The Bayesian view of knowledge representation, inference and learning

P.Antal

antal@mit.bme.hu

BME

From ”rational" preferences to probabilities: ”as if" I.

1. Definition. *A decision problem is defined by the elements $\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq$, where:*

(i) $\mathcal{E}$ *is an algebra of events,* E_j*;*

(ii) $\mathcal{C}$ *is a set of possible consequences,* c_j*;*

(iii) $\mathcal{A}$ *is a set of possible acts, which are mapping of partitions of the events to consequences;*

(iv) $\leq$ *is a binary preference relation between some of the elements of* $\mathcal{A}$*.*

With further ”rational" assumptions on comparability, transitivity, consistency and quantification the following suggestive result can be derived.

1. Proposition. *Given an uncertainty relation* $\leq$*, there exists a unique real number P(E) for each event E (defined as the probability of E) that they are compatible with* $\leq$ *(i.e.* $E \leq F; iff; P(E) \leq P(F)$*) and they form a finitely additive probability measure.*

Consequently, $P(A|\xi)$ denotes the subjective/personal beliefs in a given space-time-information context ξ vs. the ”frequentist" interpretation that $P(A) \triangleq \lim_{N\to\infty} N_A/N$.

Another axiomatic derivation: Cox-Jaynes axioms

From preferences to utilities: ”as if" II.

The parallel result for the existence and uniqueness of utilities (or losses) is stated only for decision problems with bounded consequences.

2. Proposition. *For any decision problem* $\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq$ *with bounded consequences* $c_* < c^*$,

(i) *for all* c, $u(c|c_*, c^*)$ *exists and unique;*

(ii) *the value of* $u(c|c_*, c^*)$ *is unaffected by the assumed occurrence of an event G;*

(iii) $0 = u(c_*|c_*, c^*) \leq u(c|c_*, c^*) \leq u(c^*|c_*, c^*) = 1.$

From exchangeability to parameters and "i.i.d." : "as if" III.

3. Proposition. *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random quantities with probability measure P, that is for any n and permutation $\pi(1), \ldots, \pi(n)$ the joint mass function of P $p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$, there exists a distribution function $\mathcal{Q}$ such that $p(x_1, \ldots, x_n)$ has the form*

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1 - \theta^{1-x_i}) d\mathcal{Q}(\theta),$$

where,

$$\mathcal{Q}(\theta) = \lim_{n \to \infty} P[y_n/n \leq \theta],$$

with $y_n = x_1 + \cdots + x_n$, and $\theta = \lim_{n \to \infty} y_n/n$.

The Bayesian statistical framework

1. Specify a joint distribution $p(x, \theta)$ over the observable quantity x and parameter θ having equal status by specifying $p(\theta)$ the prior distribution or prior, the $p(x|\theta)$ is the sampling distribution that also defines the likelihood and the likelihood function $\mathcal{L}(\theta; x)$ (the discrete model parameter is denoted with $\mathcal{M}_k$).
2. Perform a prior predictive inference

$$p(x) = \int p(x|\theta)p(\theta)d\theta \; or \; p(x) = \sum_k p(\mathcal{M}_k) \int p(x|\mathcal{M}_k) \tag{1}$$

or a posterior predictive inference after observing the data set D as

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta \; or \; p(x|D) = \sum_k p(x|\mathcal{M}_k)p(\mathcal{M}_k|D) \tag{2}$$

3. Perform a parametric inference by the Bayes rule

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \propto p(x|\theta)p(\theta) \; or \; p(\mathcal{M}_k|x) =\propto p(x|\mathcal{M}_k)p(\mathcal{M}_k) \tag{3}$$

On hierarchic priors (”Bayesian knowledge representation")

In an idealistic Bayesian approach the family of the included models in $p(\theta)$ and $p(x|\theta)$ should be as broad as necessary for expressing beliefs in any potentially relevant model. However three issues have to be considered: the potential violation of a scientific principle of *Ockham's razor* , the computational difficulties to cope with such large class of models and the pragmatic aspects of specifying a priori beliefs at the extreme for all the computable distributions.
A frequently occuring form in practice, that the specification usually achieved in the structured specification of the relevant model structures $\mathcal{S}_k$ or $\mathcal{M}_k$ and parameters θ_k. Correspondingly the a priori belief $p(\theta_k, \mathcal{M}_k)$ in a given model with structure k and parameters θ_k^i is expressed as a product

$$p(\theta_k, \mathcal{M}_k) = p(\mathcal{M}_k)p(\theta_k^i|\mathcal{M}_k) \tag{4}$$

