Stochastic inference methods in Bayesian networks

P. Antal, P. Sárközy

antal@mit.bme.hu

BME

Bayesian inference with Monte Carlo I.

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

$$\bar{f} = E_{\pi(X)}[f(X)] \tag{1}$$

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

$$P(lim_{N\to\infty}\hat{f}_N = \bar{f}) = 1 \tag{2}$$

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

$$\frac{\hat{f}_N - \bar{f}}{\sigma_N} \to N(0,1)\ as\ N \to \infty\ \ where\ \sigma_N = Var(f(X))/\sqrt{N}. \tag{3}$$

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

$$p(|\hat{f}_N - \bar{f}| \geq \epsilon) \quad \leq \quad 2\exp(-2\epsilon^2 N) \leq \delta \tag{4}$$

$$E[\ |\hat{f}_N - \bar{f}|\ ] \quad \leq \quad \sqrt{c_0/N}. \tag{5}$$

Consistency and convergence speed II.

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

$$Var_{q(X)}(f(X))/N = 1/N \int (\frac{f(x)\pi(x)}{q(x)} - \bar{f})^2 q(x)dx \tag{6}$$

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 $p(D)$:

$$w^*(X_t) = \frac{p(X_t|D)}{p(X_t)} = \frac{p(D|X_t)}{p(D)} \tag{7}$$

$$p(D) = \int p(D|x)p(x)dx \approx 1/N \sum_{t=1}^{N} p(D|X_t) \tag{8}$$

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.

1. 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

$$p'^{(n)} = p'^{(0)} P^{(n)}, \; where \; P^{(n)} = P^n. \tag{9}$$

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

2. 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

3. 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.

4. 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 $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$.*

5. 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*

$$\sum_j |p_{ij}^{(t)} - \pi_j| \leq V(i)\lambda^t \ for\ all\ i \tag{10}$$

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

$$\tau^2 = \sigma^2 + 2\sum_{k=1}^{\infty} E_\pi[(f(X_0) - \bar{f})(f(X_k) - \bar{f})] \tag{11}$$

exists, nonnegative and finite, and

$$\sqrt{N}\frac{\hat{f}_N - \bar{f}}{\tau} \to N(0,1)\ in\ distribution\ as\ N \to \infty. \tag{12}$$

Reversibility

6. 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*

$$\forall\, i,j\; p_i^{inv} P_{ij} = p_j^{inv} P_{ji}. \tag{13}$$

By summation it gives $\boldsymbol{p^{inv} P_{\cdot j}} = p_j^{inv}$, 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

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right); \forall i \neq j \tag{14}$$

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

$$\pi_i p_{ij} = \pi_i \frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \pi_j q_{ji} = \pi_j p_{ji} \tag{15}$$

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

$$p_{ij} = q_{ij} \min(1, \frac{\pi_j}{\pi_i}); \forall i \neq j. \tag{16}$$

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.]

Convergence and confidence issues

The Metropolis-Hastings algorithm offers complete freedom to design the proposal distribution specific to the domain, because it is ensured that the distribution and the averages will converge asymptotically. However for a Metropolis-Hastings algorithm with a specified proposal Q and target $\pi(X)$ distributions there are no analytic results with general, practical applicability for the rate of convergence to the target distribution, for forgetting the starting values or for the Monte Carlo variance of the average Eq. 11. Consequently, the length of the necessary simulation is usually determined by observing and analyzing simulations, practically based on the actual sampling.

These two problems of the convergence to the limiting distribution and the convergence of the ergodic average shows the dual usage of MCMC methods: generation of samples from the target distribution for its exploration and computing ergodic averages for approximating expectations. Clearly, to solve the second in general much easier than the first in many respect, e.g. the induced distribution of the target quantity can converge much faster than the target distribution. Furthermore, as our primary goal is the reliable approximation of expectations of the target quantities and not per se the convergence of the induced distribution of the target quantity, an optimal method would provide an estimate with a confidence interval without answering the question of convergence to the limiting distribution

Convergence diagnostic I.

The first method related to burn in is based on a single chain and the mean**?**. It tests the convergence of a single realization from the sequence $\{Y_i; i = 1..N\}$ exploiting that after burn-in (i.e. in case of convergence) the distribution of an ergodic average is asymptotically Gaussian. Formally, define the averages $\hat{Y}_b$ after a (putative) burn-in m and $\hat{Y}_a$ at the end of sequence

$$\hat{Y}_b = \frac{1}{N_b} \sum_{i=m+1}^{m+N_b} Y_i \text{ and } \hat{\mathrm{Y}}_\mathrm{a} = \frac{1}{\mathrm{N_a}} \sum_{\mathrm{i=N-N_a+1}}^{\mathrm{N}} \mathrm{Y_i} \tag{17}$$

with no overlap ($m + N_b + N_a < N$). If N_a/N and N_b/N are fixed, then

$$z_G = \frac{\hat{Y}_b - \hat{Y}_a}{\sqrt{\hat{Var}(Y_b) + \hat{Var}(Y_a)}} \rightarrow N(0,1) \text{ in distribution.} \tag{18}$$

The second method is using independently initialized multiple chains and analyze the variance **?**. Its test is based on the relation of the estimators of the (Monte Carlo) variance of the target quantity using a between-sequence estimation (i.e. the variance of the (independent) estimates for the chains) and a within-sequence estimation (i.e. average of the within-sequence estimates of the variance).

