Computational Statistics & Machine Learning

Lecture 6

Markov chain Monte Carlo

Mark Girolami
mag92@cam.ac.uk

Department of Engineering
University of Cambridge

October 30, 2021

Overview

- Product Transition Kernels
- Gibbs Sampler
- Mixture Transition Kernels
- Data Augmentation

Product Transition Kernels

- Simple illustrative examples are all very well - what if target density is multivariate?
- Breaking the vector up into sub-blocks is a good strategy to design proposals?
- Does this ensure convergence to correct target?
- Let $x = (x_1, x_2)$ where $x \in \mathcal{R}^d$, and each $x_i \in \mathcal{R}^{d_i}$
- Conditional transition kernel $P_1(x_1, dy_1|x_2)$ with invariant density $\pi_{1|2}(\cdot|x_2)$ for fixed value of x_2
- Conditional transition kernel $P_2(x_2, dy_2|x_1)$ with invariant density $\pi_{2|1}(\cdot|x_1)$ for fixed value of x_1
- Standard conditions apply $\pi^*_{1|2}(dy_1|x_2) = \int P_1(x_1, dy_1|x_2)\pi_{1|2}(x_1|x_2)dx_1$
- Standard conditions apply $\pi^*_{2|1}(dy_2|x_1) = \int P_2(x_2, dy_2|x_1)\pi_{2|1}(x_2|x_1)dx_2$
- The product kernel $P_1(x_1, dy_1|x_2)P_2(x_2, dy_2|y_1)$ has invariant density $\pi(dy_1, dy_2)$
- Wakeup Now for exercise - prove the above.

Product Transition Kernels

$$
\begin{aligned}
& \int\int P_1(x_1, dy_1|x_2) P_2(x_2, dy_2|y_1) \pi(x_1, x_2) dx_1 dx_2 \\
= & \int P_2(x_2, dy_2|y_1) \left[\int P_1(x_1, dy_1|x_2) \pi_{1|2}(x_1|x_2) dx_1 \right] \pi_2(x_2) dx_2 \\
= & \int P_2(x_2, dy_2|y_1) \pi^*_{1|2}(dy_1|x_2) \pi_2(x_2) dx_2 \\
= & \int P_2(x_2, dy_2|y_1) \frac{\pi_{2|1}(x_2|y_1)\pi^*_1(dy_1)}{\pi_2(x_2)} \pi_2(x_2) dx_2 \\
= & \pi^*_1(dy_1) \int P_2(x_2, dy_2|y_1) \pi_{2|1}(x_2|y_1) dx_2 = \pi^*_1(dy_1) \pi^*_{2|1}(dy_2|y_1) \\
= & \pi^*(dy_1, dy_2)
\end{aligned}
$$

Gibbs Sampler

- Set kernels $P_1(x_1, dy_1|x_2) = \pi^*_{1|2}(dy_1|x_2)$ & $P_2(x_2, dy_2|y_1) = \pi^*_{2|1}(dy_2|y_1)$
- Then $\alpha(x, y) = 1$ for the transition of the first coordinate x_1 given x_2:

$$
\begin{aligned}
\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} &= \frac{\pi(y_1,y_2)q([y_1,y_2],[x_1,x_2])}{\pi(x_1,x_2)q([x_1,x_2],[y_1,y_2])} \\
&= \frac{\pi(y_1|y_2)\pi(y_2) \times \pi(x_1|x_2)}{\pi(x_1|x_2)\pi(x_2) \times \pi(y_1|x_2)} \\
&= \frac{\pi(y_1|x_2)\pi(x_2) \times \pi(x_1|x_2)}{\pi(x_1|x_2)\pi(x_2) \times \pi(y_1|x_2)} \\
&= 1
\end{aligned}
$$

Gibbs Sampler

- Target density is $\mathcal{N}(\mathbf{0}, \mathbf{C})$ where $C_{1,1} = C_{2,2} = 1$ and $C_{1,2} = C_{2,1} = \rho$
- MH Algorithm with proposal $\mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma\mathbf{I})$
- MH acceptance ratio

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left[\frac{\exp(-0.5 \times (\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}))}{\exp(-0.5 \times (\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}))}, 1\right]$$

- What issues will arise with the choice of proposal ?
- MH with exact conditionals as proposals = Gibbs Sampler
-

$$\begin{aligned} y_1|x_2 &\sim \mathcal{N}(\rho \times x_2, 1 - \rho^2) \\ y_2|y_1 &\sim \mathcal{N}(\rho \times y_1, 1 - \rho^2) \end{aligned}$$