Predictive inference (”Bayesian inference")

The specification of the a priori beliefs over relevant models allows us to perform (prior) predictive inferences over the observable quantity x

$$p(x) = \sum p(\mathcal{M}_k) \int p(x|\theta_k) p(\theta_k|\mathcal{M}_k) d\theta_k \tag{5}$$

(6)

The operation of integration and/or summation over models and/or their parameterization implements marginalization and termed in this context as Bayesian model averaging . We can write the posterior predictive distribution conditioned on the data set D as

$$p(x|D) = \sum p(\mathcal{M}_k|D) \int p(x|\theta_k) p(\theta_k|D, \mathcal{M}_k) d\theta_k \approx p(x|D, \mathcal{M}_k^{MAP}) \tag{7}$$

in which $\mathcal{M}_k^{MAP} = argmax_k p(\mathcal{M}_k|D)$ is called the maximum a posteriori (MAP) model.

Parametric inference (”Bayesian learning")

In the discrete case the posterior of the model $p(\mathcal{M}_k|D)$ is given by

$$p(\mathcal{M}_k|D) = \frac{p(D|\mathcal{M}_k)p(\mathcal{M}_k)}{p(D)} \tag{8}$$

where the marginal model likelihood or evidence for $\mathcal{M}_k$ is

$$p(D|\mathcal{M}_k) = \int p(D|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k \tag{9}$$

and the marginal data likelihood

$$p(D) = \sum_k p(D|\mathcal{M}_k)p(\mathcal{M}_k) \tag{10}$$

The *Bayes factor* shows the change of the ratio of prior belief to the ratio of the posteriors, i.e. the ratios of marginal likelihoods of models $\mathcal{M}_i$ and $calM_j$

2. Definition.

$$Bayesfactor(\mathcal{M}_i, \mathcal{M}_j) = \frac{p(D|\mathcal{M}_i)}{p(D|\mathcal{M}_j)} = \frac{p(\mathcal{M}_j)}{p(\mathcal{M}_i)} \frac{p(\mathcal{M}_i|D)}{p(\mathcal{M}_j|D)} \tag{11}$$

Bayesian decision theory I.

Frequently, the full report of the posterior over observable quantity or model or model parameters is not adequate.
If only a value $\hat{(x)}$ of the observable quantity can be reported (interpreted as a decision) whose utility is specified by a utility or *loss function* $L(x, \hat{x})$, then the optimal decision x^* based on the posterior predictive distribution is

$$x^* = argmin_{\hat{x}} \int L(x, \hat{x}) p(x|D) dx \tag{12}$$

A frequent choice for the loss function both in the case of inference about observable quantities and parameters are the 0-1 loss L_0, the absolute loss L_1, the quadratic loss L_2 and if the reported values be interpreted as discrete probability distribution the logarithmic loss, which in this case is the *Kullback-Leibler* (semi)distance KL(). The optimal values for various loss functions are as follows

$$L_0(x, \hat{x}) = I\{x \neq \hat{x}\}(mode) \tag{13}$$

$$L_1(x, \hat{x}) = |x - \hat{x}|(median) \tag{14}$$

$$L_2(x, \hat{x}) = (x - \hat{x})^2(mean) \tag{15}$$

$$KL(\underline{x}||\underline{\hat{x}}) = \sum_i x_i \log(x_i/\hat{x}_i) \tag{16}$$

Bayesian decision theory II.

In the case of parameter estimation with loss function $L(\theta, \hat{\theta})$ and observation x, the optimal estimation $\hat{\theta}$ minimizes the *posterior expected loss*

$$\varrho(p(\theta), \hat{\theta}|x) = \int L(\theta, \hat{\theta}) p(\theta|x) d\theta. \tag{17}$$

If this property holds for every observation x for a given decision rule (estimator) $\delta(x)$ (from the space of observations to the decision space of the parameters), then the estimator $\delta(x)$ minimizes the *integrated risk*

$$r(p(\theta), \delta) = \int \int L(\theta, \delta(x)) p(x|\theta) dx p(\theta) d\theta. \tag{18}$$

and it is called *Bayes estimator* and the corresponding value is the *Bayes risk* . In the context of predicting a binary class label Y after observing X, the optimal decision function (based on the posterior $p(Y|X)$) is called *Bayes decision* and the Bayes risk with 0-1 loss, i.e. the probability of missclassification, is called the *Bayes error* .

