Foundations of MCMC

Alexander Fengler
Applied Statistics Reading Group

November 5, 2018

The purpose of these notes is to provide a self-contained introduction to the principles behind Markov Chain Monte Carlo samplers. We confine ourselves to the realm of finite state Markov Chains to provide the building blocks, however the ideas and intuitions then carry over to (what is more commonly the case) the infinite state scenarios.

In the beginning section I will try to provide a road-map of the arguments ahead to elucidate the motivation behind the concepts introduced.

1 The Goal

Markov Chain Monte Carlo methods are a popular technique, employed to gather samples from complicated probability distributions, which are defined up to their normalization constant. Especially in the context of Bayesian Inference, where posterior distributions over parameters are usually defined up to a proportionality constant, MCMC is a standard procedure to produce posterior samples, which can them be used for purposes of inference and prediction.

To define the problem we want to solve more formally, let's begin with our beloved Bayes Rule.

Let us assume we have a parameter $\theta \in \Theta \subset \mathbb{N}$, that lives in a finite set of possible parameter values, that defines a statistical model $(\mathbf{X}, (\mathbb{P}_\theta)_{\theta \in \Theta})$, where $\mathbf{X}$ stands for the **Evidence**, that a specific Data set provides.

This statistical model, given θ and $\mathbf{X}$, provides us with a **likelihood function**, $\pi_l(\mathbf{X}|\theta)$.

Moreover, we need a **prior distribution**, $\pi(\theta)$ for θ. Then by Bayes rule, we know that,

$$\pi_p(\theta|\mathbf{X}) \propto \pi_l(\mathbf{X}|\theta)\pi(\theta)$$

Note that we **do not** have access to the **normalization constant**,

$$\pi(\mathbf{X}) = \int_{\theta \in \Theta} \pi_l(\mathbf{X}|\theta)\pi(\theta)$$

This is the common scenario, since usually evaluation of such an integral will be computationally very expensive. Importantly, if we could compute $\pi(\mathbf{X})$, we would thereafter be able to compute posterior **predictive distributions** and **mean, mode, etc.** of the **posterior**, directly.

It is precisely because evaluation of this integral is difficult, that we resort to **sampling from the posterior** immediately. As we will see, this allows us to **evaluate the posterior where it matters**, instead of attempting to span the whole parameter space indiscriminately with our function evaluations (which as a rough approximation is what evaluation of an integral boils down to).

Lastly, note that if we had access to the **normalization constant**, we would be able to use more efficient procedures to get samples from the posterior afterwards.

2 Properties of Stochastic Matrices

To begin the journey, let's first develop some understanding of what will be our main workhorse later on: *right stochastic matrices*.

Definition: Right Stochastic Matrix. A right stochastic matrix $\mathbf{P} \in [0,1]^{n \times n}$ is defined by the following two properties.

$$(1) \sum_{j=1}^{n} P_{ij} = 1 \quad \forall i \in \{1, 2, ..., n\}$$
$$(2) P_{ij} \geq 0 \quad \forall (i,j)$$

Let's see some general properties of such matrices.

Proposition: 1 Every *RSM*, has *right eigenvector*, $\mathbb{1}$ with *eigenvalue* $\lambda = 1$.

Proof. The proposition states that $\mathbf{P}\mathbb{1} = \mathbb{1}$. This follows directly from the properties by which a *RSM* is defined. The matrix equation gives us back a vector of *row sums* of $\mathbf{P}$, which by property (1), comes out to $\mathbb{1}$. □

Proposition: 2 If we multiply any *RSM* $\mathbf{P}$, with any other *RSM* $\mathbf{P}^{'}$, we get back a *right stochastic matrix*.

Proof. We know from the previous proposition that,

$$\begin{aligned} \mathbf{P}\mathbf{P}^{'}\mathbb{1} &= \mathbf{P}(\mathbf{P}^{'}\mathbb{1}) \\ &= \mathbf{P}\mathbb{1} \\ &= \mathbb{1}. \end{aligned}$$

Since $\mathbf{P}_{ij} \geq 0 \ \ \forall (i,j)$, and $\mathbf{P}^{'}_{ij} \geq 0 \ \ \forall (i,j)$, we also have $(\mathbf{P}\mathbf{P}^{'}) \geq 0 \ \ \forall (i,j)$. Together this implies that $(\mathbf{P}\mathbf{P}^{'})$ is a *RSM*. □

Proposition 2 tell us that we can multiply an arbitrary amount of of stochastic matrices with each other and we are assured that we get a valid stochastic matrix back.

