Personal notes – Bayesian machine learning

Tuan Anh Le

September 5, 2014

Contents

1 Probability distributions

1.1 Uniform distribution

1.2 Beta distribution

1.3 Bernoulli distribution

1.4 Binomial distribution

1.5 Beta-binomial distribution

1.6 Categorical distribution

1.7 Dirichlet distribution

1.8 Multinomial distribution

1.9 Pareto distribution

2 Bayesian parameter estimation

2.1 Beta-Bernoulli model

2.1.1 Summary

The model

$$
\begin{aligned}
X_i &\sim \mathrm{Ber}(\theta), \text{for } i \in \{1, \dots, N\} && (2.1)\\
\mathcal{D} &= \{x_1, \dots, x_N\} && (2.2)\\
N_1 &= \sum_{i=1}^{N} \mathbb{I}(x_i = 1) && (2.3)\\
N_0 &= \sum_{i=1}^{N} \mathbb{I}(x_i = 0) && (2.4)
\end{aligned}
$$

Likelihood

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{2.5}$$

Prior

$$p(\theta) = \mathrm{Beta}(\theta|a, b) \tag{2.6}$$

Posterior

$$p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a' = N_1 + a, b' = N_0 + b) \tag{2.7}$$

Posterior predictive

$$p(\tilde{x} = 1|\mathcal{D}) = \frac{a'}{a' + b'} \tag{2.8}$$

Evidence

2.1.2 Derivations

2.2 Beta-binomial model

2.2.1 Summary

The model

$$N_1 \sim \mathrm{Bin}(N, \theta) \tag{2.9}$$

$$
\begin{align}
\mathcal{D} &= \{N_1, N\} \tag{2.10}\\
N_1 &= \text{number of successes} \tag{2.11}\\
N &= \text{total number of trials} \tag{2.12}\\
\tilde{\mathcal{D}} &= \{\tilde{N}_1, \tilde{N}\} \tag{2.13}\\
\tilde{N}_1 &= \text{number of successes in a new batch of data} \tag{2.14}\\
\tilde{N} &= \text{total number of trials in a new batch of data} \tag{2.15}
\end{align}
$$

Likelihood

$$p(\mathcal{D}|\theta) = \mathrm{Bin}(N_1|N, \theta) \tag{2.16}$$

Prior

$$p(\theta) = \mathrm{Beta}(\theta|a, b) \tag{2.17}$$

Posterior

$$p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a' = N_1 + a, b' = N_0 + b) \tag{2.18}$$

Posterior predictive

$$p(\tilde{\mathcal{D}}|\mathcal{D}) = \mathrm{Bb}(\tilde{N}_1; a', b', \tilde{N}) \tag{2.19}$$

Evidence

2.2.2 Derivations

2.3 Dirichlet-categorical model

2.3.1 Summary

The model

$$
\begin{align}
X_i &\sim \mathrm{Cat}\left(\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T\right), \text{for } i \in \{1, \ldots, N\} \tag{2.20}\\
\mathcal{D} &= \{x_1, \ldots, x_N\} \tag{2.21}\\
n_k &= \sum_{i=1}^{N} \mathbb{I}(x_i = k) \tag{2.22}
\end{align}
$$

Likelihood

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{n_k} \tag{2.23}$$

Prior

$$p(\theta) = \mathrm{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) \tag{2.24}$$

Posterior

$$p(\theta|\mathcal{D}) = \text{Dir}\left(\boldsymbol{\theta}; \boldsymbol{\alpha}' = \boldsymbol{\alpha} + (n_1, \dots, n_K)^T\right) \tag{2.25}$$

Posterior predictive

$$p(\tilde{X} = j|\mathcal{D}) = \frac{\alpha'_j}{\sum_{k=1}^K \alpha'_i} \tag{2.26}$$

$$= \frac{\alpha_j + n_j}{\alpha_0 + N} \tag{2.27}$$

$$\text{where } \alpha_0 = \sum_{k=1}^K \alpha_k \tag{2.28}$$

Evidence

2.3.2 Derivations

2.4 Dirichlet-multinomial model

2.4.1 Summary

The model

$$\mathbf{N} \sim \text{Mult}(N, \boldsymbol{\theta}) \in \mathbb{R}^K \tag{2.29}$$

$$\mathcal{D} = \{\mathbf{n} = \text{vector of counts of successes}\} \tag{2.30}$$

$$N = \sum_{i=1}^K n_i \tag{2.31}$$

$$\tilde{\mathcal{D}} = \{\tilde{\mathbf{n}} = \text{vector of counts of successes in a new batch of data}\} \tag{2.32}$$

$$\tilde{N} = \sum_{i=1}^K \tilde{n}_i \tag{2.33}$$

Likelihood

$$p(\mathcal{D}|\theta) = \text{Mult}(\mathbf{n}; N, \boldsymbol{\theta}) \tag{2.34}$$

