Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Harshvardhan
Instructor: Purushottam Kar
Date: March 11, 2018

LECTURE
25

MCMC, Metropolis Hastings - II

1 Introduction

The previous sets of lectures have been devoted to the analysis of probabilistic methods like EM. The previous lecture laid foundations for analysing vey widely used and powerful family of sampling algorithms namely, Markov Chain Monte Carlo (MCMC) and Metropolis-hastings (MH) algorithms. These allow us to generate samples from hard to sample from probability distributions.

Both these methods define a transition law for a Markov chain and claim that after taking sufficient number of steps, the samples generated will be from the required distribution. MH methods use a proposal distribution for sampling with an acceptance criteria while MCMC methods start with an initial distribution and with certain conditions on the transition matrix converge to a stationary distribution. As seen before, upon imposing certain conditions on the proposal distribution and acceptance criteria, the MH method reduces to the MCMC method.

This lecture provides proofs of convergence for the MCMC algorithm. We start with providing a brief recap of the notations and assumptions used in last lecture followed by a proof based on point-wise convergence of each matrix entry and a proof by construction.

2 Recap

- Domain : $\mathcal{X}$, $|\mathcal{X}| = N < \infty$ ($N >> 1$)
- Transition Law : $P \in \mathbb{R}^{N \times N}$, $\forall i \in [N], P^i \in \Delta_{N-1}$
- Right Stochastic Matrix : A matrix $P \in \mathbb{R}^{N \times N}$ is right stochastic if $P\mathbf{1} = \mathbf{1}$.
- Reversible : It is ensured by the condition of detailed balance, according to which for some $\pi \in \Delta_{N-1}, \forall i, j \in [N], \pi_i P_{ij} = \pi_j P_{ji}$. $\implies$ π is invariant with respect to P or $\pi P = \pi$.
- Ergodicity: An aperiodic and irreducible Markov chain is ergodic.

For clarifications on these definitions, the readers are directed to the references. (2)(1)(4)(3)

2.1 Algorithm

Algorithm 1: MCMC

Input: Transition Law P
Output: $S^0, \ldots, S^t, \ldots$
1: $\pi^0 \leftarrow INIT$
2: $S^0 \sim \pi^0$
3: **for** $t = 1, 2, \ldots,$ **do**
4: $\quad S^t \sim \pi^0 P^t$
5: **end for**

3 Proof by point-wise convergence of matrix entries

Remark 25.1. Reversibility and Ergodicity ensure that $\exists$a unique$\pi^* \in \Delta_{N-1}$ such that $\pi^* P = \pi^*$

Claim 25.1. $\pi^t = \pi^0 P^t = \mathbb{P}\left[S^t_{MCMC} = x\right] \to \pi^*$

This claim cannot be proved without using ergodicity of the transition law P. We provide a counter example for this claim.

Example 25.1. Let $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. This matrix will never converge to a stationary distribution independent of π^0 although it is reversible as it is not ergodic.

To simplify our proof, we work with $\mathbf{1}\pi^*$ and we try to prove $P^t \to \mathbf{1}\pi^*$. This results in $\forall \pi^0 \in \Delta_{N-1} \pi^0 P^t \to \pi^0 \mathbf{1}\pi^* = \pi^*$. (Since $\pi^0 \mathbf{1} = 1$)

Theorem 25.2. $P^t - \mathbf{1}\pi^* = (P - \mathbf{1}\pi^*)^t$

Proof. Proof strategy is Induction. For t = 1, the statement is always true. Assuming it to be true for $t = T$, $\implies$ $P^T - \mathbf{1}\pi^* = (P - \mathbf{1}\pi^*)^T$

$$(P - \mathbf{1}\pi^*)^{T+1} = P^{T+1} - (P\mathbf{1})\pi^* - \mathbf{1}(\pi^* P^T) + \mathbf{1}(\pi^* \mathbf{1})\pi^* = P^{T+1} - \mathbf{1}\pi^*$$

Since $P\mathbf{1} = \mathbf{1}$, $\pi^* P^T = \pi^*$ and $\pi^* \mathbf{1} = 1$. □

This claim was required to develop our notion of error as the error after t-steps is same as the error in one step amplified t times. We need a scalar form for the error, thus we would require matrix norms. The Frobenius norm won't work here as we have a vector with fixed L_1 norm, so we need to extend the Frobenius like norms to L_1 norms.

