LEC 05

# MCMC

## I. Markov chains

A Markov chain is a discrete-time stochastic process $\theta_1, \theta_2, \dots, \theta_t, \dots$ that satisfies the Markov property

sequence of random variables where the index corresponds to time OR just a collection of R.Vs.

$$p(\theta_t \mid \theta_{t-1}, \dots, \theta_1) = p(\theta_t \mid \theta_{t-1}) \quad \forall t.$$

states of the chain

$$\to \theta_{t-1} \to \theta_t \to$$

A Markov chain is fully defined by

- $\pi_0(\theta)$, the initial distribution.
- $T(\theta_{t-1}, \theta_t) = p(\theta_t \mid \theta_{t-1})$, the transition kernel.

$$\hookrightarrow T(i,j) = P(\theta_t = j \mid \theta_{t-1} = i)$$

# Basic limit theorem (of Markov chains)

$\pi_n(\theta) = \int \pi_{n-1}(\theta')T(\theta',\theta)\,d\theta'$

Let $\pi_n = \pi_0 \underbrace{TT\dots T}_{n \text{ times}}$ be the probability over the states after $n$ iterations.

Let $\pi_*(\theta)$ be the stationary distribution of the Markov chain with kernel $T$, i.e. such that $\pi_* T = \pi_*$.

**Thm** $\lim_{n\to\infty} \pi_n(\theta) = \pi_*(\theta) \quad \forall \theta$

$\rightsquigarrow$ "$\pi_n$ converges to $\pi_*$"

Assumptions:

- $\pi_*$ exists
- The chain is irreducible, i.e. if any state can be reached from any other state with positive probability in a finite number of steps. $\Leftrightarrow \forall i,j \; \exists n$ s.t. $P(\theta_n = j \mid \theta_0 = i) > 0$.

- The chain is aperiodic.

A state $\theta$ is periodic with period $k$ if the number of steps to return to $\theta$ is always divisible by $k \geq 2$.

A Markov chain is aperiodic if none of its states is periodic with $k \geq 2$.

$\Leftrightarrow$ it does not make deterministic visits to a subset of the states

## Time reversibility

A Markov chain is time reversible if

$$(\theta_0, \theta_1, \dots, \theta_n) \overset{D}{=} (\theta_n, \theta_{n-1}, \dots, \theta_0)$$

"equal in distribution"

this definition implies

$\Rightarrow$ i) $(\theta_0, \theta_1) \overset{D}{=} (\theta_1, \theta_0)$

The sequence of states moving forward is equal in distribution to the sequence of states moving backward.

and ii) $(\theta_0) \overset{D}{=} (\theta_1)$

They have the same joint probability

$$\Downarrow$$

$$\pi_0 = \pi_1$$

$$= \pi_0 T \Rightarrow \pi_0 \text{ is the stationary distribution}$$

Also $(\theta_0, \theta_1) \overset{?}{=} (\theta_1, \theta_0)$

$$\Leftrightarrow P(\theta_0 = i, \theta_1 = j) = P(\theta_1 = i, \theta_0 = j) \quad \forall i, j$$

$$\Leftrightarrow P(\theta_0 = i) P(\theta_1 = j \mid \theta_0 = i) = P(\theta_0 = j) P(\theta_1 = i \mid \theta_0 = j)$$

$$\Leftrightarrow \underbrace{\pi_0(i) T(i, j) = \pi_0(j) T(j, i)}_{\text{Local balanced equation}}$$

If the local balanced equation hold for $\pi$ and $T$, then $\pi$ is the stationary distribution governed by $T$. $\to$ to be used in MH.

Time reversibility will give us a way to construct a chain that converge to a target stationary distribution.

# II. Metropolis - Hastings (MH)

MH sampling works by deriving a Markov chain $\theta_1, \theta_2, \dots$ whose stationary distribution is the target $\pi(\theta)$.

candidate

current state

① $\theta' \sim q(\theta' | \theta_t)$

proposal distribution

② $\alpha(\theta' | \theta_t) = \min \left\{ \frac{\pi(\theta')}{\pi(\theta_t)} \frac{q(\theta_t | \theta')}{q(\theta' | \theta_t)}, 1 \right\}$

acceptance ratio

③ $\theta_{t+1} = \begin{cases} \theta' & \text{with probability } \alpha(\theta' | \theta_t) \\ \theta_t & \text{otherwise} \end{cases}$

$u \sim U[0;1]$
if $u < \alpha$, accept

⚠ MH only requires the ratio of the target density $\dfrac{\pi(\theta')}{\pi(\theta_t)}$

$\Rightarrow$ We can use an unnormalised density $\pi^*(\theta) = \frac{1}{Z}\pi(\theta)$ with an unknown normaliser $\frac{1}{Z}$.

