TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Monte-Carlo Markov Chain (MCMC) Sampling

Sampling From the Model

For backpropagation through $P_s(\hat{\mathcal{Y}}) = \frac{1}{Z}e^{s(\hat{\mathcal{Y}})}$ we have

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

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

MCMC Sampling

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

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

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

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

Metropolis MCMC

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

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

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

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

The Metropolis Markov Chain

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

Let $Q(\hat{\mathcal{Y}})$ 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 $Q(\hat{\mathcal{Y}}) = P_s(\hat{\mathcal{Y}})$.

Stationarity Condition

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

Detailed Balance

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

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

Without loss generality assume $s(\hat{\mathcal{Y}}') \geq s(\hat{\mathcal{Y}})$.

Metropolis is defined by

$$P_{\mathrm{Trans}}(\hat{\mathcal{Y}}' \to \hat{\mathcal{Y}}) \;=\; \frac{1}{N}\, e^{-\Delta s} \;=\; P_{\mathrm{Trans}}(\hat{\mathcal{Y}} \to \hat{\mathcal{Y}}')\, \frac{P_s(\hat{\mathcal{Y}})}{P_s(\hat{\mathcal{Y}}')}$$

Gibbs Sampling

The Metropolis algorithm wastes time by rejecting proposed moves.

Gibbs sampling avoids this move rejection.

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

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

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

Gibbs Sampling

Markov Blanket Property:

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

Gibbs Sampling, Repeat:

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

This algorithm does not require knowledge of Z.

The stationary distribution is P_s.

END