Machine Learning in ICT

TECHNISCHE UNIVERSITÄT DARMSTADT

Programming Exercise 2: Sampling Methods and Graphical Models

SoSe 2022
Due date: May 25, 2022

Department of Electrical Engineering and Information Technology
Instructors: Profs A. Klein, H. Koeppl
Teaching Assistant: Y. Eich, yannick.eich@tu-darmstadt.de, G. Ekinci, gizem.ekinci@tu-darmstadt.de

Problem 1 (20 pts)

Algorithm 1 Metropolis-Hastings algorithm

Initialize $X^0 = x^0$
for $i = 0$ to $L - 1$ **do**
  Generate $X' \sim q(x'|X^i = x^i)$
  Evaluate the acceptance probability $r = \alpha(x'|x^i) = \min\{1, \frac{q(x^i|x')f(x')}{q(x'|x^i)f(x^i)}\}$
  Generate $U \sim \mathcal{U}(0, 1)$
  if $U < r$ **then**
    Accept the proposal: $X^{i+1} := X'$
  else
    Reject the proposal
  end if
end for

(a) Implement Metropolis-Hastings algorithm to sample the sequence of two-dimensional RVs with a Rosenbrock density function $f(x) = \frac{1}{Z}f(x_1, x_2) = \frac{1}{Z}e^{-\frac{100(x_2-x_1^2)^2+(1-x_1)^2}{20}}$, where Z is a normalization constant. For the proposal distribution, take $q(x'|x) = \mathcal{N}\left(x, \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}\right)$.
Assume $L = 10^5$, $\sigma = 0.5$, $x^0 = (0, 10)$.
Plot

(i) the generated two-dimensional sequence **(5 pts)**

(ii) the evolution of each coordinate versus iteration (the trace plot) **(5 pts)**

(iii) the acceptance rate, i.e. the ratio between accepted proposals and the number of samples, versus σ^2, where the standard deviation is varied as $\sigma = \sigma_{min} + i\Delta\sigma, i \in [0, 20], i \in N_+, \sigma_{min} = 0, \Delta\sigma = 0.05$. You should observe that if the standard deviation is small, the acceptance rate will be high but the progress through the state space is slow. **(5 pts)**

(b) The samples produced by MCMC are auto-correlated. We can quantify this the following way. Define the sample-based auto-correlation function at a lag k of a set of samples $s_1, s_2, \ldots, s_L$ as follows:

$$r_k = \frac{\frac{1}{L-k}\sum_{i=1}^{L-k}(s_i - \overline{s})(s_{i+k} - \overline{s})}{\frac{1}{L}\sum_{i=1}^{L}(s_i - \overline{s})^2}$$

where $\overline{s} = \frac{1}{L}\sum_{i=1}^{L} s_i$. Compute the auto-correlation function for the sequence from (a) for both dimensions for $\sigma \in \{0.1, 0.5, 1\}$ and produce plots of r_k vs k. Assume $k = 1, \ldots, 2000$. You should see that with the increase of k, the auto-correlation function decreases. This information helps you to decide upon what sub-sampling size should be applied in order to get an uncorrelated sequence. Note also that the analysis of the sequence can only be performed after the required distribution reaches its equilibrium. That is why all the samples, produced till that point (during the burn-in time), have to be discarded. Compute the auto-correlation function for both dimensions of the sequence $X^{\frac{L}{2}}, \ldots, X^L$ from (a). **(5 pts)**

Problem 2

Algorithm 2 Gibbs Sampling

Initialize $X^0 = x^0$
for $i = 1$ to L **do**
  Sample $X_1^{i}{}' \sim p(x_1 | X_2 = x_2^{i-1})$
  Sample $X_2^{i}{}' \sim p(x_2 | X_1 = x_1^{i-1})$
end for

Let $X^0, X^1, \ldots, X^L$ be a sequence of two-dimensional RVs, i.e. $X^i = (X_1^i, X_2^i)$, for $i = 0, \ldots, L$. Implement Gibbs sampling algorithm to sample this sequence from a bivariate normal distribution $\mathcal{N}\left((\mu_1, \mu_2), \begin{bmatrix} \sigma^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma^2 \end{bmatrix}\right)$. In order to sample from this distribution using a Gibbs sampler, we need to have the conditional distributions $p(x_1|X_2 = x_2)$ and $p(x_2|X_1 = x_1)$. They are given by $X_1|X_2 = x_2 \sim \mathcal{N}(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_{2\mu 2}), (1-\rho^2)\sigma_1^2)$ and $X_2|X_1 = x_1 \sim \mathcal{N}(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_{1\mu 1}), (1-\rho^2)\sigma_2^2)$ Generate a sequence using this algorithm and plot

(i) the generated two-dimensional sequence

(ii) the evolution of each coordinate versus iteration

(iii) the autocorrelation function from Problem 1 for each dimension of $X^{\frac{L}{2}}, \ldots, X^L$.

Assume $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1, \rho = 0.8$.

Problem 3

Recall that a directed acyclic graph (DAG) is a directed graph that contains no directed cycles. A directed graphical model (DGM) is a DAG where each node represents an RV and the joint distribution factorizes as $p(x_1, \ldots, x_n) = \prod_{k=1}^{D} p(x_k | x_{pa}(k))$, where $x_{pa}(k)$ is the set of values x_i for which i is a parent node of k. Suppose that the vertices of a DGM are topologically ordered. The ancestral sampling algorithm allows us to draw a sample $x_1, \ldots, x_n$ from the joint distribution of a DGM. We start with the nodes that do not have any parents and draw a sample from their marginal distributions. We then visit each other node so that for a node x_i we draw a sample from the conditional distribution $p(x_i|pa(x_i))$, where the parent variables have been set to their sampled values. Consider now the Burglar-alarm network discussed in the lecture and in and suppose all variables are binary (1 $\equiv$ true, 0 $\equiv$ false). Assume the prior probabilities $p(E = 1) = 0.002$ and $P(B = 1) = 0.001$. From the factorization property of Bayesian networks, the joint distribution is fully defined by the following conditional probability tables:

$p(A = 1\|B, E)$	B	E
0.95	1	1
0.94	1	0
0.29	0	1
0.001	0	0

$p(C = 1\|A)$	A
0.9	1
0.05	0

$p(R = 1\|E)$	E
0.85	1
0.02	0

Implement the ancestral sampling algorithm for the given Burglar-alarm network.