Prior

$$p(\theta) = \text{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) \tag{2.35}$$

Posterior

$$p(\theta|\mathcal{D}) = \text{Dir}\left(\boldsymbol{\theta}; \boldsymbol{\alpha}' = \boldsymbol{\alpha} + (n_1, \dots, n_K)^T\right) \tag{2.36}$$

Posterior predictive

$$p(\tilde{\mathcal{D}}|\mathcal{D}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_0 + N + \tilde{N})} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + n_k + \tilde{n}_k)}{\Gamma(\alpha_k + n_k)} \tag{2.37}$$

$$\text{where } \alpha_0 = \sum_{k=1}^{K} \alpha_k \tag{2.38}$$

Evidence

$$p(\mathcal{D}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)} \tag{2.39}$$

2.4.2 Derivations

2.5 Poisson-gamma model

2.5.1 Summary

The model

$$x \sim \text{Poi}(\lambda) \tag{2.40}$$

$$\mathcal{D} = \{x_1, \ldots, x_N\} \tag{2.41}$$

Likelihood

$$p(\mathcal{D}|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \tag{2.42}$$

Prior

$$p(\lambda) = \text{Gamma}(\lambda; a, b) \tag{2.43}$$

Posterior

$$p(\lambda|\mathcal{D}) = \text{Gamma}\left(\lambda; a' = a + \sum_{i=1}^{N} x_i, b' = b + N\right) \tag{2.44}$$

Posterior predictive

$$p(\tilde{x}|\mathcal{D}) = \text{NB}\left(\tilde{x}|a', \frac{1}{1+b'}\right) \tag{2.45}$$

Evidence

$$p(\mathcal{D}) = \tag{2.46}$$

2.5.2 Derivations

3 Sampling algorithms

3.1 Introduction

Let p be a probability distribution with a pdf $p(\mathbf{x}), \mathbf{x} \in \mathcal{X}$ (usually $\mathcal{X} = \mathbb{R}^D, D \in \mathbb{N}$), which we assume can be evaluated within a multiplicative factor (i.e. we can only evaluate $p^*(\mathbf{x}) = Z_p p(\mathbf{x})$, where $Z_p = \int_{\mathcal{X}} p^*(\mathbf{x})\,\mathrm{d}\mathbf{x}$). We want to achieve the following:

Problem 1 Generate samples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(R)}\}, R \in \mathbb{N}$ (we will use the shorthand notation $\{\mathbf{x}^{(r)}\}$ from now) from the probability distribution p.

Problem 2 Estimate the expectation of an arbitrary function f given $\mathbf{x} \sim p$, $\mathrm{E}_{\mathbf{x}\sim p}\left[f(\mathbf{x})\right]$ (we will use the shorthand notation $\mathrm{E}[f]$ from now).

3.2 Rejection sampling

Assume we can sample from a proposal distribution q with a pdf $q(\mathbf{x})$, which can be evaluated within a multiplicative factor (i.e. we can only evaluate $q^*(\mathbf{x}) = Z_q q(\mathbf{x})$). Also assume we know the value of a constant c such that

$$cq^*(\mathbf{x}) > p^*(\mathbf{x}) \text{ for all } \mathbf{x} \tag{3.1}$$

The procedure that generates a sample $\mathbf{x} \sim p$ is described in Algorithm 1 below.

Algorithm 1 Rejection sampling

1: Generate $\mathbf{x} \sim q$.
2: Generate $u \sim \mathrm{Unif}(0, cq^*(\mathbf{x}))$.
3: If $u > p^*(\mathbf{x})$ it is rejected, otherwise it is accepted.

3.2.1 Why it works?

Assume $\mathbf{x} \in \mathbb{R}^D$. Define sets $\mathcal{X}$ and $\mathcal{X}'$ to be

$$\mathcal{X} = \left\{\boldsymbol{\alpha} \in \mathbb{R}^{d+1} : \alpha_{1:d} \in \mathbb{R}^d, \alpha_{d+1} \in [0, cq^*(\boldsymbol{\alpha})]\right\} \tag{3.2}$$

$$\mathcal{X}' = \left\{\boldsymbol{\alpha} \in \mathbb{R}^{d+1} : \alpha_{1:d} \in \mathbb{R}^d, \alpha_{d+1} \in [0, p^*(\boldsymbol{\alpha})]\right\} \tag{3.3}$$

Note that $\mathcal{X}' \subseteq \mathcal{X}$.

