Computational Statistics & Machine Learning

Lecture 5

Markov chain Monte Carlo

Mark Girolami
mag92@cam.ac.uk

Department of Engineering
University of Cambridge

September 16, 2021

Overview

- Bayesian Probabilistic Inference
- Markov chain Monte Carlo
- Metropolis-Hastings Algorithms
- Derivation of Metropolis-Hastings

Probability Inversion

- Rev Thomas Bayes - Probability Inversion

- The Rev Thomas Bayes FRS, 1701 - 1761

Probability Inversion

- Data $\mathbf{x} \in \mathcal{X}$ and parameters, including latent variables $\boldsymbol{\theta} \in \Theta$
- In model a *Prior* distribution for all unknowns is defined $\pi_0(\boldsymbol{\theta})$
- The data under the model has *Likelihood* $p(\mathbf{x}|\boldsymbol{\theta})$
- The *Posterior* distribution follows as

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- Inference requires integration w.r.t. the posterior e.g. expectations

$$E_{\boldsymbol{\theta}|\mathbf{x}}\{f(\boldsymbol{\theta})\} = \int_{\Theta} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}$$

- In most cases $\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}$ is non-analytic
- Progress in Bayesian inference halted beyond *Conjugate* prior and likelihood functions

Probability Inversion

- Monte Carlo estimation of intractable integrals of the following kind

$$I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}|\mathbf{x}}\{f(\boldsymbol{\theta})\} = \int_{\boldsymbol{\Theta}} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}$$

- Estimates can be obtained by stochastic simulation (hence the Monte Carlo moniker)
- Obtain independent (i.i.d) samples $\boldsymbol{\theta}^{(n)} \sim \pi(\boldsymbol{\theta}|\mathbf{x})$
- Obtain the *plug-in* Monte-Carlo estimate

$$\hat{I}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\theta}^{(n)}) \rightarrow E_{\boldsymbol{\theta}|\mathbf{x}}\{f(\boldsymbol{\theta})\}$$

- Convergence: *Unbiased*, *Consistent* estimator as $N \rightarrow \infty$, $\hat{I}_N(\boldsymbol{\theta}) \rightarrow I(\boldsymbol{\theta})$
- Convergence of MC error $\sqrt{N}[\hat{I}_N(\boldsymbol{\theta}) - I(\boldsymbol{\theta})] \rightarrow \mathcal{N}(0, \sigma_f^2)$

Markov chains

- Now have a means to obtain unbiased, consistent estimators of expectations w.r.t. intractable distributions
- Require to obtain i.i.d. samples from the target distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$
- Most general scheme is to employ Markov chains as a means of stochastic simulation from $\pi(\boldsymbol{\theta}|\mathbf{x})$

- Andrey Markov 1856 - 1922

Markov chains

- Studies of stochastic transitions on discrete and continuous state spaces governed by the Markov property

$$p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^{(i-2)}, ..., \boldsymbol{\theta}^{(1)}) = T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)})$$

- Homogenous transitions for $T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)})$ remains invariant for all (i).
- Transitions satisfy $\sum_{\boldsymbol{\theta}^{(i)}} T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)}) = 1$ or $\int_{\boldsymbol{\theta}^{(i)} \in \Theta} T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)}) d\boldsymbol{\theta}^{(i)} = 1$
- Discrete example

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{pmatrix}$$

- If $\pi(\boldsymbol{\theta}^{(1)}) = [0.5, 0.2, 0.3]$ then $\pi(\boldsymbol{\theta}^{(2)}) = \pi(\boldsymbol{\theta}^{(1)})^T \mathbf{T} = [0.18, 0.64, 0.18]$
- $\pi(\boldsymbol{\theta}^{(3)}) = \pi(\boldsymbol{\theta}^{(2)})^T \mathbf{T} = \pi(\boldsymbol{\theta}^{(1)})^T \mathbf{T}\mathbf{T} = \pi(\boldsymbol{\theta}^{(1)})^T \mathbf{T}^2$, $\pi(\boldsymbol{\theta}^{(i)}) = \pi(\boldsymbol{\theta}^{(1)})^T \mathbf{T}^{(i-1)}$

