C19 : Lecture 3 : Markov Chain Monte Carlo

Frank Wood

University of Oxford

January, 2017

Many figures from PRML [Bishop, 2006]

Markov Chain Monte Carlo (MCMC)

- Rejection and importance sampling fail in high dimensions
- MCMC works better in high dimensions
- Various Algorithms
  - Metropolis Hastings
  - Gibbs
  - Metropolis-Hastings within Gibbs
  - Hamiltonian Monte Carlo (HMC)
- Can mix and match

Remember : Inference is all about integration and Monte Carlo integration is all about sampling

- *Important :* MCMC works by constructing and simulating a **Markov chain** whose equilibrium distribution is the distribution of interest

Metropolis Hastings (MH)

Algorithm

Initialize $\tau \leftarrow 1, \mathbf{x}^{(\tau)} \leftarrow ?$

Repeat Forever Yielding $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$

1. Propose $\mathbf{x}^* \sim q(\mathbf{x}^*|\mathbf{x}^{(\tau)})$
2. Accept $\mathbf{x}^*$ w.p. $A(\mathbf{x}^*, \mathbf{x}^{(\tau)}) = \min\left(1, \frac{p(\mathbf{x}^*)q(\mathbf{x}^{(\tau)}|\mathbf{x}^*)}{p(\mathbf{x}^{(\tau)})q(\mathbf{x}^*|\mathbf{x}^{(\tau)})}\right)$
3. If $\mathbf{x}^*$ accepted set $\mathbf{x}^{(\tau+1)} \leftarrow \mathbf{x}^*$ else $\mathbf{x}^{(\tau+1)} \leftarrow \mathbf{x}^{(\tau)}$
4. Increment τ

Common choices of proposal include $q(\mathbf{x}^*|\mathbf{x}^{(\tau)}) = \mathcal{N}(\mathbf{x}^{(\tau)}|\sigma^2\mathbf{I})$ (random-walk Metropolis) and/or $q(\mathbf{x}^*|\mathbf{x}^{(\tau)}) = q(\mathbf{x}^*)$ (independent MH). Rules of thumb suggest aiming for acceptance rates of between 25% and 50% by tuning the proposal distribution.

http://www.stat.duke.edu/~km68/materials/214.7%20(MH).pdf

Gibbs

Given a joint $p(\mathbf{x}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ s.t. $\mathbf{x}_1 \cup \cdots \cup \mathbf{x}_i \cup \cdots \cup \mathbf{x}_M = \mathbf{x}$ are subsets of the dimensions of $\mathbf{x}$ consider the case where we can sample exactly from $p(\mathbf{x}_i | \mathbf{x} \setminus \mathbf{x}_i)$

Algorithm

Initialize $\tau \leftarrow 1, \mathbf{x}^{(\tau)} \leftarrow ?$

Repeat Forever Yielding $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$

1. Sample $\mathbf{x}_1^{(\tau+1)} \sim p(\mathbf{x}_1 | \mathbf{x}^{(\tau)} \setminus \mathbf{x}_1)$
2. Sample $\mathbf{x}_2^{(\tau+1)} \sim p(\mathbf{x}_2 | \mathbf{x}_1^{(\tau+1)} \cup \mathbf{x}^{(\tau)} \setminus \{\mathbf{x}_1 \cup \mathbf{x}_2\})$
3. Sample $\mathbf{x}_3^{(\tau+1)} \sim p(\mathbf{x}_3 | \mathbf{x}_2^{(\tau+1)} \cup \mathbf{x}_1^{(\tau+1)} \cup \mathbf{x}^{(\tau)} \setminus \{\mathbf{x}_1 \cup \mathbf{x}_2 \cup \mathbf{x}_3\})$
4. $\vdots$
5. Sample $\mathbf{x}_M^{(\tau+1)} \sim p(\mathbf{x}_M | \mathbf{x}_{M-1}^{(\tau+1)} \cup \cdots \cup \mathbf{x}_1^{(\tau+1)})$
6. Increment τ

Understanding MCMC

- MH and Gibbs are example MCMC sampling algorithms
- MCMC sampling is based on simulating Markov chains with carefully designed, special, "general purpose" **transition operators**
- Understanding Markov chains and the design of such operators leads to an understanding of sampling and Monte Carlo integration
- MCMC = default choice for anytime inference algorithm

Warning

In the following we are very lazy mathematically. In much of what follows the *algorithms* have been proved to be correct in general settings whereas the *arguments* provided generally apply only to Markov chains defined on discrete state spaces. One justification for this is that all Markov chain simulation on digital computers is actually performed over discrete spaces; however, deeper consideration *must* be given to proof and justification in (common) situations where densities and continuous variables are used.