Convergence diagnostic II.

Formally, for M chains with N samples $\{Y_i, j; i = 1..N\, j = 1 \ldots, M\}$ define

$$B = \frac{N}{M-1}\sum_{j=1}^{M}(\bar{Y}_{+,j} - \bar{(Y)}_{+,+})^2, \text{ where } \bar{\mathrm{Y}}_{+,\mathrm{j}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{Y}_{\mathrm{ij}}, \bar{\mathrm{Y}}_{+,+} = \frac{1}{\mathrm{M}}\sum_{\mathrm{j}=1}^{\mathrm{M}}\mathrm{Y}_{+,\mathrm{j}}$$

$$W = \frac{1}{M}\sum_{j=1}^{M}s_j^2 \text{ where } \mathrm{s}_{\mathrm{j}}^2 = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\bar{\mathrm{Y}}_{\mathrm{ij}} - \bar{(\mathrm{Y})}_{+,\mathrm{j}})^2.$$

If the simulations are started independently from an overdispersed starting distribution, then the quantity

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{var}^+(Y)}{W}}, \text{ where } \hat{\mathrm{var}}^+(\mathrm{Y}) = \frac{\mathrm{N}-1}{\mathrm{N}}\mathrm{W} + \frac{1}{\mathrm{NB}} \tag{19}$$

called ”potential scale reduction" can be used to monitor convergence, because $\hat{var}^+(Y)$ overestimates the variance as the chains are still overdispersed and W underestimates the variance as they are still confined to small regions. Various approximate distributions can be constructed for B/W **???**, which could be used to construct statistical tests or to select a task specific constant and continue the simulation until $\sqrt{\hat{R}}$ decline below this for all the target quantities using small number of chains.

Confidence estimation I.

The second task after the determination and elimination of the burn-in period is to determine the stopping time and/or providing confidence measure(s) for the estimate(s) (see Eq. 11 for the ”MCMC" variance and Eq. **??** for an ”MCMC" central limit theorem). The first method is related to the between-sequence variance of the earlier method, though using a single chain $\{Y_i; i = 1, \ldots, NM\}$. It partitions a sufficiently long chain into M parts with length N such that the ergodic averages are approximately independently Gaussian with mean $E_\pi[f(X)]$ and variance τ^2/N (see Eq. 11). Then approximate τ^2 as follows

$$\hat{\tau}^2 = \frac{N}{(M-1)} \sum_{j=1}^{M} (\bar{Y}_j - \bar{\bar{Y}})^2, \text{ where } \bar{\mathrm{Y}}_\mathrm{j} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=(\mathrm{j}-1)\mathrm{N}+1}^{\mathrm{jN}} \mathrm{Y}_\mathrm{i}, \bar{\bar{\mathrm{Y}}} = \frac{1}{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{M}} \bar{\mathrm{Y}}_\mathrm{j}$$

Confidence estimation II.

Another method is based on the direct estimation of the autocovariance terms $\gamma_k = E_\pi[(f(X_0) - \bar{f})(f(X_k) - \bar{f})]$ in the Eq. 11 of the Monte Carlo variance with

$$\hat{\gamma}_k = \frac{1}{N-k} \sum_{i=1}^{N-k} (Y_i - \bar{f})(Y_{i+k} - \bar{f}) \tag{20}$$

and use a special weighting to eliminate the not reliable autocorrelation terms as follows

$$\hat{\tau}_N^2 = \hat{\gamma}_0 + 2 \sum_{i=1}^{\infty} w_N(i) \hat{\gamma}_0, \text{ where } 0 \leq \mathrm{w_N(i)} \leq 1. \tag{21}$$

Speeding up MCMC

The Metropolis-Hastings algorithm allows the incorporation of any proposal distribution satisfying only mild conditions (see **??**). However, the general improvement of a proposal distribution for faster convergence to the limiting distribution and for better estimates of the scalar quantities is an open issue. A general empirical method for random walk Metropolis, where the proposal distribution depends on only some distance between the current state x and a proposed state x is to calibrate this distribution as follows. The average step size is increased till the acceptance rate (the average acceptance probability, see Eq. 12)

$$a_q = E_{\pi(X)}[E_{q(Y|X)}[\alpha(X,Y)]] \tag{22}$$

is close to 1. But this is not enough as the following example from **?** shows.

1. Example. *Let the target distribution be a Gaussian with a diagonal covariance matrix in which the maximum and minimum are σ^{max} and σ^{min}. Assume a proposal distribution corresponding to a random walk Metropolis that is a centralized Gaussian with an identical diagonal covariance matrix with value ϵ corresponding to an acceptance rate close to 1. Then the coordinates evolves as independent one-dimensional random walks and ϵ is in the range of σ^{min}. Consequently, after T steps it explored a region less than $\epsilon\sqrt{2T \ln \ln T}$ and located about $\epsilon\sqrt{T}$. So an estimate for the necessary number of steps to explore the direction with σ^{max} is ${\sigma^{max}}/{\sigma^{min}}^2$.*

This problem can be solved with an (offline) coordinate transformation (reparameterezation) and/or dimension reduction (model projection).