Beta distribution

3. Definition. *A family $\mathcal{F}$ of prior distributions $p(\theta)$ is said to be conjugate for a class of sampling distributions $p(x|\theta)$, if the posteriors $p(\theta|x)$ also belongs to $\mathcal{F}$.*

1. Example. *Assume that x denotes the sum of 1s of n independent and identically distributed (i.i.d.) Bernoulli trials, that is we assume a binomial sampling distribution. If the prior is specified using a Beta distribution, the posterior remains a Beta distribution with updated parameters.*

$$p(x|\theta) = Bin(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{19}$$

$$p(\theta) = Beta(\alpha,\beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \; where \; c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{20}$$

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = c'\theta^{\alpha-1+x}(1-\theta)^{\beta-1+n-x} = Beta(\alpha+x,\beta+n-x)$$

In general a conjugate prior is updated to posterior using only an appropriate statistics of the observations to update its parametrization. It shows that the parameters frequently has an intuitive interpretation based on observations, that is in the prior specification the parameters corresponds to real or virtual past observations.

The Dirichlet distribution

2. Example. *Assume that the observed sequence* $D_n = \{X_i; i = 1, 2 \ldots, n\}$ *contains i.i.d. multinomial samples with* L *discrete values. The prior is a Dirichlet prior with hyperparameters* $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_L$ *and* $\alpha_\cdot = \sum_i \alpha_i$.

$$p(\theta) = Di(\boldsymbol{\alpha}) = c \prod_i \theta^{\alpha_i - 1} \; where \; c = \frac{\Gamma(\alpha_\cdot)}{\prod_i \Gamma(\alpha_i)} \tag{21}$$

Dirichlet distribution II.

It is conjugate for multinomial sampling (see Example **??**), so the posterior predictive distributions of the defined Bayesian forecasting system are the updated Dirichlet with hyperparameters $\boldsymbol{\alpha}_j$ at step j and the posterior prediction for x_j (i.e. the marginal posterior probability $E[\theta_{x_j}]$) is

3. Example.

$$
\begin{aligned}
p(x_j|x_1,\ldots,x_{j-1}) &= \int p(x_j|\theta)p(\theta|x_1,\ldots,x_{j-1})d\theta && (22)\\
&= \int p(x_j|\theta)Dir(\boldsymbol{\theta}|\boldsymbol{\alpha_j})d\theta && (23)\\
&= c\int \prod_{i=1}^{L}\theta_i^{1(x_j=r_i)}\prod_i \theta^{\alpha_{ji}-1}d\theta \; where\; c=\frac{\Gamma(\alpha_{j,.})}{\prod_i\Gamma(\alpha_{j,i})} && (24)\\
&= c\int \prod_i \theta^{\alpha_{j+1}-1}d\theta && (25)\\
&= \frac{\Gamma(\alpha_{j.})}{\Gamma(\alpha_{j+1,.})}\frac{\prod_i\Gamma(\alpha_{j+1,i})}{\prod_i\Gamma(\alpha_{ji})} && (26)\\
&= \frac{\alpha_{j,x_j}}{\alpha_{j,.}} && (27)\\
& && (28)
\end{aligned}
$$

Dirichlet distribution III.

The marginal probability of the data set D_n with n_i occurrences of value r_i

4. Example.

$$p(x_1,\ldots,x_n|Dir(\boldsymbol{\alpha}_1)) = \prod_{i=1}^{n} p_i(x_i|x_1,\ldots,x_{i-1}) \quad (29)$$

$$= \frac{\prod_{i=1}^{L} \alpha_{1,i}..(\alpha_{1,i}+n_i)}{\alpha_{1,.}\ldots(\alpha_{1,.}+n)} \quad (30)$$

$$= \frac{\Gamma(\alpha_{1,.})}{\Gamma(\alpha_{1,.}+n)} \frac{\prod_{i=1}^{L}\Gamma(\alpha_{1,i}+n_i)}{\prod_{i=1}^{L}\Gamma(\alpha_{1,i})} \quad (31)$$

Bayesian inference with Monte Carlo I.

