TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Monte-Carlo Markov Chain (MCMC) Sampling

Sampling From the Model

For backpropagation through an exponential softmax we will estimate the model marginals by sampling from the exponential model distribution.

Theorem:

$$s_n.\text{grad}[i, c] = P_{\hat{y} \sim P_s}(\ \hat{y}[i] = c\ ) \\ -\mathbf{1}[\ y[i] = c\ ]$$

$$s_e.\text{grad}[\langle i, j \rangle, c, c'] = P_{\hat{y} \sim P_s}(\ \hat{y}[i] = c \wedge \hat{y}[j] = c'\ ) \\ -\mathbf{1}[\ y[i] = c \wedge y[j] = c'\ ]$$

MCMC Sampling

The model marginals, such as the node marginals $P_s(\hat{y}[i] = c)$, can be estimated by sampling $\hat{y}$ from $P_s(\hat{y})$.

There are various ways to design a Markov process whose states are node labelings $\hat{y}$ and whose stationary distribution is P_s.

Given such a process we can sample $\hat{y}$ from P_s by running the process past its mixing time.

We will consider Metropolis MCMC and the Gibbs MCMC. But there are more (like Hamiltonian MCMC).

Metroplis MCMC

We assume a neighor relation on node assignments and let $N(\hat{y})$ be the set of neighbors of assignment $\hat{y}$.

For example, $N(\hat{y})$ can be taken to be the set of assignments $\hat{y}'$ that differ form $\hat{y}$ on exactly one node.

For the correctness of Metropolis MCMC we need that all states have the same number of neighbors and that the neighbor relation is symmetric — $\hat{y}' \in N(\hat{y})$ if and only if $\hat{y} \in N(\hat{y}')$.

Metropolis MCMC

Pick an initial state $\hat{y}_0$ and for $t \geq 0$ do

1. Pick a neighbor $\hat{y}' \in N(\hat{y}_t)$ uniformly at random.

2. If $P_s(\hat{y}') > P_s(\hat{y}_t)$ then $\hat{y}_{t+1} = \hat{y}'$

3. If $P_s(\hat{y}') \leq P_s(\hat{y})$ then with probability

$$e^{-\Delta s} = e^{-(s(\hat{y})-s(\hat{y}'))} = \frac{e^{s(\hat{y}')}}{e^{s(\hat{y})}} = \frac{P_s(\hat{y}')}{P_s(\hat{y})}$$

do $\hat{y}_{t+1} = \hat{y}'$ and otherwise $\hat{y}_{t+1} = \hat{y}_t$

The Metropolis Markov Chain

We need to show that P_s is a stationary distribution of this process.

We must show that if we select $\hat{y}_t$ from P_s, and then select $\hat{y}_{t+1}$ using the transition probabilities, then the distribution on $\hat{y}_{t+1}$ is also P_s.

Stationarity Condition

$$
\begin{aligned}
P'(\hat{y}) &= \sum_{\hat{y}'} P_s(\hat{y}') P_{\text{Trans}}(\hat{y} \mid \hat{y}') \\
&= P_s(\hat{y}) + \text{flow-in} - \text{flow-out} \\
&= P_s(\hat{y}) + \sum_{\hat{y}' \in N(\hat{y})} P_s(\hat{y}') P_{\text{Trans}}(\hat{y} \mid \hat{y}') - P_s(\hat{y}) P_{\text{Trans}}(\hat{y}' \mid \hat{y})
\end{aligned}
$$

Detailed Balance

Detailed balance means that for each pair of neighboring assignments $\hat{y}$, $\hat{y}'$ we have equal flows in both directions.

$$P_s(\hat{y}')P_{\text{Trans}}(\hat{y} \mid \hat{y}') = P_s(\hat{y})P_{\text{Trans}}(\hat{y}' \mid \hat{y})$$

Without loss generality assume $P_s(\hat{y}') \geq P_s(\hat{y})$.

Metropolis is defined by

$$P_{\text{Trans}}(\hat{y} \mid \hat{y}') = e^{-\Delta s}\ P_{\text{Trans}}(\hat{y}' \mid \hat{y}) = \frac{P_s(\hat{y})}{P_s(\hat{y}')}\ P_{\text{Trans}}(\hat{y}' \mid \hat{y})$$

Gibbs Sampling

The Metropolis algorithm wastes time by rejecting proposed moves.

Gibbs sampling avoids this move rejection.

In Gibbs sampling we select a node i at random and change that node by drawing a new node value conditioned on the current values of the other nodes.

We let $\hat{y}\backslash i$ be the assignment of labels given by $\hat{y}$ except that no label is assigned to node i.

We let $\hat{y}[N(i)]$ be the assignment that $\hat{y}$ gives to the nodes (pixels) that are the neighbors of node i (connected to i by an edge.)

Gibbs Sampling

Markov Blanket Property:

$$P_s(\hat{y}[i] \mid \hat{y} \backslash i) = P_s(\hat{y}[i] \mid \hat{y}[N(i)])$$

Gibbs Sampling, Repeat:

- Select i at random
- draw c from $P_s(\hat{y}[i] \mid y \backslash i) = P_s(\hat{y}[i] \mid \hat{y}[N(i)])$
- $\hat{y}[i] = c$

This algorithm does not require knowledge of Z.

The stationary distribution is P_s.

END