By definition, $\mathcal{X}$ is the support of $(\mathbf{x}, u)$. The probability of $(\mathbf{x}, u)$ can be expressed as

$$
\begin{aligned}
\Pr(\mathbf{x}, u) &= \Pr(\mathbf{x}) \Pr(u) && (3.4)\\
&= q(\mathbf{x}) \frac{1}{c q^*(\mathbf{x})} && (3.5)\\
&= q(\mathbf{x}) \frac{1}{c Z_q q(\mathbf{x})} && (3.6)\\
&= \frac{1}{c Z_q} && (3.7)
\end{aligned}
$$

which is constant w.r.t. $(\mathbf{x}, u)$, i.e.

$$
(\mathbf{x}, u) \sim \mathrm{Unif}(\mathcal{X}) \tag{3.8}
$$

Let $(\mathbf{x}', u')$ be the value of $(\mathbf{x}, u)$ that gets accepted. By definition, $\mathcal{X}'$ is the support of $(\mathbf{x}', u')$:

$$
(\mathbf{x}', u') = \begin{cases} (\mathbf{x}, u) & \text{if } (\mathbf{x}, u) \in \mathcal{X}' \\ \text{nothing} & \text{otherwise.} \end{cases} \tag{3.9}
$$

The probability of $(\mathbf{x}', u')$ can be expressed as

$$
\Pr(\mathbf{x}', u') = \begin{cases} \Pr(\mathbf{x}, u) & \text{if } (\mathbf{x}, u) \in \mathcal{X}' \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}
$$

which means

$$
(\mathbf{x}', u') \sim \mathrm{Unif}(\mathcal{X}') \tag{3.11}
$$

Working backwards

$$
\begin{aligned}
\Pr(\mathbf{x}') &= \frac{\Pr(\mathbf{x}', u')}{\Pr(u')} && (3.12)\\
&\propto \frac{1}{1/p^*(\mathbf{x}')} && (3.13)\\
&\propto p^*(\mathbf{x}') && (3.14)
\end{aligned}
$$

Hence the accepted $\mathbf{x}$, $\mathbf{x}'$ is $\sim p$.

3.3 Importance sampling

Assume we can sample from a proposal distribution q with a pdf $q(\mathbf{x})$, which can be evaluated within a multiplicative factor (i.e. we can only evaluate $q^*(\mathbf{x}) = Z_q q(\mathbf{x})$). To solve problem 2, we follow Algorithm 2 below.

Algorithm 2 Importance sampling

1: Generate samples from q, $\{\mathbf{x}^{(r)}\}$.
2: Calculate importance weights $w_r = \frac{p^*(\mathbf{x}^{(r)})}{q^*(\mathbf{x}^{(r)})}$.
3: $\hat{\mathbf{y}} = \frac{\sum_r w_r f(\mathbf{x}^{(r)})}{\sum_r w_r}$ is the estimator of $\mathrm{E}[f]$.

3.3.1 Convergence of estimator as R increases

We want to prove that if $q(\mathbf{x})$ is non-zero for all $\mathbf{x}$ where $p(\mathbf{x})$ is non-zero, the estimator $\hat{\mathbf{y}}$ converges to $\mathrm{E}[f]$, as R increases. We consider the the expectations of the numerator and denominator separately:

$$
\begin{aligned}
\mathrm{E}_q[\text{numer}] &= \mathrm{E}_q\left[\sum_r w_r f(\mathbf{x}^{(r)})\right] && (3.15)\\
&= \sum_r \mathrm{E}_q\left[w_r f(\mathbf{x}^{(r)})\right] && (3.16)\\
&= \sum_r \mathrm{E}_q\left[\frac{p^*(\mathbf{x}^{(r)})}{q^*(\mathbf{x}^{(r)})} f(\mathbf{x}^{(r)})\right] && (3.17)\\
&= \sum_r \mathrm{E}_q\left[\frac{Z_p p(\mathbf{x}^{(r)})}{Z_q q(\mathbf{x}^{(r)})} f(\mathbf{x}^{(r)})\right] && (3.18)\\
&= \frac{Z_p}{Z_q} \sum_r \int_{\mathbf{x}^{(r)}} p(\mathbf{x}^{(r)}) f(\mathbf{x}^{(r)})\,\mathrm{d}\mathbf{x}^{(r)} && (3.19)\\
&= \frac{Z_p}{Z_q} \sum_r \mathrm{E}_p\left[f(\mathbf{x}^{(r)})\right] && (3.20)\\
&= \frac{Z_p}{Z_q} R\, \mathrm{E}_p\left[f(\mathbf{x})\right] && (3.21)\\
\mathrm{E}_q[\text{denom}] &= \mathrm{E}_q\left[\sum_r w_r\right] && (3.22)\\
&= \sum_r \mathrm{E}_q\left[\frac{p^*(\mathbf{x}^{(r)})}{q^*(\mathbf{x}^{(r)})}\right] && (3.23)\\
&= \sum_r \mathrm{E}_q\left[\frac{Z_p p(\mathbf{x}^{(r)})}{Z_q q(\mathbf{x}^{(r)})}\right] && (3.24)\\
&= \frac{Z_p}{Z_q} \sum_r \int_{\mathbf{x}^{(r)}} p(\mathbf{x}^{(r)})\,\mathrm{d}\mathbf{x}^{(r)} && (3.25)\\
&= \frac{Z_p}{Z_q} R && (3.26)
\end{aligned}
$$