## Random walk MH (= "Metropolis algorithm")

$$\theta' = \theta_t + \varepsilon$$

$\varepsilon \rightarrow$ from a symmetric and centred distribution $g$ (e.g. $\mathcal{N}$)

In this case, $q(\theta' | \theta_t) = g(\varepsilon)$

$= \big($ $q(\theta_t | \theta') = g(-\varepsilon) = g(\varepsilon)$

$$\alpha(\theta' | \theta_t) = \min\left\{ \frac{\pi(\theta')}{\pi(\theta_t)} \frac{\cancel{q(\theta_t | \theta')}}{\cancel{q(\theta' | \theta_t)}}, 1 \right\}$$

## Why does it work?

**Thm:** The MH acceptance probability $\alpha(\theta'|\theta_t)$ for the proposal $q(\theta'|\theta_t)$ and target $\pi(\theta)$ defines a Markov chain with the transition kernel

$$T(\theta_t, \theta') = \alpha(\theta'|\theta_t)\, q(\theta'|\theta_t).$$

This chain satisfies the detailed balanced condition.

**Proof:**

$$\pi(\theta_t)\, T(\theta_t, \theta') = \pi(\theta_t)\, \alpha(\theta'|\theta_t)\, q(\theta'|\theta_t)$$

$$= \pi(\theta_t) \min\left\{ \frac{\pi(\theta')}{\pi(\theta_t)} \frac{q(\theta_t|\theta')}{q(\theta'|\theta_t)}, 1 \right\} q(\theta'|\theta_t)$$

$$= \min\left\{ \pi(\theta')\, q(\theta_t|\theta'), \pi(\theta_t)\, q(\theta'|\theta_t) \right\}$$

$$= \min\left\{ \pi(\theta_t)\, q(\theta'|\theta_t), \pi(\theta')\, q(\theta_t|\theta') \right\}$$

$$= \pi(\theta')\, \underbrace{\alpha(\theta_t|\theta')\, q(\theta_t|\theta')}_{T(\theta', \theta_t)} \qquad \square$$

# III. Bayesian inference with MCMC

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)}$$

$$= \int p(x|\theta)\, p(\theta)\, d\theta$$

Difficult to evaluate!

$$= \frac{1}{Z(x)}\, p(x|\theta)\, p(\theta) = \pi^*(\theta)$$

$\Rightarrow$ Use MCMC for posterior sampling

$$\log \pi^*(\theta) = \log p(x|\theta) + \log p(\theta) + C$$

Diagnostics:

Gelman-Rubin diagnostic

* Convergence checking based on $m$ chains from widely diverse starting points.

→ Eventually, all chains should look like the stationary distribution and "look" all the same.

Assuming $\theta_{ij}$ $(i=1..n,\ j=1...m)$

- Between - chains variance

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\theta_{:j}-\theta_{::}\right)^2$$ (variance of the mean)

where $$\theta_{:j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}$$ (chain mean)

As $n\to\infty$, $B \searrow 0$.

$$\theta_{::} = \frac{1}{m}\sum_{j=1}^{m}\theta_{:j}$$ (grand mean)

- Within - chain variance

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2$$ (mean of chain variances)

where $$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_{ij}-\theta_{:j}\right)^2$$ (within chain variance)

$\rightsquigarrow \hat{R} = \left(\frac{n-1}{n}W + \frac{1}{n}B\right)/W$

$$\Rightarrow \quad \mathbb{V}[\theta] = \frac{n-1}{n}W + \frac{1}{n}B$$

overestimates $\pi$'s variance

Gelman-Rubin statistic

$$\Rightarrow \hat{R} = \frac{\mathbb{V}[\theta]}{W} \geq 1$$

$\hat{R}$ approaches 1 as $n$ increases

underestimate $\pi$'s variance

$\longrightarrow$ Stop when $\hat{R} < 1.1$

* Effective sample size

$$\rho_k = \frac{\mathbb{E}\left[(\theta_i - \mu)(\theta_{i+k} - \mu)\right]}{\sigma^2}$$

over the time index i

Number of samples

$$ESS = \frac{M}{\underbrace{1 + 2\sum_{k=1}^{\infty} \rho_k}_{= \sum \text{autocorrelation}}}$$

$\Rightarrow$ Uncorrelated samples $\rightarrow \rho_k = 0$

and $ESS = M$

$\Rightarrow$ Correlated samples $\rightarrow$ small ESS