Definition 25.1. Induced Norms : For every $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$, we define induced norm on $\mathbb{R}^{d \times d}$ matrices as $\|A\|_{p \to q} \triangleq \max_{\|\mathbf{x}\|_p = 1} \|A\mathbf{x}\|_q$

We will be using the symmetric induced norm with $p = 1, q = 1$. $\|A\|_{1 \to 1} = \max_{\|\mathbf{x}\|_1 = 1} \|A\mathbf{x}\|_1$

Theorem 25.3. For $A \in \mathbb{R}^{d \times d}$, $\left\|A^t\right\|_{1 \to 1} \leq (\|A\|_{1 \to 1})^t$

This theorem finds application in the Power method and its basic proof idea is also similar.

We want to show that $\left\|P^t - \mathbf{1}\pi^*\right\|_{1\to 1} = (\|P - \mathbf{1}\pi^*\|_{1\to 1})^t \leq \epsilon$. This is equivalent to proving $\forall j \in [N], \left|(\pi^0 P^t)_j - \pi^*_j\right| \to 0$

$$\sum_{j=1}^{N} \left|(\pi^0 P^t)_j - \pi^*_j\right| = \left\|\pi^0 P^t - \pi^*\right\|_{1\to 1} \qquad = \left\|\pi^0(P^t - \mathbf{1}\pi^*)\right\|_1 \leq \epsilon \left\|\pi^0\right\|_1 \qquad \leq \epsilon$$

Thus Total Variance between $\pi^0 P^t$ and π^* is ϵ.

Definition 25.2. Total Variance between 2 distributions p and q over $\mathcal{X}$ is defined as $TV(p,q) \triangleq \sup_{S\subseteq\mathcal{X}} |p(S) - q(S)|$

We want this quantity ϵ to be less than 1.

Suppose for some $n_0 > 0$, $\|P^{n_0} - \mathbf{1}\pi^*\|_{1\to 1} = \eta < 1$.

Then, for $t \geq kn_0, \left\|P^t - \mathbf{1}\pi^*\right\|_{1\to 1} \leq (\|P^{n_0} - \mathbf{1}\pi^*\|_{1\to 1})^k \leq \eta^k$

We get a bound of $(\eta^{\frac{1}{n_0}})^t$, where n_0 plays the role of condition number. Higher is n_0, closer is $\eta^{\frac{1}{n_0}}$ to 1.

Ergodicity ensures that $\exists n$ such that $\forall i, j \in [N], P^n_{ij} > 0$

Let $\beta \leq \min_{i,j} \frac{P^{n_0}_{ij}}{\pi^*_j}$. The trick behind this assumption is $\forall i, j P^{n_0}_{ij} \geq \beta\pi^*_j$ which can be interpreted as having a healthy probability of lying in any state after running the chain n_0 times. This notion is related to the conductance of the graph or diameter of the graph with respect to the MCMC algorithm. Also, $\eta = 1 - \beta$

Notation : $P^{n_0} = Q$

$\forall i, j Q_{ij} \geq \beta\pi^*_j$

$$\begin{aligned}
\sum_j Q_{ij} = 1 &= Q_{il} + \sum_{j\neq l} Q_{ij} \\
&\geq Q_{il} + \beta \sum_{j\neq l} \pi^*_j \text{`} \\
\implies Q_{il} &\leq (1-\beta) + \beta\pi^*_l \\
\implies Q_{il} - \pi^*_l &\leq (1-\beta)(1-\pi^*_l) \leq (1-\beta)
\end{aligned}$$

We can similarly show that $Q_{il} - \pi^*_l \geq -(1-\beta)\pi^*_l \geq -(1-\beta)$ Thus, $|Q_{il} - \pi^*_l| \leq (1-\beta)$, which implies that every entry(point-wise) of matrix $Q - \mathbf{1}\pi^*$ is bounded by $(1-\beta)$

$$\implies \left\|\pi^0(Q - \mathbf{1}\pi^*)\right\|_1 \leq (1-\beta)\left\|\pi^0\right\|_1 = (1-\beta)$$

These arguments are easy to show for the discrete support case but become very involved when extended to continuous domains.