Hence $\hat{\mathbf{y}}$ converges to $\mathrm{E}_p[f]$ as R increases (but is not necessarily an unbiased estimator because $\mathrm{E}_q[\hat{\mathbf{y}}]$ is not necessarily $= \mathrm{E}_p[f]$).

3.3.2 Optimal proposal distribution

Assuming we can evaluate $p(\mathbf{x})$ and $q(\mathbf{x})$, we want to find a proposal distribution q to minimise the variance of the weighted samples

$$\operatorname{var}_q\left[\frac{p(\mathbf{x})}{q(\mathbf{x})}f(\mathbf{x})\right] = \mathrm{E}_q\left[\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})}f^2(\mathbf{x})\right] - \left(\mathrm{E}_q\left[\frac{p(\mathbf{x})}{q(\mathbf{x})}f(\mathbf{x})\right]\right)^2 \tag{3.27}$$

$$= \mathrm{E}_q\left[\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})}f^2(\mathbf{x})\right] - \left(\mathrm{E}_p\left[f(\mathbf{x})\right]\right)^2 \tag{3.28}$$

The second part is independent of q so we can ignore it. By Jensen's inequality, we have $\mathrm{E}\left[g(u(\mathbf{x}))\right] \geq g\left(\mathrm{E}\left[u(\mathbf{x})\right]\right)$ for $u(\mathbf{x}) \geq 0$ where $g : x \mapsto x^2$. Setting $u(\mathbf{x}) = p(\mathbf{x})|f(\mathbf{x})|/q(\mathbf{x})$, we have the following lower bound:

$$\mathrm{E}_q\left[\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})}f^2(\mathbf{x})\right] \geq \left(\mathrm{E}_q\left[\frac{p(\mathbf{x})}{q(\mathbf{x})}|f(\mathbf{x})|\right]\right)^2 = \left(\mathrm{E}_p[|f(\mathbf{x})|]\right)^2 \tag{3.29}$$

with the equality when $u(\mathbf{x}) = \text{const.} \implies q_{\text{optimal}}(\mathbf{x}) \propto |f(\mathbf{x})|p(\mathbf{x})$. Taking care of normalisation, we get

$$q_{\text{optimal}}(\mathbf{x}) = \frac{|f(\mathbf{x})|p(\mathbf{x})}{\int |f(\mathbf{x}')|p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'} \tag{3.30}$$

3.4 Sampling importance resampling

In Sampling importance resampling (SIR), we approximate the pdf of p as point masses and resample from them to get samples $\left\{\mathbf{x}^{(r)}\right\}$ which are approximately $\sim p$. The process is described in Algorithm 3 below.

Algorithm 3 Sampling importance resampling

1: Generate samples $\left\{\mathbf{z}^{(r)}\right\}$ from q.
2: Calculate importance weights $\left\{w_r = \frac{p^*(\mathbf{z}^{(r)})}{q^*(\mathbf{z}^{(r)})}\right\}$.
3: Calculate the normalised importance weights $\left\{\hat{w}_r = \frac{w_r}{\sum_{r'} w_{r'}}\right\}$. Note that $\sum_r \hat{w}_r = 1$.
4: Resample from a probability distribution with the pmf

$$f(\mathbf{x}) = \sum_r \hat{w}_r \delta_{\mathbf{z}^{(r)}}(\mathbf{x}) \tag{3.31}$$

5: The resulting samples $\left\{\mathbf{x}^{(r)}\right\}$ are approximately $\sim p$.