Proposition 1, is indirectly important to us. In fact we do not directly care about the *right eigenvector*, $\mathbb{1}$, but we care about the implication of it's assured existence.

Proposition: 3 Given a *RSM* $\mathbf{P}$, $\exists\pi$, s.t. π is a *left eigenvector* of $\mathbf{P}$ with *eigenvalue*, $\lambda = 1$

Proof. Note first that a *left eigenvector* of $\mathbf{P}$ is a *right eigenvector* of $\mathbf{P^T}$. So we will show that $\exists\pi$, s.t. π is a *right eigenvector* of $\mathbf{P^T}$ with *eigenvalue*, $\lambda = 1$.

For this proof we take for granted the result that for any matrix $\mathbf{A}$, we know $\mathbf{A}$ is *similar* to $\mathbf{A}^T$. In other words, $\exists$ a *non-singular (invertible)* matrix $\mathbf{B}$ s.t.,

$$\begin{aligned}\mathbf{AB} &= \mathbf{BA}^T \\ \mathbf{A} &= \mathbf{BA}^T\mathbf{B}^{-1}\end{aligned}$$

Note that this implies, $det(\mathbf{A}) = det(\mathbf{A^T})$, since

$$\begin{aligned}det(\mathbf{A}) &= det(\mathbf{BA^TB^{-1}}) \\ &= det(\mathbf{B})det(\mathbf{A^T})det(\mathbf{B^{-1}}) \\ &= det(\mathbf{A^T})\end{aligned}$$

Using $det(\mathbf{B}) = \frac{1}{det(\mathbf{B^{-1}})}$.

Now using those results, we get for any *RSM* $\mathbf{P^T}$,

$$\begin{aligned}det(\mathbf{P^T} - \lambda\mathbf{I}) &= det((\mathbf{P} - \lambda\mathbf{I})^T) \\ &= det(\mathbf{P} - \lambda\mathbf{I})\end{aligned}$$

This implies that the *eigenvalues* of $\mathbf{P}$ and $\mathbf{P^T}$ are equivalent, and therefore, since we know that $\mathbb{1}$ is an *right eigenvector* with *eigenvalue* 1, there must exist such a *left eigenvector* π as requested. □

Now what will be important to us as well, is that for any *RSM*, $1 = max\{|\lambda_1|, ..., |\lambda_n|\}$. In other words the *spectrum* $\rho(\mathbf{P}) = max\{|lambda_1|, .., |lambda_n|\}$ of any *RSM* is equal to one.

We already know that we can find π, such that $\pi\mathbf{P} = \pi$, so we know at this point that $\rho(P) \geq 1$ for any *RSM*. To show that also $\rho(P) \leq 1$, we employ the *Gershgorin Circle Theorem*.

Theorem: Gershgorin. Define for any square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, for $i \in \{1, 2, ..., n\}$,

$$R_i = \sum_{j\neq i} |a_{ij}|$$

$$D(a_{ii}, R_i) \subset \mathbb{C}$$

Then every *eigenvalue* of A lies within $D(a_{ii}, R_i)$ (*Gershgorin Disk*) for some i.

Proof. Take λ any *eigenvalue* of A and $\mathbf{x}$ the corresponding *eigenvector*.

We choose $\mathbf{x}$ s.t. $x_i = 1$ and $x_j \leq 1 \;\; \forall j \neq i$ (can always do so).

Now since, $\mathbf{Ax} = \lambda\mathbf{x}$, we know that

$$\sum_j a_{ij}x_j = \sum_{j\neq i} a_{ij}x_j + a_{ii} = \lambda$$

Now since $x_j \leq 1 \ \ \forall j \neq i$, we get,

$$\begin{aligned} |\lambda - a_{ii}| &= |\sum_{j\neq i} a_{ij}x_j| \\ &\leq \sum_{j\neq i} |a_{ij}||x_j| \\ &\leq \sum_{j\neq i} |a_{ij}| = R_i \end{aligned}$$

□

Corollary: . Eigenvalues must also lie within at least one of the *Gershgorin Discs* to the columns of $\mathbf{A}$

Proof. We just need to apply the **Gershgorin Theorem** to $\mathbf{A^T}$. □

Now let's see what this theorem tells us for *right stochastic matrices*.

