Hidden Markov Models

Justin T. Chiu

February 7, 2020

Abstract

TODO

1 Problem Setup

We apply hidden markov models (HMMs) to language modeling, where we would like to model sentences $\mathbf{x}_{1:T}$. The generative process of an HMM is as follows:

1. Choose an initial state $z_0 \sim \text{Cat}(\nu), \nu \in \mathbb{R}^K$
2. For each time $t \in \{1, \ldots, T\}$ choose a state $z_t \mid z_{t-1} = i \sim \text{Cat}(\theta_{i\cdot}), \theta \in \mathbb{R}^{K \times K}$
3. For each time $t \in \{0, \ldots, T\}$ choose a word $x_t \mid z_t = j \sim \text{Cat}(\phi_{j\cdot}), \phi \in \mathbb{R}^{V \times K}$.

This gives the following joint distribution:

$$\log p_\theta(\mathbf{x}_{0:T}, \mathbf{z}_{0:T}) = \log p_\theta(x_0, z_0) + \sum_{t=1}^{T} \log p_\theta(x_t, z_t \mid z_{t-1})$$

2 Parameter estimation

We maximize the evidence of the observed sentences $\log p(\mathbf{x}_{0:T}) = \log \sum_{\mathbf{z}_{0:T}} p(\mathbf{x}_{1:T}, \mathbf{z}_{0:T})$.

2.1 Gradient of evidence

Let $\psi_0(z_0, z_1) = \log p(\mathbf{x}_{0:1}, \mathbf{z}_{0:1})$ and $\psi_t(z_t, z_{t+1}) = \log p(x_{t+1}, z_{t+1} \mid z_t)$ for $t \in \{1, \ldots, T-1\}$. Additionally, let $\oplus$ and $\otimes$ be addition and multiplication in the log semiring. After conditioning on the observed $\mathbf{x}_{0:T}$, we can express the evidence as the following:

$$A_x = \log p(\mathbf{x}_{0:T}) = \bigoplus_{\mathbf{z}_{0:T}} \bigotimes_{t=0}^{T-1} \psi_t(z_t, z_{t+1})$$

where $A_{\mathbf{x}}$ is the clamped log partition function.

We show that the gradient of the log partition function with respect to the $\psi_t(z_t, z_{t+1})$ is the first moment (a well-known result of the log cumulant generating function and exponential family distributions). Given the value of the derivative $\frac{\partial A_{\mathbf{x}}}{\partial \psi_t(z_t,z_{t+1})}$, we can then apply the chain rule to compute the total derivative with respect to downstream parameters. For example, the gradient with respect to the transition matrix of the HMM is given by

$$\frac{\partial A_{\mathbf{x}}}{\partial \theta_{ij}} = \sum_t \frac{\partial A_{\mathbf{x}}}{\partial \psi_t(i,j)} \frac{\partial \psi_t(i,j)}{\partial \theta_{ij}}$$

Recall that the gradient of logaddexp is

$$\frac{\partial}{\partial x} x \oplus y = \frac{\partial}{\partial x} \log e^x + e^y = \frac{e^x}{e^x + e^y}.$$

The gradient of the clamped log partition function $A_{\mathbf{x}}$ is given by

$$\begin{aligned}\frac{\partial A_{\mathbf{x}}}{\partial \psi_t(i,j)} &= \frac{\partial}{\partial \psi_t(i,j)} \bigoplus_{\mathbf{z}_{0:T}} \bigotimes_t \psi_t(z_t, z_{t+1})\\ &= \frac{\partial}{\partial \psi_t(i,j)} \bigoplus_{\mathbf{z}_{0:T}} \bigotimes_t \psi_t(z_t, z_{t+1})\end{aligned}$$

2.2

2.2.1 Very high training loss

Surrogate loss is a loose bound, but that is ok. We proved gradient estimator is correct.