3.4.1 Why it works?

We consider the univariate case (to do: general case) as the number of proposal samples (particles) $R \to \infty$. We can express the number of proposal samples that are in the interval $\lim_{\delta x \to 0}[x, x+\delta x]$, $N(x)$, to be

$$N(x) = \lim_{\delta x \to 0} Rq(x)\delta x \tag{3.32}$$

We can express the probability of the one final sample, $x^{(r)}$ being in the interval $\lim_{\delta x \to 0}[x, x+\delta x]$ to be

$$\begin{aligned}
\lim_{\delta x \to 0} \Pr(x \le x^{(r)} \le x + \delta x) &= N(x)\hat{w}_r && (3.33)\\
&\propto \lim_{\delta x \to 0} Rq(x)\delta x \frac{p(x)}{q(x)} && (3.34)\\
&\propto \lim_{\delta x \to 0} p(x)\delta x && (3.35)
\end{aligned}$$

Hence (to do: why exactly does that result in an integral)

$$\begin{aligned}
\Pr(a \le x^{(r)} \le b) &\propto \int_a^b p(x)\,\mathrm{d}x && (3.36)\\
\implies x^{(r)} &\sim p && (3.37)
\end{aligned}$$

3.5 Particle filtering

3.5.1 Sequential importance sampling (SIS)

Assume the probabilistic graphical model similar to the one in HMMs, where $\mathbf{x}_t$ and $\mathbf{y}_t$ are the hidden and observed random variables at time t respectively. We want to sample from the distribution $p(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t})$. Assume we can sample from the probability distribution with the pdf of the following form

$$\begin{aligned}
q(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t}) &= q(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t})q(\mathbf{x}_{1:t-1} \mid \mathbf{y}_{1:t}) && (3.38)\\
&= q(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t})q(\mathbf{x}_{1:t-1} \mid \mathbf{y}_{1:t-1}) && (3.39)\\
&= q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}_t) && (3.40)
\end{aligned}$$

If we express the pdf of p in the form of

$$\begin{aligned}
p(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t}) &= \frac{p(\mathbf{y}_{1:t} \mid \mathbf{x}_{1:t})p(\mathbf{x}_{1:t})}{p(\mathbf{y}_{1:t})} && (3.41)\\
&= \frac{p(\mathbf{y}_t \mid \mathbf{x}_{1:t}, \mathbf{y}_{1:t-1})p(\mathbf{y}_{1:t-1} \mid \mathbf{x}_{1:t})p(\mathbf{x}_{1:t})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})p(\mathbf{y}_{1:t-1})} && (3.42)\\
&= \frac{p(\mathbf{y}_t \mid \mathbf{x}_{1:t}, \mathbf{y}_{1:t-1})p(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} && (3.43)\\
&= \frac{p(\mathbf{y}_t \mid \mathbf{x}_{1:t}, \mathbf{y}_{1:t-1})p(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1})p(\mathbf{x}_{1:t-1} \mid \mathbf{y}_{1:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} && (3.44)\\
&= \frac{p(\mathbf{y}_t \mid \mathbf{x}_t)p(\mathbf{x}_t \mid \mathbf{x}_{t-1})p(\mathbf{x}_{1:t-1} \mid \mathbf{y}_{1:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} && (3.45)
\end{aligned}$$

$$\propto p(\mathbf{y}_t \mid \mathbf{x}_t)p(\mathbf{x}_t \mid \mathbf{x}_{t-1})p(\mathbf{x}_{1:t-1} \mid \mathbf{y}_{1:t-1}) \tag{3.46}$$

we can write the weight of the sample $\mathbf{x}_{1:t}^{(r)}$ from the proposal q to be

$$w_t^{(r)} \propto \frac{p\left(\mathbf{x}_{1:t}^{(r)} \mid \mathbf{y}_{1:t}\right)}{q\left(\mathbf{x}_{1:t}^{(r)} \mid \mathbf{y}_{1:t}\right)} \tag{3.47}$$

$$\propto \frac{p\left(\mathbf{y}_t \mid \mathbf{x}_t^{(r)}\right) p\left(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}\right) p\left(\mathbf{x}_{1:t-1}^{(r)} \mid \mathbf{y}_{1:t-1}\right)}{q\left(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t\right) q\left(\mathbf{x}_{1:t-1}^{(r)} \mid \mathbf{y}_{1:t-1}\right)} \tag{3.48}$$

$$= w_{t-1}^{(r)} \frac{p\left(\mathbf{y}_t \mid \mathbf{x}_t^{(r)}\right) p\left(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}\right)}{q\left(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t\right)} \tag{3.49}$$

The algorithm for SIS is shown in Algorithm 4 below. The reason why it works is the

Algorithm 4 Sequential importance sampling

1: Initialise weights $\left\{w_0^{(r)} = \frac{1}{R}\right\}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Observe $\mathbf{y}_t$.
4: Sample $\left\{\mathbf{x}_{1:t}^{(r)}\right\}$ from $q\left(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t}\right)$.
5: Calculate weights $\left\{w_t^{(r)}\right\}$ according to (3.49).
6: Calculate normalised weights $\left\{\hat{w}_t^{(r)} = \frac{w_t^{(r)}}{\sum_{r'} w_t^{(r')}}\right\}$.
7: ▷ The pmf $\sum_r \hat{w}_t^{(r)} \delta_{\mathbf{x}_{1:t}^{(r)}}(\mathbf{x}_{1:t})$ approximates the pdf $p(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t})$. Hence we can approximate the pdf $p(\mathbf{x}_t \mid \mathbf{y}_{1:t})$ by $\sum_r \hat{w}_t^{(r)} \delta_{\mathbf{x}_t^{(r)}}(\mathbf{x}_t)$.

same as in the case of Sampling importance resampling described in section 3.4.