Proposition: 4 For any *right stochastic matrix* $\mathbf{P}$, $\rho(\mathbf{P}) = 1$.

Proof. The difference between a general matrix square $\mathbf{A}$ and a *RSM* $\mathbf{P}$, is that we can make R_i specific. We have $R_i = \sum_{j\neq i} |a_{ij}|$. However we know that $a_{ij} \in \mathcal{R}^+$ so $R_i = \sum_{j\neq i} a_{ij}$. But since $\forall i$, $\sum_j a_{ij} = 1$, therefore $R_i = 1 - a_{ii}$.

But this implies that $D(a_{ii}, R_i) = D(a_{ii}, 1 - a_{ii})$. Hence, all *Gershgorin Discs* lie within the *circle of unity*. Since by the **Gershgorin Theorem** all *eigenvalues* need to lie within at least one *Gershgorin Disc*, this completes the proof. □

At this point we have shown that every *right stochastic matrix* has at least one *left eigenvector* with *eigenvalue* 1, and no *eigenvector* with *eigenvalue* greater than 1.

It will become clear later, why this is useful.

To avoid confusion later, let us define an alternative name for our *left eigenvector* with *eigenvalue* 1.

Definition: Invariant Distribution. We call a *left eigenvector* with *eigenvalue* 1, of a *right stochastic matrix*, an **invariant distribution**.

3 Ergodic Markov Chains

Let us start in general by defining a stochasting process in discrete time, with finite state space.

Definition: Stochastic Process (Discrete Time, Finite State Space). A discrete time, finite state stochastic process is defined as an indexed (ordered) set of *random variables*, $(X, t \in \mathcal{T})$, where $X \in \mathcal{X}$, s.t. $|\mathcal{X}| \leq N$ for some $\mathbb{N} \in \mathcal{N}$, and the *index set* set $\mathcal{T}$ is countable.

Usually we look at $\mathcal{T}$ as *time index*, this is however not necessary. Now in principle, the random variables in $(X, t \in \mathcal{T})$, could stand in any relation to each other. One particular *relation* that is of interest for us, is what is called a *Discrete Time Markov Chain*.

Definition: Discrete Time Markov Chain. A discrete-time Markov Chain is a sequence of random variables $X_1, X_2, ...$ with the *Markov Property*. The probability of moving to the next state given the current state, depends only on the current state and is independent of any other past states.

$$\mathbb{P}(X_{t+1}|X_1 = x_1, ..., X_t = x_t) = \mathbb{P}(X_{t+1} = x|X_t = x_t)$$

Moreover, we impose the restriction of **time-homogeneity**, namely that

$$\mathbb{P}(X_{t+1}|X_t = y) = \mathbb{P}(X_t = x|X_{t-1} = y), \quad \forall t \in \mathcal{T}$$

Now, we represent the probabilities $\mathbb{P}(X_t = x_i|X_{t-1} = x_j)$, $\forall i$, $\forall j$, $\forall t \in \mathcal{T}$ as a single *state-transition matrix* $\mathbf{P}$, where $\mathbf{P}$ is a *RSM*.

Our goal is to use such *Markov Chains*, to perform *Markov Chain Monte Carlo*. Two more restrictions on our *Markov Chain* are needed before we can move towards this purpose. These are provided in the following two definitions.

Definition: Irreducible Markov Chain. A *Markov Chain* is said to be *irreducible*, if it is possible to get from any state to any other state.

Definition: Period of a state. A state i has *period* k if any return to state i must occur in multiples of k steps. Formally,

$$k = gcd\{n > 0 : \mathbb{P}(X_n = i|X_0 = i) > 0\}$$

If $k = 1$ for state i, then we call state i **aperiodic**. If all states $i \in \mathcal{X}$, are *aperiodic*, we say that our *Markov Chain*, is an *aperiodic chain*.

This brings us to the type of *Markov Chain* that we finally consider in the following. The *Ergodic Chain*.

Definition: Ergodic Chain. An *aperiodic*, *irreducible Markov Chain*, is called *Ergodic*.