Markov chains

- Discrete example

$$\mathbf{T}^2 = \begin{pmatrix} 0 & 0.1 & 0.9 \\ 0.54 & 0.37 & 0.09 \\ 0 & 0.64 & 0.36 \end{pmatrix} \quad \mathbf{T}^5 = \begin{pmatrix} 0.1998 & 0.2485 & 0.5517 \\ 0.3310 & 0.4453 & 0.2237 \\ 0.1123 & 0.4672 & 0.4205 \end{pmatrix}$$

$$\mathbf{T}^8 = \begin{pmatrix} 0.2405 & 0.2410 & 0.4185 \\ 0.2511 & 0.4210 & 0.3069 \\ 0.1767 & 0.4164 & 0.4079 \end{pmatrix} \quad \mathbf{T}^{16} = \begin{pmatrix} 0.2174 & 0.4066 & 0.3760 \\ 0.2256 & 0.4085 & 0.3659 \\ 0.2189 & 0.4133 & 0.3678 \end{pmatrix}$$

$$\mathbf{T}^{32} = \begin{pmatrix} 0.2213 & 0.4099 & 0.3688 \\ 0.2213 & 0.4098 & 0.3689 \\ 0.2213 & 0.4098 & 0.3688 \end{pmatrix} \quad \mathbf{T}^{64} = \begin{pmatrix} 0.2213 & 0.4098 & 0.3689 \\ 0.2213 & 0.4098 & 0.3689 \\ 0.2213 & 0.4098 & 0.3689 \end{pmatrix}$$

- Therefore $\pi(\boldsymbol{\theta}^{(\infty)}) = [0.2213, 0.4098, 0.3689]$ irrespective of $\pi(\boldsymbol{\theta}^{(1)})$

Markov chains

- The Markov chain converges to the Invariant Distribution of the chain $\pi(\boldsymbol{\theta}^{(\infty)}) \rightarrow \pi(\boldsymbol{\theta})$ irrespective of $\pi(\boldsymbol{\theta}^{(1)})$ provided
- The transition operator $\mathbf{T}$ is Irreducible i.e. the transition graph is fully connected therefore cannot be reduced to smaller transition sub-graphs meaning non-zero probability to arrive at all other states.
- The Chain is Aperiodic i.e. no cycles cannot escape from.
- A sufficient condition for an Invariant distribution to exist is if the chain is Reversible also refered to as satisfying Detailed Balance.

$$\pi(\boldsymbol{\theta}^{(i)})T(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(i)}) = \pi(\boldsymbol{\theta}^{(i-1)})T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)})$$

$$\sum_{\boldsymbol{\theta}^{(i-1)}} \pi(\boldsymbol{\theta}^{(i)})T(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(i)}) = \sum_{\boldsymbol{\theta}^{(i-1)}} \pi(\boldsymbol{\theta}^{(i-1)})T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)})$$

$$\pi(\boldsymbol{\theta}^{(i)}) = \sum_{\boldsymbol{\theta}^{(i-1)}} \pi(\boldsymbol{\theta}^{(i-1)})T(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)})$$

- Study of Markov chain consider conditions for existence of associated invariant distribution for transition operator
- MCMC knows the invariant distribution and considers definition of the required transition operator

Metropolis-Hastings Algorithm

- You met Nick Metropolis in a previous lecture who along with co-authors described a method in the paper *Equation of State Calculations by Fast Computing Machines* in 1953
- Wilfred Hastings generalised the original algorithm in 1970 when at the University of Toronto

Metropolis-Hastings Algorithm

- Presentation now uses continuous domain arguments, since they are more general. These arguments also hold for the discrete case
- Draw samples from $\pi^*(dy) = \int_{\mathcal{R}^d} P(x, dy)\pi(x)dx$, $\pi^*(dy) = \pi(y)dy$
- As in discrete case $P^{(n)}(x, A) = \int_{\mathcal{R}^d} P^{(n-1)}(x, dy)P^{(1)}(y, A)$
- Require to define the transition kernel operator $P(x, dy)$ so that it has $\pi(\cdot)$ as its invariant density
- Consider some function $p(x, y)$ and define a transition operator as

$$P(x, dy) = p(x, y)dy + r(x)\delta_x(dy)$$

- If $p(x, x) = 0$ and $\int_{\mathcal{R}^d} P(x, dy) = 1$ then $r(x) = 1 - \int_{\mathcal{R}^d} p(x, y)dy$, probability that chain remains at x
- If $p(x, y)$ satisfies reversibility $\pi(x)p(x, y) = \pi(y)p(y, x)$ then $\pi(\cdot)$ is the invariant density of the transition kernel $P(x, \cdot)$.
- This is cool, need to prove it though.