Markov Chain Review

A first order Markov chain is on which, for $m \in 1, \ldots, M$ and for a sequence of random variables $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ the following conditional *in*dependence property holds

$$p(\mathbf{x}^{(m+1)}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}) \equiv p(\mathbf{x}^{(m+1)}|\mathbf{x}^{(m)})$$

Markov Chain review material largely from [Neal, 1993].

Markov Chains for MCMC II

Such a Markov chain can be specified by the initial distribution $p(\mathbf{x}^{(0)})$ and a stochastic transition function

$$T_m(\mathbf{x}^{(m)}|\mathbf{x}^{(m+1)}) \equiv p(\mathbf{x}^{(m+1)}|\mathbf{x}^{(m)})$$

Definition : Homogenous Markov Chain

A Markov chain is **homogenous** if

$$T_1 = T_2 = \ldots T_M = T$$

i.e. the transition functions are the same for all m.

Markov Chains for MCMC III

The marginal distribution of a particular var can be expressed as

$$p(\mathbf{x}^{(m+1)}) = \sum_{\mathbf{x}^{(m)}} p(\mathbf{x}^{(m+1)}|\mathbf{x}^{(m)})p(\mathbf{x}^{(m)})$$

Definition : Invariant/Stationary Distribution

A distribution is said to be **invariant** or **stationary** w.r.t a Markov chain if the transition function of that chain leaves that distribution unchanged. For example, the distribution $p^*(\mathbf{x})$ is the invariant distribution of the Markov chain with transition operator $T(\mathbf{x}', \mathbf{x})$ if

$$p^*(\mathbf{x}) = \sum_{\mathbf{x}'} T(\mathbf{x}', \mathbf{x}) p^*(\mathbf{x}')$$

Note :

- Trivial transition distributions (identity) are not of interest.
- *Designing* general purpose *transition operators* that can have *any* distribution of interest be their stationary distribution is the key to MCMC.

Definition : Detailed Balance

If a transition operator T satisfies *detailed balance* it means that

$$p^*(\mathbf{x})\,T(\mathbf{x},\mathbf{x}') = p^*(\mathbf{x}')\,T(\mathbf{x}',\mathbf{x})$$

Markov Chains for MCMC VI

If a transition operator satisfies detailed balance w.r.t. a particular distribution then that distribution will be invariant under T.

$$\begin{aligned}
\sum_{\mathbf{x}'} p^*(\mathbf{x}')\,T(\mathbf{x}',\mathbf{x}) &= \sum_{\mathbf{x}'} p^*(\mathbf{x})\,T(\mathbf{x},\mathbf{x}') && \text{def. detailed balance} \\
&= p^*(\mathbf{x}) \sum_{\mathbf{x}'} p(\mathbf{x}'|\mathbf{x}) && \text{def. of } T \\
&= p^*(\mathbf{x}) \quad \square
\end{aligned}$$

So a Markov chain *designed* to satisfy detailed balance will have $p^*(\mathbf{x})$ as its stationary distribution.

Markov Chains for MCMC VII

We must further restrict our choice of T to those that that for $m \to \infty$ the distribution $p(\mathbf{x}^{(m)}|\mathbf{x}^{(0)})$ converges to the invariant distribution $p^*(\mathbf{x})$ regardless of choice of $p(\mathbf{x}^{(0)})$. This Markov chain property is called **ergodicity**.

Among other things, poorly designed T's could partition space (the "set of states") such that some subsets are unreachable from others.

Homework :

- What do *irreducible* and *aperiodic* mean when applied to Markov chains?

Fundamental Theorem

If a homogeneous Markov chain on a *finite* state space with transition probability $T(z, z')$ has π as an invariant distribution and

$$\nu = \min_{z} \min_{z':\pi(z')>0} T(z, z')/\pi(z') > 0$$

then

1. that Markov chain is *ergodic*, i.e. for all z regardless of the initial distribution $p_0(z)$
$$\lim_{n\to\infty} p_n(z) = \pi(z)$$
2. if $g(z)$ is a real-valued function of the state, then the expectation of g w.r.t. p_n, denoted $\mathbb{E}_n[g] = \sum g(z)p_n(z)$ converges to its expectation w.r.t. π, i.e. $\sum g(z)\pi(z)$.