Importantly, given a *state transition matrix* $\mathbf{P}$, we can check if it defines an *Ergodic Chain*, by checking whether $\mathbf{P}$ is **quasi-positive**.

Definition: Quasi-positive Matrix. A non-negative matrix $\mathbf{A}$ is *quasi-positive*, if $\exists n_0 \in \mathbb{N}$, s.t. $\mathbf{A^{n_0}} > 0$.

Now let's see what purchase we get from considering such *Ergodic Chains*.

Theorem: Perron-Frobenius, Ergodic RSM Version. Take any *right stochastic matrix* $\mathbf{P}$, which defines an *Ergodic Chain*. Then, $\exists \mathbf{x} > 0$ an *eigenvector* of $\mathbf{P}$ with corresponding *eigenvalue* $\lambda_1 = 1$, s.t.

$$\forall \;\; i(\lambda_1 > |\lambda_i|)$$

The **Perron-Frobenius Theorem** implies that the **eigenvector x**, corresponding to the *eigenvalue*, $\lambda = 1$, is *positive* and *unique* (since it implies *geometric multiplicity* of 1 for $\lambda_1 = 1$). Now this purchases us the next, theorem.

Theorem: Convergence of Ergodic Markov Chains. If $\mathbf{P}$ is the *state transition matrix* of an ergodic markov chain, then the powers of $\mathbf{P}$ converge.

Proof. To proof this theorem, first note that by existence of the *Jordan Normal Form of a square Matrix*, we can always find a complex, non-singular matrix $\mathbf{S}$, s.t.,

$$\mathbf{U} = \mathbf{S^{-1}PS}$$

where $\mathbf{U}$ is of the form (note that we know that the *algebraic multiplicity of the leading eigenvalue* is 1,

$$\mathbf{U} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & ... & 0 \\ 0 & \lambda_2 & 1 & 0 & ... & 0 \\ 0 & 0 & \lambda_2 & 0 & ... & 0 \\ 0 & 0 & 0 & \lambda_3 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & \lambda_p \end{bmatrix}$$

Moreover, we know that $\lambda_i \leq 1 \;\; \forall i$. Now this implies that,

$$lim_{n\to\infty}\mathbf{U^n} = \mathbf{N}$$

, where,

$$\mathbf{N} = \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & 0 \end{bmatrix}$$

Now since $\mathbf{N} = \mathbf{S^{-1}P^{\infty}S}$, we conclude that the powers of $\mathbf{P}$ converge. □

Corollary: Rows of $\mathbf{P^{\infty}}$. If $\mathbf{P}$ is a *right stochastic matrix* that defines an *ergodic chain*, then all rows of $\mathbf{P^{\infty}}$ are equal to the unique invariant distribution $\mathbf{P}$

Proof. First note that, by definition of a *right stochastic matrix*, every row of $\mathbf{P}$ is a valid distribution. Moreover, we by **Proposition 2**, we know that $\forall n \in \mathbb{N}$, $\mathbf{P^n}$ is a *right stochastic matrix*.

Now first we show that for any *right stochastic matrix* $\mathbf{Q}$, for which the powers $\mathbf{Q^n}$ converge, we have that the rows of $\mathbf{Q}^\infty$ are equal to *invariant distributions* of $\mathbf{Q}$. For this purpose we need to show that $\mathbf{Q}^\infty\mathbf{Q} = \mathbf{Q}^\infty$. The following chain of equalities provides this,

$$\begin{aligned}\mathbf{Q}^\infty\mathbf{Q} &= (\lim_{n\to\infty} \mathbf{Q^n})\mathbf{Q} \\ &= \lim_{n\to\infty} \mathbf{Q^{n+1}} \\ &= \lim_{n\to\infty} \mathbf{Q^n} \\ &= \mathbf{Q}^\infty\end{aligned}$$

We already concluded that for any *right stochastic matrix* $\mathbf{P}$, which defines an *Ergodic Chain*, the *invariant distribution* is *unique*. Hence, all rows of $\mathbf{P}^\infty$, must be the same, which concludes the proof. □

Now we are the position to show the most important result in this section.