3.5.2 The degeneracy problem

Because the support of the pdf we are approximating ($p(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t})$) is growing, the constant number of weights we use (R) won't be sufficient after a while. This is because many weights will become very negligible, wasting our resources. An **effective sample size** is used to measure this degeneracy is defined to be and approximated by the following:

$$S_{\text{eff}} \triangleq \frac{S}{1 + \text{var}\left[w_t^{(r)*}\right]} \tag{3.50}$$

$$\hat{S}_{\text{eff}} \approx \frac{1}{\sum_r \left(w_t^{(r)}\right)^2} \tag{3.51}$$

where $w_t^{(r)^*} = p(\mathbf{x}_t^{(r)} \mid \mathbf{y}_{1:t})/q(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t)$ is the "true weight" of particle r.

There are (among others) two solutions to this problem – introduce the resampling step, and using a good proposal distribution.

3.5.3 The resampling step

Whenever the effective sample size drops below some threshold, resample to get new R samples from the approximation of the pdf. This step is also called **rejuvenation**. The full algorithm for a generic particle filter is shown in Algorithm 5 below.

Algorithm 5 Generic particle filter

Initialise weights $\left\{ w_0^{(r)} = \frac{1}{R} \right\}$.
for $t = 1, \ldots, T$ **do**
    Observe $\mathbf{y}_t$.
    Sample $\left\{ \mathbf{x}_{1:t}^{(r)} \right\}$ from $q\left(\mathbf{x}_{1:t} \mid \mathbf{y}_{1:t}\right)$.
    Calculate weights $\left\{ w_t^{(r)} \right\}$ according to (3.49).
    Calculate normalised weights $\left\{ \hat{w}_t^{(r)} = \frac{w_t^{(r)}}{\sum_{r'} w_t^{(r')}} \right\}$.
    Calculate the effective sample size, $\hat{S}_{\text{eff}}$, according to (3.51).
    if $\hat{S}_{\text{eff}} < S_{\min}$ **then**
        Resample R particles, $\left\{ \mathbf{x}_t^{(r)} \right\}$ from the pmf $\sum_r \hat{w}_t^{(r)} \delta_{\mathbf{x}_t^{(r)}}(\mathbf{x}_t)$.
        Reassign $w_t^{(r)} = \frac{1}{R}$ for $r = 1, \ldots, R$.

3.5.4 Particle filter animation

Particle filter

3.5.5 The proposal distribution

It is common to use the following proposal distribution

$$
\begin{aligned}
q(\mathbf{x}_{1:t}^{(r)} \mid \mathbf{y}_{1:t}) &= q(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t) && (3.52)\\
&= p(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}) && (3.53)
\end{aligned}
$$

Hence the weight equation in (3.49) becomes

$$
\begin{aligned}
w_t^{(r)} &\propto w_{t-1}^{(r)} \frac{p\left(\mathbf{y}_t \mid \mathbf{x}_t^{(r)}\right) p\left(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}\right)}{q\left(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t\right)} && (3.54)\\
&= w_{t-1}^{(r)} p\left(\mathbf{y}_t \mid \mathbf{x}_t^{(r)}\right) && (3.55)
\end{aligned}
$$

This approach can be **i**nefficient because the likelihood, $p\left(\mathbf{y}_t \mid \mathbf{x}_t^{(r)}\right)$, can be very small at many places meaning many of the particles will be very small.