Mixture Transition Kernels

- For two transition kernels $P_1(x, dy)$ and $P_2(x, dy)$ that both have $\pi(x)$ as invariant density
- and a probability $0 \leq \gamma \leq 1$
- what is the invariant density of the kernel composed of a mixture $\gamma \times P_1(x, dy) + (1 - \gamma) \times P_2(x, dy)$?
- Convince yourself that it is $\pi(x)$
- Can you think of situations where this might be useful ?
- Imagine a target density that may have multiple modes

Properties

- Monte Carlo estimates require i.i.d samples
- From our MCMC runs will the samples be i.i.d ?
- We know that repeated application of the transition kernel will lead to the desired target density
- How long will it be before we can be sure that the chain has converged ?

Data Augmentation

- Consider a target density $\pi(\theta)$ where $\theta \in \mathbb{R}^D$
- Let us now augment our model by introducing $\phi \in \mathbb{R}^D$
- Then the augmented target density is $\pi(\theta, \phi)$
- The desired density can be recovered as $\pi(\theta) = \int \pi(\theta, \phi) d\phi$
- Therefore devising a Markov chain whose invariant density is $\pi(\theta, \phi)$
- and we draw samples $\theta^{(n)}, \phi^{(n)} \sim \pi(\theta, \phi)$
- then for each $\theta^{(n)}$ it follows that $\int \pi(\theta^{(n)}, \phi) d\phi = \pi(\theta^{(n)})$
- Nice result..... each $\theta^{(n)}$ is marginally distributed with density $\pi(.)$

Data Augmentation

- The main point is that we need to find a representation or completion that will admit exact conditionals for $\pi(\phi|\theta)$ and $\pi(\theta|\phi)$ that can be sampled from directly.
- Consider a Binary Probit Regression example popular in Machine Learning for classification problems.
- The probability for binary response variable t is $p(t = 1|x, \beta) = \Phi(\beta' x)$, where Φ is the standard normal CDF
- Employ DA by introducing the auxiliary variable y such that
- $y = \beta' x + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$
- Further define $t_i = 1 : y_i > 0$ and $t_i = 0 : y_i \leq 0$
- The joint density $p(t_i, y_i|x_i, \beta) = p(t_i|y_i)p(y_i|x_i, \beta)$
- By defintion $p(t_i = 1|y_i) = \delta_{(t_i=1)}\delta_{(y_i>0)}$ and $p(t_i = 0|y_i) = \delta_{(t_i=0)}\delta_{(y_i\leq 0)}$
- Likewise $p(y_i|x_i, \beta) = \mathcal{N}(y_i|\beta' x_i, 1)$

Data Augmentation

- The joint density $p(\mathbf{t}, \mathbf{y}|\beta, \mathbf{X})$ follows as

$$\prod_i [\delta_{(t_i=1)}\delta_{(y_i>0)} + \delta_{(t_i=0)}\delta_{(y_i\leq 0)}] \times \mathcal{N}(y_i|\beta' x_i, 1)$$

- Now derive full conditional $p(\beta|\mathbf{t}, \mathbf{y}, \mathbf{X})$
- $p(\beta|\mathbf{t}, \mathbf{y}, \mathbf{X}) \propto \prod_i \mathcal{N}(y_i|\beta' x_i, 1)$
- $p(\beta|\mathbf{t}, \mathbf{y}, \mathbf{X}) = \mathcal{N}(\beta|(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, (\mathbf{X}'\mathbf{X})^{-1})$
- Now derive full conditional for $p(\mathbf{y}|\mathbf{X}, \mathbf{t}, \beta)$
- This is a product of simple truncated Normals so that

$$\begin{aligned} y_i &\sim TN_{(0,\infty)}(\beta' x_i, 1) : t_i = 1 \\ y_i &\sim TN_{(\infty,0)}(\beta' x_i, 1) : t_i = 0 \end{aligned}$$

Data Augmentation

- The overall Gibbs sampling scheme follows as

$$
\begin{aligned}
\beta|\mathbf{t},\mathbf{y},\mathbf{X} &\sim \mathcal{N}(\beta|(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},(\mathbf{X}'\mathbf{X})^{-1}) \\
y_i &\sim TN_{(0,\infty)}(\beta' x_i,1) : t_i = 1 \\
y_i &\sim TN_{(\infty,0)}(\beta' x_i,1) : t_i = 0 \quad for \quad all \quad i
\end{aligned}
$$

- A prior can be placed on β i.e. $\pi_0(\beta) = \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$
- In the examples paper you will derive the Gibbs sampler with a prior on β and this is implemented in the notebooks.