Theorem: Convergence from any distribution. If $\mathbf{P}$, is the state transition matrix for an *ergodic markov chain*, then from any initial distribution q we have, $\lim_{t\to\infty} q\mathbf{P^t} = s$, where s is the unique invariant distribution of $\mathbf{P}$

Proof. Given what we know about *ergodic chains* at this point, this result is easy to show. We have seen that $\mathbf{P}^\infty$ is of the form,

$$\mathbf{P}^\infty = \begin{bmatrix} s_1 & s_2 & \dots & s_n \\ s_1 & s_2 & \dots & s_n \\ . & . & . & . \\ . & . & . & . \\ s_1 & s_2 & \dots & s_n \end{bmatrix}$$

where s is the *invariant distribution* of $\mathbf{P}$. Now since q is a valid distribution, $q\mathbf{P}^\infty$, is a weighted sum over the rows of $\mathbf{P}^\infty$. However this must return s, for any q. □

What follows is, from an operational viewpoint the result we are exploiting in the context of *Markov Chain Monte Carlo* methods.

Theorem: Ergodic Theorem. If $\mathbf{P}$ defines an *Ergodic Chain* with *invariant distribution* π, then

$$\frac{\{t \leq n : X_t = i\}}{n} \longrightarrow \pi_i \quad \text{as} \quad n \to \infty$$

Let's take stock of what we know at this point. If we have a *right stochastic matrix* $\mathbf{P}$, which defines the *state transition probabilities* of an *Ergodic Markov Chain*, then we know that $\mathbf{P}$ has a unique *invariant distribution*, which it reaches from any starting distribution. Moreover, the **Ergodic Theorem** tells us that a single sample path from an *Ergodic Markov Chain* yields us the invariant distribution in the limit, by virtue of it's site visitation history.

4 Markov Chain Monte Carlo

The idea behind *Markov Chain Monte Carlo* methods is to generate a sample path from an *Ergodic Markov Chain*, which serve as samples from a distribution of interest. In our case this distribution of interest is the *posterior distribution over parameters* in a Bayesian inference problem. So what we need to do is to come up with an *Ergodic Markov Chain* that has as it's unique invariant distribution our distribution of interest. In other words, we need to come up with a *state transition probability matrix*, $\mathbf{P}$, s.t.

$$\pi_p \mathbf{P} = \pi_p$$

where π_p is our posterior distribution of interest.
In principle, we only need to assure *Global Balance*.

Definition: Global Balance. At the *invariant distribution*, the flow of probability mass into and out of each state is balanced:

$$\sum_{i=1}^{N} \pi_i \mathbf{P}_{ij} = \pi_j = \sum_{j=1}^{N} \pi_j \mathbf{P}_{ij}$$

This states nothing other than the fact that π_p should be an *eigenvector* of $\mathbf{P}$ with *eigenvalue* equal to 1.

However the validity of most *Markov Chain Monte Carlo* algorithms is justified by a more stringent (buy supposedly easier to prove) condition, **Detailed Balance**.

Definition: Detailed Balance / Reversibility. At the *invariant distribution*, the flow of probability mass between each pair of states is balanced.

$$\pi_i \mathbf{P}_{ij} = \pi_j \mathbf{P}_{ji} \quad \forall i, j$$

Proposition: 5 Detailed balance implies that *pi* is the *invariant distribution*

Proof. Under detailed balance we have, $\forall i$

$$\sum_{j=1}^{N} \pi_i \mathbf{P}_{ij} = \sum_{j=1}^{N} \pi_j \mathbf{P}_{ji}$$

$$\pi_i = \sum_{j=1}^{N} \pi_j \mathbf{P}_{ji}$$

But this is just the definition of an *invariant distribution*. □

Now we can put all the ingredients we gathered together in the following theorem.

Theorem: Detailed Balance Theorem. Take $\mathbf{P}$ to define an *Ergodic Markov Chain*, and take π s.t., $\mathbf{P}_{ij}\pi_i = \mathbf{P}_{ji}\pi_j \quad \forall i, j$. Then, regardless of where the chain starts, we get,

$$\mathbb{P}(X_t = i) \approx \pi_i \quad \text{for } t \text{ large enough}$$

5 Metropolis Hastings Algorithm

We are now in a position to appreciate the justification for any *Metropolis Hastings type sampler*.

Consider the following *setup*:

1. A discrete, finite state space $\mathcal{X} = \{s_1, ..., s_n\}$
2. A *target distribution* π, s.t. $\forall i \;\; \pi(s_i) = \pi_i \geq 0$ (up to normalization constant).