Metropolis-Hastings Algorithm

- $\int_{\mathcal{R}^d} P(x, A)\pi(x)dx$ is equal to
-

$$\begin{aligned}
& \int \left[\int_A p(x,y)dy\right] \pi(x)dx + \int r(x)\delta_x(A)\pi(x)dx \\
= & \int_A \left[\int p(x,y)\pi(x)dx\right] dy + \int_A r(x)\pi(x)dx \\
= & \int_A \left[\int p(y,x)\pi(y)dx\right] dy + \int_A r(x)\pi(x)dx \\
= & \int_A (1 - r(y))\pi(y)dy + \int_A r(x)\pi(x)dx \\
= & \int_A \pi(y)dy = \pi^*(A)
\end{aligned}$$

- The Reversibility condition for $p(x, y)$ is sufficient in designing a transition operator to target a specific invariant distribution

Metropolis-Hastings Algorithm

- Now we have the form of the transition kernel that will leave our desired distribution invariant require algorithm
- Now need to define the form of $p(x, y)$ explicitly
- Consider a Proposal density $q(x, y)$ s.t. $\int q(x, y)dy = 1$, if $q(x, y)$ is reversible we are finished
- This may hold $\pi(x)q(x, y) > \pi(y)q(y, x)$ x to y transitions too frequent
- Rebalance both sides by introducing an Acceptance probability $\alpha(x, y)$
- In this case $\alpha(x, y) < 1$ so now reversibility is established as $\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$
- Need form of $\alpha(x, y)$, maximum value of α is 1, set $\alpha(y, x) = 1$
- Then $\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)$ and

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

- What about if $\pi(x)q(x, y) < \pi(y)q(y, x)$?$\alpha(x, y) = 1$

Metropolis-Hastings Algorithm

- We now have that $p_{MH}(x,y) = q(x,y)\alpha(x,y)$ for $x \neq y$ with
$$\alpha(x,y) = min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right)$$
- Overall transition kernel $P_{MH}(x,dy)$ is equal to
$$q(x,y)\alpha(x,y)dy + \left[1 - \int_{\mathcal{R}^d} q(x,y)\alpha(x,y)dy\right]\delta_x(dy)$$
is reversible by construction and hence has $\pi(x)$ as its invariant density.
- Note that MH is defined by the proposal density $q(x,y)$ which requires to be chosen
- If proposed move is rejected the current value is taken as the next one in the chain
- The acceptance probability is $\pi(y)q(y,x)/\pi(x)q(x,y)$ for any density $\pi(x) = \phi(x)/\int \phi(x)dx$
- then
$$\alpha(x,y) = min\left(\frac{\phi(y)q(y,x)}{\phi(x)q(x,y)}, 1\right)$$
troublesome normalising constants not required. Awesome.

Metropolis-Hastings Algorithm

- What if $q(x, y)$ is symmetric i.e. $q(x, y) = q(y, x)$ e.g. $\mathcal{N}(x|y, \Sigma) = \mathcal{N}(y|x, \Sigma)$?
- then

$$\alpha(x, y) = min\left(\frac{\phi(y)}{\phi(x)}, 1\right)$$

  ratio of unnormalised densities only..... Nice
- Suggesting that moves to regions of higher density always accepted
- Moves to lower density regions accepted with probability $\phi(y)/\phi(x)$

Metropolis-Hastings Algorithm

- Overall Metropolis-Hastings algorithm follows as

for $j = 1 \rightarrow N$ **do**
  Simulate y from $q(x^{(j)}, \cdot)$
  Simulate u from $U(0, 1)$
  if $u \leq \alpha(x^{(j)}, y)$ **then**
    Set $x^{(j+1)} = y$
  else
    Set $x^{(j+1)} = x^{(j)}$
  end if
end for
Return $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(N)}\}$