The optimal proposal distribution has the form

$$
\begin{aligned}
q(\mathbf{x}_{1:t}^{(r)} \mid \mathbf{y}_{1:t}) &= q(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t) && (3.56)\\
&= p(\mathbf{x}_t^{(r)} \mid \mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t) && (3.57)\\
&= \frac{p(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{x}_{t-1}^{(r)}) p(\mathbf{x}_t, \mathbf{x}_{t-1}^{(r)})}{p(\mathbf{x}_{t-1}^{(r)}, \mathbf{y}_t)} && (3.58)\\
&= \frac{p(\mathbf{y}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(r)})}{p(\mathbf{y}_t \mid \mathbf{x}_{t-1}^{(r)})} && (3.59)
\end{aligned}
$$

The weight equation in (3.49) becomes

$$
\begin{aligned}
w_t^{(r)} &\propto w_{t-1}^{(r)} p\left(\mathbf{y}_t \mid \mathbf{x}_{t-1}^{(r)}\right) && (3.60)\\
&= w_{t-1}^{(r)} \int p(\mathbf{y}_t, \mathbf{x}_t' \mid \mathbf{x}_{t-1}^{(r)})\, \mathrm{d}\mathbf{x}' && (3.61)\\
&= w_{t-1}^{(r)} \int p(\mathbf{y}_t \mid \mathbf{x}_t', \mathbf{x}_{t-1}^{(r)}) p(\mathbf{x}_t' \mid \mathbf{x}_{t-1}^{(r)})\, \mathrm{d}\mathbf{x}' && (3.62)\\
&= w_{t-1}^{(r)} \int p(\mathbf{y}_t \mid \mathbf{x}_t') p(\mathbf{x}_t' \mid \mathbf{x}_{t-1}^{(r)})\, \mathrm{d}\mathbf{x}' && (3.63)
\end{aligned}
$$

The proposal distribution is optimal because for any fixed $\mathbf{x}_{t-1}^{(r)}$, the new weight $w_t^{(r)}$ takes the same value regardless of the value drawn for $\mathbf{x}_t^{(r)}$. Hence, conditional on the old values, the variance of true weights is zero.

3.6 Sequential Monte Carlo

(to do: improve to be more rigorous)

Assume that at time t, we can extend a particle's path using a Markov kernel M_t:

$$p_t(x_t) = p_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t) \tag{3.64}$$

Also assume that

$$\tilde{p}_t(x_{0:t}) = p_t(x_t) \sum_{k=1}^{t} L_k(x_k, x_{k-1}) \tag{3.65}$$

where $\{L_k\}$ is a sequence of auxiliary Markov transition kernels.

The generic algorithm for Sequential Monte Carlo (SMC) can be found in Algorithm 6.

Algorithm 6 Generic Sequential Monte Carlo

Initialisation, $t = 0$:
for $r = 1, \ldots, R$ **do** ▷ Sample.
    Sample $\tilde{x}_0^{(r)} \sim q_0(\cdot)$.
for $r = 1, \ldots, R$ **do**
    Calculate normalised weights $\hat{w}_0^{(r)} \propto \frac{p_0\left(\tilde{x}_0^{(r)}\right)}{q_0\left(\tilde{x}_0^{(r)}\right)}$, such that $\sum'_r \hat{w}_0^{(r')} = 1$.
Resample from the pmf $\sum_r \hat{w}_0^{(r)} \delta_{\tilde{x}_0^{(r)}}(\cdot)$ to get R samples $\left\{x_0^{(r)}\right\}$. ▷ Resample.

Iterate, $t = 1, \ldots, T$:
for $t = 1, \ldots, T$ **do**
    for $r = 1, \ldots, R$ **do** ▷ Sample.
        Set $\tilde{x}_{0:t-1}^{(r)} = x_{0:t-1}^{(r)}$.
        Sample $\tilde{x}_t^{(r)} \sim M_t\left(\tilde{x}_{0:t-1}^{(r)}, \cdot\right)$.
    for $r = 1, \ldots, R$ **do**
        Calculate normalised weights $\hat{w}_t^{(r)} \propto \frac{p_t(x_t) L_t(x_t, x_{t-1})}{p_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t)}$.
    Resample from the pmf $\sum_r \hat{w}_t^{(r)} \delta_{\tilde{x}_t^{(r)}}(\cdot)$ to get R samples $\left\{x_t^{(r)}\right\}$. Reset the weights to $1/R$. ▷ Resample.

3.7 Markov chain Monte Carlo methods

3.7.1 Definitions

Definition 3.7.1. Markov chain *(MC) is defined via a state space* $\mathcal{X}$ *and a model that defines, for every state* $\mathbf{x} \in \mathcal{X}$ *a next-state distribution over* $\mathcal{X}$*. More precisely, the transition model* $\mathcal{T}$ *specifies for each pair of state* $\mathbf{x}, \mathbf{x}'$ *the probability* $\mathcal{T}(\mathbf{x} \to \mathbf{x}')$ *of* *going from* $\mathbf{x}$ *to* $\mathbf{x}'$*, i.e.* $\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \Pr(\mathbf{x}' \mid \mathbf{x})$*. This transition probability applies whenever the chain is in state* $\mathbf{x}$*.*

If the MCMC generates a sequence of states $\mathbf{x}_0, \ldots, \mathbf{x}_T$, the state at time t, $\mathbf{x}_t$ can be viewed as a random variable $\mathbf{X}_t$ for $t = 1, \ldots, T$.

