Perfect Sampling

Yvan Quinn

The content of this pulls from various sources, but is ultimately based on Propp and Wilson (1996). From here on, assume that we are working with finite dimensional transition matrix P which produces an irreducible, aperiodic, recurrent Homogeneous Markov chain $\{X_i\}_{\mathbb{Z}}$ with finite state space S. Note that this guarantees the existence of an invariant π. In many cases it is possible to do something similar with a continuous state space, although convergence to the stationary distribution is only approximate (see Murdoch and Green, 1998).

1 setup

First we define maps $f_t : S \times [0,1]$ so that for any state $i \in S$, we have that $f_t(i, U_t)$ is distributed according to P_i given uniform U_t. Thus, we may use f_t to generate the next step in a chain with transition probabilities according to P, i.e. $X_{t+1} := f_t(X_t) := f_t(X_t, U_t)$. What this looks like in practice is that we use one random variable at each time step to determine what happens to each possible state in S under the action of f_t. This means that we can actually generate a whole family of chains from different initial states. In the simplest case, we choose $f_t(i)$ independently of $f_t(j)$ for all $i \neq j$, however this isn't strictly necessary. For (2) we will use this particular choice out of convenience.

Given a sequence of independent $\{U_i\}_{i\in\mathbb{Z}}$, we may now define the composition of maps $F_s^t = f_{t-1} \circ ... \circ f_s$. Since the same f_i generate all the chains in a family, these compositions traverse any of the chains in our family over time, i.e.

$$F_s^t(X_s) = (f_{t-1} \circ ... \circ f_s)(X_s) = (f_{t-1} \circ ... \circ f_{s+1})(X_{s+1}) = X_t \tag{1}$$

2 coupling from the past

One corollary proved in class stated that for an irreducible HMC with period d, for any two states i, j, there are constants $n_0(i)$ and $m(i,j) < d$ such that for any $n \geq n_0$:

$$p_{ij}(m + nd) > 0$$

Aperiodic chains have $d = 1$, and thus $m = 0$. By the finiteness of S, we define $L := \max_i n_0(i)$ for which $p_{ij}(L) > 0 \ \forall i, j \in S$. Since our chain has a nonzero probability of transitioning between any two states in L steps, there's a nonzero probability that F_{-L}^0 is constant (sends all of S to the same state):

$$\mathbb{P}\left(\exists j \in S : F_{-L}^0(S) = j\right) = \sum_j \mathbb{P}\left(F_{-L}^0(S) = j\right) = \sum_j \prod_i p_{ij}(L) > 0 \tag{2}$$

Since all f_i are distributed according to P and the U_i independent, all f_i are i.i.d. Thus all of $F_{-L}^0, F_{-2L}^{-L}, ...$ are i.i.d., so the probability of any one being constant is identically nonzero. Thus the probability of at least one being constant is

$$1 - \prod_{k=0}^{\infty} \mathbb{P}\left(F_{-(k+1)L}^{-kL} \text{is not constant}\right) = 1 - \prod \left(1 - \mathbb{P}\left(F_{-L}^0 \text{is constant}\right)\right) = 1$$

Once one of these compositions has settled on being constant, all successive maps are simply juggling us from one state to the next. Thus all possible initial states have coalesced by the time we reach time

0: the family of chains is coupled. In particular, there is an M for which F_{-M}^{-M+L} is constant, and thus $F_{-M}^{0} = F_{-L}^{0} \circ ... \circ F_{-M}^{-M+L}$ is constant.

Moreover, no matter what we input to a constant F_{-M}^{0} we will always get the same output, so $F_{-N}^{0} = F_{-M}^{0} \circ F_{-N+M}^{-M} = F_{-M}^{0}$ for all $N \geq M$. Consequently, we define $F_{-\infty}^{0} := F_{-M}^{0}$.

3 sampling π

By the time we reach time 0, any distribution of initial states at time $-M$ is now indistinguishable from if we had initially chosen according to the invariant distribution π. By the unique invariance of π under P (and thus the f_i and F), after coupling we must actually be sampling from π directly.

This can also be seen by taking the limit as $N \to \infty$ of X_N, which must converge to π regardless of the input. Since F_0^N, F_{-N}^0 are i.i.d., we have per (1)

$$\mathbb{P}\left(F_{-N}^{0}(X_0) = i\right) = \mathbb{P}\left(F_0^N(X_0) = i\right) = \mathbb{P}(X_N = i) \longrightarrow \pi(i)$$

In particular $F_{-N}^{0} = F_{-\infty}^{0}$ for all $N \geq M$, so for all the chains $\{X_i\}$ generated by the $\{f_i\}$,

$$\mathbb{P}(X_0 = i) = \mathbb{P}\left(F_{-\infty}^{0}(\,\cdot\,) = i\right) = \pi(i)$$

4 monotone chains, ex Ising Monte Carlo

Suppose our state space S is partially ordered and the maps f_t preserve this partial order (i.e. that $f_t(i) \leq f_t(j) \Leftrightarrow i \leq j$). It suffices to only check the extremal states for coupling, as they continue to bound the rest of the states under the maps!