Sketch proof in [Neal, 1993]

Markov Chains for MCMC IX

Towards designing general purpose MCMC transition operators we have

- One way to ensure T the is ergodic, i.e. a way to ensure T avoids traps and can visit everywhere in our state space
- A sufficient condition to ensure that T has the equilibrium distribution π^* we want (via detailed balance)

We also need to know that averaging over simulations of / samples from a Markov chain with such a T and stationary distribution π^* average nicely.

A Paraphrase of the Strong LLN for Markov Chains

For $z^{(0)}, z^{(1)}, \ldots$ generated by simulating a "nice" Markov chain having stationary distribution $\pi^*(\cdot)$.

$$\lim_{n \to \infty} \frac{\mathbb{I}(z^{(n)} = i)}{n} = \pi^*(i)$$

See [Breiman, 1960] for a proper statement and proof. Neal [1993] discusses this more readably but less precisely

Two Resulting Views on Computing Expectations

Computing expectations can be done via the

- Fundamental theorem, sampling from $p_n(z)$ via "parallel chains" and forward simulation
- Strong LLN, sample by simulating a single Markov chain for a long time, and use all samples from sequence

Problem

- Don't know when chains have "burned-in" (i.e. when the set of samples is close-enough distributed according to p^* rather than being heavily dominated by initial choice of starting point)
- Don't know when to stop (i.e. when the expectation approximation to the integral good enough, particularly if samples are not iid)
- Worse : can't parameterize stopping time n by quality of estimate desired

Reality

- Sometimes you don't care

`http://users.stat.umn.edu/~geyer/mcmc/one.html` has an amusing take on this tension

Metropolis Hastings as Markov Chain I

Remember

MH algorithm for sampling from p

Initialize $\tau \leftarrow 1$,$\mathbf{x}^{(\tau)} \leftarrow$?, repeat forever yielding $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$

1. Propose $\mathbf{x}^* \sim q(\mathbf{x}^*|\mathbf{x}^{(\tau)})$
2. Accept $\mathbf{x}^*$ w.p. $A(\mathbf{x}^*, \mathbf{x}^{(\tau)}) = \min\left(1, \frac{p(\mathbf{x}^*)q(\mathbf{x}^{(\tau)}|\mathbf{x}^*)}{p(\mathbf{x}^{(\tau)})q(\mathbf{x}^*|\mathbf{x}^{(\tau)})}\right)$
3. If $\mathbf{x}^*$ accepted set $\mathbf{x}^{(\tau+1)} \leftarrow \mathbf{x}^*$ else $\mathbf{x}^{(\tau+1)} \leftarrow \mathbf{x}^{(\tau)}$
4. Increment τ

and

Detailed Balance

If a transition operator T satisfies *detailed balance* it means that

$$p(\mathbf{x})T(\mathbf{x}, \mathbf{x}') = p(\mathbf{x}')T(\mathbf{x}', \mathbf{x})$$

Metropolis Hastings as Markov Chain II

To show that the MH algorithm satisfies detailed balance note that $T(\mathbf{x}, \mathbf{x}') = q(\mathbf{x}'|\mathbf{x})A(\mathbf{x}, \mathbf{x}')$ so

$$\begin{aligned} p(\mathbf{x})T(\mathbf{x}, \mathbf{x}') &= p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})A(\mathbf{x}, \mathbf{x}') \\ &= \min\left(p(\mathbf{x})q(\mathbf{x}'|\mathbf{x}), p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')\right) \\ &= \min\left(p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}'), p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})\right) \\ &= p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')A(\mathbf{x}', \mathbf{x}) \\ &= p(\mathbf{x}')T(\mathbf{x}', \mathbf{x}) \quad \square \end{aligned}$$

Which, provided q is chosen so that

$$\min_{\mathbf{x}} \min_{\mathbf{x}':p(\mathbf{x}')>0} \left(q(\mathbf{x}'|\mathbf{x})A(\mathbf{x}, \mathbf{x}')\right)/p(\mathbf{x}') > 0$$

means that MH is a general purpose Markov chain for simulating from and thereby computing expectations against arbitrary distributions p.

Study Suggestions

- Show that Gibbs is MH with a transition operator that always accepts
- Relate the invariant distribution statement to an eigenvector problem
- Prove that the independent MH operator transition yields a valid sampler
- Implement the MH algorithm for sampling from a multivariate Gaussian
- Implement the Gibbs algorithm for sampling from a multivariate Gaussian

Bibliography I

C M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Leo Breiman. The strong law of large numbers for a class of markov chains. *The Annals of Mathematical Statistics*, 31(3):801–803, 1960.

Radford M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report*, 1993.