TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Monte-Carlo Markov Chain (MCMC) Sampling

Sampling From the Model

For back-propagation of $-\ln P_s(\mathcal{Y})$ through the exponential softmax defined by $P_s(\mathcal{Y}) = \frac{1}{Z}e^{s(\mathcal{Y})}$ we have

$$
\begin{aligned}
s^N.\mathrm{grad}[n, y] \;=\; & P_{\mathcal{Y}' \sim P_s}(\;\mathcal{Y}'[n] = y\;) \\
& -\mathbf{1}[\;\mathcal{Y}[n] = y\;]
\end{aligned}
$$

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

MCMC Sampling

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

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

Given such a process we can sample $\mathcal{Y}$ 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 $N(\mathcal{Y})$ be the set of neighbors of assignment $\mathcal{Y}$.

For example, $N(\mathcal{Y})$ can be taken to be the set of assignments $\mathcal{Y}'$ that differ form $\mathcal{Y}$ 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 — $\mathcal{Y}' \in N(\mathcal{Y})$ if and only if $\mathcal{Y} \in N(\mathcal{Y}')$.

Metropolis MCMC

Pick an initial state $\mathcal{Y}_0$ and for $t \geq 0$ do

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

2. If $s(\mathcal{Y}') > s(\mathcal{Y}_t)$ then $\mathcal{Y}_{t+1} = \mathcal{Y}'$

3. If $s(\mathcal{Y}') \leq s(\mathcal{Y})$ then with probability $e^{-\Delta s} = e^{-(s(\mathcal{Y})-s(\mathcal{Y}'))}$ do $\mathcal{Y}_{t+1} = \mathcal{Y}'$ and otherwise $\mathcal{Y}_{t+1} = \mathcal{Y}_t$

The Metropolis Markov Chain

We need to show that $P_s(\mathcal{Y}) = \frac{1}{Z}e^{s(\mathcal{Y})}$ is a stationary distribution of this process.

Let $Q(\mathcal{Y})$ be the distribution on states defined by drawing a state from P_s and applying one stochastic transition of the Metropolis process.

We must show that $Q(\mathcal{Y}) = P_s(\mathcal{Y})$.

The Stationary Distribution

Let $P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}')$ denote the probability of transitioning from $\mathcal{Y}$ to $\mathcal{Y}'$, or more formally,

$$P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}') = P(\mathcal{Y}_{t+1} = \mathcal{Y}' \mid \mathcal{Y}_y = \mathcal{Y})$$

We can then write $Q(\mathcal{Y}')$ as

$$Q(\mathcal{Y}') = \sum_{\mathcal{Y}} P_s(\mathcal{Y}) P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}')$$

The Stationary Distribution

$$
\begin{aligned}
Q(\mathcal{Y}') &= \sum_{\mathcal{Y}} P_s(\mathcal{Y}) P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}') \\
&= P_s(\mathcal{Y}') P_{\text{Trans}}(\mathcal{Y}' \to \mathcal{Y}') + \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_s(\mathcal{Y}) P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}') \\
&= \begin{cases} P_s(\mathcal{Y}') \left(1 - \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_{\text{Trans}}(\mathcal{Y}' \to \mathcal{Y})\right) \\ + \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_s(\mathcal{Y}) P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}') \end{cases}
\end{aligned}
$$

The Stationary Distribution

$$Q(\mathcal{Y}') = \begin{cases} P_s(\mathcal{Y}')\left(1 - \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_{\mathrm{Trans}}(\mathcal{Y}' \to \mathcal{Y})\right) \\ \\ + \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_s(\mathcal{Y}) P_{\mathrm{Trans}}(\mathcal{Y} \to \mathcal{Y}') \end{cases}$$

$$= \begin{cases} P_s(\mathcal{Y}') \\ \\ - \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_s(\mathcal{Y}') P_{\mathrm{Trans}}(\mathcal{Y}' \to \mathcal{Y}) \\ \\ + \sum_{\mathcal{Y} \in N(\mathcal{Y}')} P_s(\mathcal{Y}) P_{\mathrm{Trans}}(\mathcal{Y} \to \mathcal{Y}') \end{cases}$$

$$= P_s(\mathcal{Y}') - \text{flow out} + \text{flow in}$$

Detailed Balance

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

$$P_s(\mathcal{Y}')P_{\text{Trans}}(\mathcal{Y}' \to \mathcal{Y}) = P_s(\mathcal{Y})P_{\text{Trans}}(\mathcal{Y} \to \mathcal{Y}')$$

If we can show detailed balance we have that the flow out equals the flow in and we get $Q(\mathcal{Y}') = P_s(\mathcal{Y}')$ and hence P_s is the stationary distribution.

Detailed Balance

To show detailed balance we can assume without loss generality that $s(\mathcal{Y}') \geq s(\mathcal{Y})$.

We then have

$$
\begin{aligned}
P_s(\mathcal{Y}')P_{\mathrm{Trans}}(\mathcal{Y}' \to \mathcal{Y}) &= \frac{1}{Z}e^{s(\mathcal{Y}')} \quad \left(\frac{1}{N}\, e^{-\Delta s}\right) \\
&= \frac{1}{Z}e^{s(\mathcal{Y})} \quad \frac{1}{N} \\
&= P_s(\mathcal{Y})P_{\mathrm{Trans}}(\mathcal{Y} \to \mathcal{Y}')
\end{aligned}
$$

Gibbs Sampling

The Metropolis algorithm wastes time by rejecting proposed moves.

Gibbs sampling avoids this move rejection.

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

We let $\mathcal{Y}\backslash n$ be the assignment of labels given by $\mathcal{Y}$ except that no label is assigned to node n.

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

Gibbs Sampling

Markov Blanket Property:

$$P_s(\mathcal{Y}[n] \mid \mathcal{Y}\backslash n) = P_s(\mathcal{Y}[n] \mid \mathcal{Y}[N(n)])$$

Gibbs Sampling, Repeat:

- Select n at random
- draw y from $P_s(\mathcal{Y}[n] \mid \mathcal{Y}\backslash n) = P_s(\mathcal{Y}[n] \mid \mathcal{Y}[N(n)])$
- $\mathcal{Y}[n] = y$

This algorithm does not require knowledge of Z.

The stationary distribution is P_s.

END