Integration/summation is a central operation in Bayesian statistics (c.f. optimization in the frequentist approach)

$$\bar{f} = E_{\pi(X)}[f(X)] \tag{32}$$

For example

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}, D_N) &= E_{p(G|D_N)}[E_{p(\boldsymbol{\theta}|G,D_N)}[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, G)]] \\ L_{\hat{G}|D_N} &= E_{p(G|D_N)}[L(G, \hat{G})] = \sum_G L(G, \hat{G}) p(G|D_N), \\ p(\alpha(G)|D_N) &= \sum_G 1(\alpha(G)\ is\ true) p(G|D_N) \end{aligned}$$

Idea:

1. sampling from $\pi(X)$ to generate i.i.d random samples $\{X_t, t = 1..N\}$;
2. computation of the estimate $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$;
3. providing confidence measure for $|\bar{f} - \hat{f}_N|$, where $\bar{f} = E_{\pi(X)}[f(X)]$.

Consistency and convergence speed I.

The estimate $\hat{f}_N$ is strongly consistent (by the "strong law of large number"), that is

$$P(lim_{N\to\infty}\hat{f}_N = \bar{f}) = 1 \tag{33}$$

The standardized of $\hat{f}_N$ has asymptotically Gaussian distribution (by the "central limit theorem"), that is

$$\frac{\hat{f}_N - \bar{f}}{\sigma_N} \to N(0,1)\ as\ N \to \infty\ \ where\ \sigma_N = Var(f(X))/\sqrt{N}. \tag{34}$$

If $f(X)$ is bounded, then non-asymptotic results about the speed of convergence are also available by the Hoeffding's inequality including the bound and by the Bernstein's inequality. Specifically, if $f(X)$ is within $[0,1]$, then

$$p(|\hat{f}_N - \bar{f}| \geq \epsilon) \quad \leq \quad 2\exp(-2\epsilon^2 N) \leq \delta \tag{35}$$

$$E[\ |\hat{f}_N - \bar{f}|\ ] \quad \leq \quad \sqrt{c_0/N}. \tag{36}$$

Averaging with importance sampling

If efficient sampling exists for another density $q(X)$ called importance density that has the same support as of $\pi(X)$, then using the identity

$$\bar{f} = E_{\pi(X)}[f(X)] = E_{q(X)}[\frac{f(X)\pi(X)}{q(X)}] \tag{37}$$

we can use the samples $\{X_t, t = 1..N\}$ from $q(X)$ to evaluate the original expectation with the following weighted average

$$\hat{f}_N = 1/N \sum_{t=1}^{N} w^*(X_t) f(X_t) = \frac{1/N \sum_{t=1}^{N} w(X_t) f(X_t)}{1/N \sum_{t=1}^{N} w(X_t)} \tag{38}$$

where $w^*(X_t)$ are the importance weights

$$w(X_t) = \frac{\pi(X_t)}{q(X_t)} \; and \; w^*(X_t) = \frac{w(X_t)}{1/N \sum_{t=1}^{N} w(X_t)} \tag{39}$$

Consistency and convergence speed II.

It is strongly consistent, its standardized is asymptotically Gaussian, additionally its variance

$$Var_{q(X)}(f(X))/N = 1/N \int (\frac{f(x)\pi(x)}{q(x)} - \bar{f})^2 q(x)dx \tag{40}$$

that is the quadratic estimation error, can be smaller than of the standard Monte Carlo with $Var_{\pi(X)}(f(X))/N$ and shown to be minimized by the selection of $q(X) \propto f(x)\pi(x)$, but in general it is advisable to select $q(X)$ as close as possible to $\pi(X)$ maintaining efficient sampling from $q(X)$

In the case of using the prior $p(X)$ as importance distribution for the (normalized) posterior $p(X|D)$ the (normalized) weights are the likelihoods normalized by the estimation of the data likelihood $p(D)$:

$$w^*(X_t) = \frac{p(X_t|D)}{p(X_t)} = \frac{p(D|X_t)}{p(D)} \tag{41}$$

$$p(D) = \int p(D|x)p(x)dx \approx 1/N \sum_{t=1}^{N} p(D|X_t) \tag{42}$$

Markov chains I.

Let $\mathcal{X} = \{X_0, X_1, \ldots\}$ is a sequence of random variables. The values of X_t are frequently interpreted as states from a state space, the index parameter frequently has a temporal or in biological sequence analysis a location interpretation.

