{\mathcal{Y}} \sim P_s}(\hat{\mathcal{Y}}[n] = y) - \mathbf{1}[\mathcal{Y}[n] = y]
\end{aligned}
$$

## Model Marginals Theorem

Theorem:

$$s^N.\text{grad}[n, y] \;=\; P_{\hat{y}\sim P_s}(\;\hat{\mathcal{Y}}[n] = y\;) \\ -\mathbf{1}[\;\mathcal{Y}[n] = y\;]$$

$$s^E.\text{grad}[\langle n, m\rangle, y, y'] \;=\; P_{\hat{y}\sim P_s}(\;\hat{\mathcal{Y}}[n] = y \;\wedge\; \hat{\mathcal{Y}}[m] = y'\;) \\ -\mathbf{1}[\;\mathcal{Y}[n] = y \;\wedge\; \mathcal{Y}[m] = y'\;]$$

# END