We want to find a *state transition matrix* $\mathbf{P}$, which defines an *Ergodic Markov Chain*, and has as its *unique invariant distribution*, π_p.

Algorithm 1 Metropolis Hastings Algorithm

We have,
A target distribution π (specified up to proportionality constant)
A finite state space $\mathcal{X}$
A proposal distribution $\mathbf{Q_{i,j}} \;\; \forall i, j$ (specifies probability of moving from i to j
An initial state $X_0 = i$

for $t > 0$ **do**
(1) Propose state $X_t \sim Q_{X_{t-1},\cdot}$
(2) Compute acceptance probability,

$$\alpha(X_t, X_{t-1}) = \min\left(1, \frac{\mathbf{Q_{X_t, X_{t-1}}} \cdot \pi_{X_{t-1}}}{\mathbf{Q_{X_{t-1}, X_t}} \cdot \pi_{X_t}}\right)$$

(3) Accept X_t with probability $\alpha(X_t, X_{t-1})$, otherwise $X_t = X_{t-1}$
(4) Return to (1)

Let us see how this algorithm satisfies in fact the *Detailed Balance Condition*.

Theorem: MH satisfies detailed balance. The *metropolis hastings algorithm* satisfies the detailed balance condition.

Proof. Consider the *transition probability matrix* $\mathbf{P}$, associated with the *metropolis hasting algorithm*. We can define it as follows,

$$P_{ij} = \alpha(i,j)\mathbf{Q_{ij}} + \left(\sum_{l=1}^{N}(1 - \alpha(i,l))\mathbf{Q_{i,l}}\right)\mathbb{1}_{i=j}$$

To transition from state i to state j, where $i \neq j$, we need to *propose* j from i (using $\mathbf{Q_{ij}}$), and *accept* j when proposed from i (using $\alpha(i,j)$). So,

$$P_{ij} = \alpha(i,j)\mathbf{Q_{ij}} \;\; i \neq j$$

I we have $i = j$, then additionally *rejecting* the proposal when proposing any state will lead us to remain at i. So,

$$P_{ii} = \alpha(i,i)\mathbf{Q_{ii}} + \sum_{l=1}^{N}(1-\alpha(i,l))\mathbf{Q_{i,l}}$$
$$= \mathbf{Q_{ii}} + \sum_{l=1}^{N}(1-\alpha(i,l))\mathbf{Q_{i,l}}$$

Since $\alpha(i,i) = 1$.

Assuming that $\mathbf{Q} > 0$, and therefore $\mathbf{P} > 0$, we know that our *state transition matrix* defines an *ergodic markov chain*. It remains to be shown, that the *invariant distribution* of $\mathbf{P}$ is π_p. For this purpose we will show that it satisfies the *detailed balance condition*.

$$\mathbf{P}_{ij}\pi_i = \mathbf{P}_{ji}\pi_j \quad \forall i,j$$

If $i = j$ detailed balance holds trivially. Lets consider the case where $i \neq j$. First, we note that if $\alpha(i,j) \leq 1$, then $\alpha(j,i) = 1$. Now lets assume that $\alpha(i,j) \leq 1$. Now,

$$\begin{aligned}\pi_i\mathbf{P}_{ij} &= \frac{\mathbf{Q}_{ji}\cdot\pi_j}{\mathbf{Q}_{ij}\cdot\pi_i}\cdot\mathbf{Q}_{ji}\cdot\pi_i \\ &= \mathbf{Q}_{ji}\cdot\pi_i \\ \pi_j\mathbf{P}_{ji} &= \alpha(j,i)\cdot\mathbf{Q}_{ji}\cdot\pi_j \\ &= \mathbf{Q}_{ji}\cdot\pi_j\end{aligned}$$

Note that the case where $\alpha(j,i) \leq 1$ is symmetrical. We can conclude that detailed balance holds for the *metropolis hastings algorithm*. □

Remark: Note that it is enough to specify the *target distribution* only up to the normalization constant. In the metropolis hasting algorithm the *target distribution* shows up only in the computation of the *acceptance ratio*, where it appears in the *numerator* as well as the *denominator*, and therefore cancels out. Moreover note that it also does not mess with any *detailed balance* calculations, since it would show up on the *LHS* and the *RHS*, and therefore can be cancelled too.