4. Definition. *A sequence of random variables $\mathcal{X} = \{X_0, X_1, \ldots\}$ is called a (first-order) Markov chain, if $p(X_t|X_{t-1}, \ldots, X_0) = p(X_t|X_{t-1})$. The Markov chain is (time-)homogeneous, if the so called transition kernels $p(X_t|X_{t-1})$ does not depend on t.*

In this section, unless otherwise stated that the values of X_t are discrete and finite, denoted by nonnegative integers $S = \{0, 1, \ldots, K\}$. We use the notation $p^{(t)}$ for the distribution of X_t and $p(X_t = i) = p_i^{(t)}$. We always assume homogeneity and these allows a shorthand notation p_{ij} for the transition probabilities as $p_{ij} = p_{ij}^{(1)} = P(X_{t+1} = j|X_t = i)$, which are forming the (one-step) transition probability matrix $P = P^{(1)}[p_{ij}]$ (a stochastic matrix).

Markov chains II.

The "n-step" transition probability matrix $P^{(n)}$ containing $p_{ij}^{(n)} = P(X_{t+n} = j | X_t = i)$ is the nth power of P and

$$p'^{(n)} = p'^{(0)} P^{(n)}, \; where \; P^{(n)} = P^n. \tag{43}$$

A special distribution is the so called invariant distribution p^{inv}.

5. Definition. *The distribution p'^{inv} is called an invariant distribution of a homogeneous Markov chain $\mathcal{X}$ with transition probability matrix P, if $p'^{inv} = p'^{inv} P$ (Consequently, if $p^{(0)} = p'^{inv}$, then $p^{(t)} = p'^{inv}$ for $\forall \; t$.)*

For a first-order Markov chain $\mathcal{X}$ the identical marginals $p^{(t)} = p^{inv}$ implies that $\mathcal{X}$ is strongly stationer, that is the distributions of time-shifted finite marginals are identical, so the invariant distribution p^{inv} is frequently called a stationer distribution.

Stability, irreducibility, aperiodicity

6. Definition. *A Markov chain $\mathcal{X}$ is stable, if $\lim_{t\to\infty} p(X_t) = p^{(\infty)}$ exists, independent of the initial distribution $p(X_0)$ and it is a distribution (called limiting distribution or equilibrium distribution).*

Now we need the concept of irreducibility and aperiodicity to state a central result about the limiting and invariant distributions.

7. Definition. *The discrete and finite state space Markov chain $\mathcal{X}$ is called*

1. *irreducible, if there exists $n_{ij} > 0$ for all i, j that $p_{ij}^{(n_{ij})} > 0$,*
2. *aperiodic, if for some i (and with irreducibility for all), there exists $n_i > 0$ that for all $n \geq n_i$ $p_{ii}^{(n)} > 0$,*

1. Theorem. *If a discrete and finite state space Markov chain $\mathcal{X}$ is irreducible and aperiodic, then the chain is stable and there is a unique invariant distribution that is also the limiting distribution (i.e p'^{∞} is a unique, nonnegative solution of $p'^{\infty} = p'^{\infty} P$ and $\sum_i p_i^{(\infty)} = 1$).*

To simplify notation, for a stable chain we denote this unique limiting and invariant distribution (p^{∞}, p^{inv} with $\pi(X)$).

Ergodicity, confidence

2. Theorem. *If a discrete and finite state space Markov chain $\mathcal{X}$ is stable and $\bar{f} = E_{\pi(X)}[f(X)] < \infty$, then $P(lim_{N\to\infty}\hat{f}_N = \bar{f}) = 1$, where $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$.*

8. Definition. *The discrete and finite state space Markov chain $\mathcal{X}$ is called geometrically ergodic, if there exists $0 \leq \lambda < 1$ and function $V(.) > 1$ such that*

$$\sum_j |p_{ij}^{(t)} - \pi_j| \leq V(i)\lambda^t \ for\ all\ i \tag{44}$$

The smallest such λ is called a rate of convergence.

3. Theorem. *If a discrete and finite state space Markov chain $\mathcal{X}$ is geometrically ergodic (so stable), started with its invariant distribution $\pi(X)$ and for a real valued function f $\bar{f} = E_\pi[f(X)]$, $\sigma^2 = Var_\pi(f(X))$, $E_\pi[f(X)^{2+\epsilon}] \leq \infty$ with some $\epsilon > 0$, then for $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$*

$$\tau^2 = \sigma^2 + 2\sum_{k=1}^{\infty} E_\pi[(f(X_0) - \bar{f})(f(X_k) - \bar{f})] \tag{45}$$

exists, nonnegative and finite, and

$$\sqrt{N}\frac{\hat{f}_N - \bar{f}}{\tau} \to N(0,1)\ in\ distribution\ as\ N \to \infty. \tag{46}$$