We define a partial order of Ising states by comparing the spin at each site: for two configurations σ, σ', we say that $\sigma \geq \sigma'$ if $\sigma_k \geq \sigma'_k$ for every site k.

4.1 Metropolis

If our map f_t accepts a transition with probability $\min\left[1, e^{-\beta \Delta E}\right]$, then for states $\sigma \geq \sigma'$ we have that $h_k \geq h'_k$. Suppose our map randomly selects a site i then flips a coin to decide whether to try setting the spin to up or down.

WLOG suppose it selects $+$: if σ' ends up at a lower energy (i.e. $\Delta E' < 0$) then σ will either stay the same (i.e. if $\sigma_i = +$ already) or it will end up at a lower energy. Thus if $\sigma'_k \to +$ with probability 1, then so will σ. If σ' winds up in a higher energy state (i.e. $\Delta E' > 0$) then σ will necessarily have a smaller (possibly negative) change in energy $\Delta E' > \Delta E$. Thus,

$$\min\left[1, e^{-\beta \Delta E}\right] \geq \min\left[1, e^{-\beta \Delta E'}\right]$$

Since we are using the same random variable to select what happens to both states, this guarantees that $f_t(\sigma) \geq f_t(\sigma')$.

On the face of it this seems like exactly what we require for perfect sampling, however coupling takes an astronomical amount of time for critically low temperature and background field since the two stable states are widely separated.

4.1.1 Hopfield

As an illustrative example, we consider the Hopfield neural network which operates in much the same way as Metropolis except that it completely ignores the stochastic portion and deterministically

selects the minimum energy state. Thus the only way we will have coupling from the past is if our two extremal states deterministically converge to the same memory. This will never occur if there are multiple memories! We really shouldn't be surprised that coupling never occurs, as Hopfield certainly does not satisfy our assumption of irreducibility because all memories are stable states.

Given $s \geq s'$, we may still verify that the partial order is invariant under the update rule:

$$h_i = \sum_j w_{ij} s_j + \theta_i \geq \sum_j w_{ij} s'_j + \theta_i = h'_i \tag{3}$$

$$f_t(s_i(t-1)) = s_i(t) = \Theta(h_i(t-1)) \geq \Theta(h'_i(t-1)) = s'_i(t) = f_t(s'_i(t-1))$$

4.2 heat bath

Now consider transitions modeled according to the heat bath algorithm. Under one of our maps, all the possible states have the same operation evaluated at the same site. In particular, the same random variable is used to decide whether to flip the spin up or down, even though the threshold is different for each state.

Given a local field $h(i)$ at site i, we have probabilities of being set

$$\pi_h^{\pm} = \frac{e^{-\beta E^{\pm}}}{e^{-\beta E^{+}} + e^{-\beta E^{-}}} = \frac{e^{\mp \beta h}}{e^{-\beta h} + e^{\beta h}} = \frac{1}{1 + e^{\mp 2\beta h}}$$

Since these transitions happen irrespective of the previous state at site i (and how the change affects the overall energy/local energy of its neighbors), it is easier to check the invariance of the partial order under the map f_t whose transition probabilities are defined by the heat bath model. As previously noted $h \geq h'$, so

$$\pi_h^{+} = \frac{1}{1 + e^{-2\beta h}} \geq \frac{1}{1 + e^{-2\beta h'}} = \pi_{h'}^{+}$$

$$\pi_h^{-} = 1 - \pi_h^{+} \leq \pi_{h'}^{-}$$

This means that the threshold π^- for our random variable U_t to set the site σ_i to $+$ is lower than that for σ'_i, thus if $f_t(\sigma'_i) = +$ then $f_t(\sigma_i) = +$. Therefore $f_t(\sigma) \geq f_t(\sigma')$, preserving the partial order. The heat bath algorithm performs similarly in practice to the Metropolis algorithm, however it has higher stochasticity and thus couples quicker, making it more suitable for perfect sampling.

4.2.1 stochastic Hopfield

The discussion for the heat bath algorithm can be applied directly to stochastic Hopfield neural networks (with $\Theta = \text{sgn}$) since their update rule is governed by the same transition probabilities and as shown in (3), the "local field" terms satisfy the appropriate inequality: $h \geq h'$.

This is directly applicable to maximum likelihood training of the network, which requires calculation of the expected value of $\sum_{i,j} s_i s_j$ over all possible states. Ordinarily this is computed by sampling many randomly instantiated simulations after a long period of time in order to approximate a sample of the invariant distribution.

5 bibliography

Krauth, Werner. *Statistical Mechanics: Algorithms and Computations*, 1.1, 5.2

`https://www.stat.berkeley.edu/~aldous/206-RWG/RWGpapers/propp_wilson.pdf`

`https://personal.math.ubc.ca/~jhermon/Mixing/Coupling_from_the_past.pdf`

`http://www.cs.cmu.edu/~bhiksha/courses/deeplearning/Spring.2018/www/slides/lec21.hopfield.pdf`