4 Proof by Construction

We have two transition laws Q which is not stationary and $\mathbf{1}\pi^*$ which is stationary. Also, with mixing constant n_0, $\forall i, j Q_{ij} \geq \beta\pi^*_j$. Then we simultaneously run the following chains -

1. Simulate 2 chains
2. First chain follows the transition law $\mathbf{1}\pi^*$

3. Second chain does the following at every step -
   (a) with probability β choose the same state as the first chain
   (b) With probability $1-\beta$, choose a state j with probability $\frac{Q_{ij}-\beta\pi^*_j}{1-\beta}$

For the second chain $\mathbb{P}[j|i] = \beta\pi^*_j + (1-\beta)\frac{Q_{ij}-\beta\pi^*_j}{1-\beta} = Q_{ij}$. The proof idea is that the 2 chains give us different states with probability $(1-\beta)$ which is ensured by design since we link the chains together with probability β. Thus, $\|Q - \mathbf{1}\pi^*\|_{1\to 1} \leq (1-\beta)$.

Such a transition law with acceptance criteria is very similar to the Metropolis Hastings transition law which has a proposal distribution with acceptance probability. $p(x,y) = q_x(y)\alpha_x(y)$ where

$$\alpha_x(y) = min\left\{1, \frac{\pi^*(y)q_y(x)}{\pi^*(x)q_x(y)}\right\}$$

There are many variants of MH algorithm with different acceptance criteria and proposal distributions. A few of these are listed below -

- Symmetric Proposal(Metropolis Algorithm) : $q_x(y) = q_y(x), \implies \alpha_x(y) = min\left(1, \frac{\pi^*(y)}{\pi^*(x)}\right)$
- Hastings Independence Proposal : $q_x(y) = q(y)\ \forall x, y$
- Random Walk : $q_x(y) = q(|x-y|) = q_y(x)$

For the Hastings Independence Proposal, if $q(y) \geq \beta\pi^*(y)\ \forall y$ (which is true if q has universal support) Then we want to show that $p(x,y) \geq \beta\pi^*(y)$

- Case 1 : $\alpha_x(y) = 1, \implies p(x,y) = q_x(y) = q(y) \geq \beta\pi^*(y)$
- Case 2: $\alpha_x(y) = \frac{\pi^*(y)}{\pi^*(x)}$
  $\implies p(x,y) = \frac{q_x(y)\pi^*(y)q_y(x)}{\pi^*(x)q_x(y)} = \frac{\pi^*(y)q(x)}{\pi^*(x)} \geq \beta\pi^*(y)$ Since $\frac{q(x)}{\pi^*(x)} \geq \beta$

We can prove that the Hastings independence proposal has a linear rate of convergence. For MH, the covergence becomes difficut to show if support is too large or not too compact.

We have discussed 2 proofs one by establishing pointwise convergence of each matrix entry and another by construction for the convergence of the MCMC algorithm. The next lectures will deal with analyses of other popular techniques in probabilistic machine learning like Langevin Dynamics.

References

[1] Ergodicity-Wiki. Ergodicity. URL `https://en.wikipedia.org/wiki/Ergodicity`. [Online; accessed 20-April-2018].

[2] Markov Chain-Wiki. Markov chain. URL `https://en.wikipedia.org/wiki/Markov_chain`. [Online; accessed 20-April-2018].

[3] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the hastings and metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 02 1996. doi: 10.1214/aos/1033066201. URL `https://doi.org/10.1214/aos/1033066201`.

[4] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0521731828, 9780521731829.