Reversibility

9. Definition. *The discrete and finite state space Markov chain $\mathcal{X}$ with transition probability matrix P and invariant distribution p^{inv} is called reversible, if it satisfies the detailed balance condition*

$$\forall\, i,j\; p_i^{inv} P_{ij} = p_j^{inv} P_{ji}. \tag{47}$$

By summation it gives $\boldsymbol{p^{inv} P_{.j}} = p_j^{inv}$, which is the defining equation of an invariant distribution. Consequently, if for a given P $\boldsymbol{q}$ satisfies detailed balance, then it is an invariant distribution and vice versa, if for a given target distribution $\boldsymbol{q}$ we can construct a P such that it satisfies detailed balance with $\boldsymbol{q}$, then $\boldsymbol{q}$ is its invariant distribution. Furthermore, if the constructed P is such that the corresponding reversible Markov chain is irreducible and aperiodic as well, then $\boldsymbol{q}$ is its unique, invariant, limiting distribution, so we can generate (dependent) samples by sequential simulation and use it to estimate expectations and to provide confidence measures.

The Metropolis-Hastings Algorithm I.

Let $\pi(X)$ denote the unnormalized, strictly positive target distribution over $S = \{0, 1, \ldots, K\}$ ($\pi_i = \pi(X = i) \geq 0$). Let Q be a transition probability matrix ($Q\mathbf{1} = \mathbf{1}$), the so called proposal distribution (for transitions), such that $(q_{ij} \geq 0)\ iff\ (q_{ji} \geq 0)$. Define a Markov chain $\mathcal{X}$ with probability transition matrix P such that

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right); \forall i \neq j \tag{48}$$

using $0/0 = 0$ and define $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. Note that the construction needs only the ratios of the target distribution, which fits to the practical case of unnormalized posterior in Bayesian analysis.

Now $\pi(X)$ is the stationary distribution of the defined Markov chain, which can be proved by showing that the detailed balance condition is satisfied. The cases $i = j$ and if $q_{ij} = q_{ji} = 0$ are trivially satisfied. For $i \neq j$ with $q_{ij} \geq 0$, suppose that $\pi_i q_{ij} \geq \pi_j q_{ji}$, then

$$\pi_i p_{ij} = \pi_i \frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \pi_j q_{ji} = \pi_j p_{ji} \tag{49}$$

The Metropolis-Hastings Algorithm II.

If Q is irreducible, so will be P and the same is true for aperiodicity. Consequently, if we provide a proposal distribution Q that (its corresponding Markov chain) is irreducible and aperiodic, then for a given target distribution $\pi(X)$ the construction above defines a stable and reversible Markov chain with (invariant) limiting distribution $\pi(X)$).
If Q is symmetric, then we fall back to the original *Metropolis algorithm* without ratio of the proposal distributions

$$p_{ij} = q_{ij} \min(1, \frac{\pi_j}{\pi_i}); \forall i \neq j. \tag{50}$$

If Q depends on only some distance between the current state x_t and a proposed state x^* ($q(x^*|x_t) = q(|x^* - x_t|)$), then we get the *random-walk Metropolis algorithm* (the distance can be semantically defined in discrete spaces). If Q is independent of the current state ($q(x^*|x_t) = q(x^*)$), then we get the *independence sampler* , which is geometrically convergent determined by $\inf q(x)/\pi(x)$ (by the closeness to the target distribution) **?**. If Q is such that changes at most one component of X based on its full conditional distribution, then we get the *Gibbs sampler* , with an acceptance probability 1.

The Metropolis-Hastings Algorithm III.

0. [Construct an approximate distribution P^S of the posterior using mixture model around modes for checking and initialization of the MCMC.]
1. Construct an irreducible and aperiodic proposal distribution Q specific to the domain.
2. Draw an initial state x_0 from P^S.
3. For $t = 1, 2, \ldots$
   (a) Draw a candidate state x^* from the proposal distribution Q given x_t.
   (b) Calculate the *acceptance probability* of a step from x_t to x^* $\alpha(x_t, x^*) = min(1, \frac{\pi_{x^*} q_{x_t x^*}}{\pi_{x_t} q_{x_t x^*}})$.
   (c) Set $x_{t+1} = x^*$ with probability $\alpha(x_t, x^*)$, otherwise $x_{t+1} = x_t$.
4. Continue until convergence and specified confidence.
5. [Evaluate speed of convergence and improve efficiency by redesigning Q. Step back to 2.]
6. [Compare against base-line method using importance resampling with P^S. Step back to 1.]