Theorem 3.7.1 (Ergodic Theorem for MC (simplified))**.** *If* $(\mathbf{X}_0, \ldots, \mathbf{X}_T)$ *is an* irreducible*,* time-homogeneous *discrete space MC with* stationary distribution π*, then*

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{X}_t) \xrightarrow[n\to\infty]{a.s.} \mathrm{E}[f(\mathbf{X})] \qquad \text{where } \mathbf{X} \sim \pi \tag{3.66}$$

for any bounded function $f : \mathcal{X} \mapsto \mathbb{R}$*.*

If further, it is aperiodic*, then*

$$\Pr(\mathbf{X}_T = \mathbf{x} \mid \mathbf{X}_0 = \mathbf{x}_0) \xrightarrow[n\to\infty]{} \pi(\mathbf{x}) \qquad \forall\, \mathbf{x}, \mathbf{x}_0 \in \mathcal{X}. \tag{3.67}$$

A MC following these conditions is ergodic

Definition 3.7.2. *A MC* $(\mathbf{X}_t)$ *is* time-homogeneous *if* $\Pr(\mathbf{X}_{t+1} = b \mid \mathbf{X}_t = a) = \mathcal{T}(a \to b)\ \forall\, t \in \{1, \ldots, T-1\}\ \forall\, a, b \in \mathcal{X}$ *for some kernel function* $\mathcal{T}$*.*

Definition 3.7.3. *A pmf* π *on* $\mathcal{X}$ *is a* stationary (invariant) distribution *(w.r.t.* $\mathcal{T}$*) if*

$$\pi(\mathbf{X} = \mathbf{x}') = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{X} = \mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x}') \qquad \forall\, \mathbf{x}' \tag{3.68}$$

Definition 3.7.4. *A MC* $(\mathbf{X}_t)$ *is* irreducible *if* $\forall\, a, b \in \mathcal{X}\ \exists\, t \geq 0$ *s.t.* $\Pr(\mathbf{X}_t = b \mid \mathbf{X}_0 = a) > 0$*.*

Definition 3.7.5. *An* irreducible *MC* $(\mathbf{X}_t)$ *is* aperiodic *if* $\forall\, a \in \mathcal{X}$*,*

$$\gcd\{t : \Pr(\mathbf{X}_t = a \mid \mathbf{X}_0 = a) > 0\} = 1. \tag{3.69}$$

Definition 3.7.6. *A MC is* regular *if there exists some number* k *such that, for every* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*, the probability of getting from* $\mathbf{x}$ *to* $\mathbf{x}'$ *in exactly* k *steps is* > 0*.*

Theorem 3.7.2. *If a finite state MC described by* $\mathcal{T}$ *is* regular*, then it has a unique* stationary distribution*.*

A MC being *ergodic* is equivalent to it being *regular* [1, p. 510].

Definition 3.7.7. *A finite state MC described by* $\mathcal{T}$ *is* reversible *if there exists a unique distribution* π *such that, for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$

$$\pi(\mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \pi(\mathbf{x}')\mathcal{T}(\mathbf{x}' \to \mathbf{x}). \tag{3.70}$$

This equation is called the detailed balance *(DB).*

Proposition 3.7.1. *If a finite state MC described by $\mathcal{T}$ is* regular *and satisfies the* detailed balance *equation relative to π, then π is the* unique stationary distribution *of $\mathcal{T}$.*

Proof. Assuming the DB equation (3.70), we want to prove the stationarity equation (3.68) to ensure π is a stationary distribution of $\mathcal{T}$. We have

$$\sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x}')\mathcal{T}(\mathbf{x}' \to \mathbf{x}) \tag{3.71}$$

$$= \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x}')\Pr(\mathbf{x} \mid \mathbf{x}') \tag{3.72}$$

$$= \pi(\mathbf{x}')\sum_{\mathbf{x}\in\mathcal{X}} \Pr(\mathbf{x} \mid \mathbf{x}') \tag{3.73}$$

$$= \pi(\mathbf{x}') \tag{3.74}$$

which proves the equation (3.68). π is the unique stationary distribution of $\mathcal{T}$ because of Theorem 3.7.2. □

Proposition 3.7.2. *Let $\mathcal{T}_1, \dots, \mathcal{T}_K$ be a set of kernels each of which satisfies detailed balance w.r.t. π. Let $p_1, \dots, p_K$ be any distribution over $\{1, \dots, K\}$. The mixture MC $\mathcal{T}$, which at each step takes a step sampled from $\mathcal{T}_k$ with probability p_k also satisfies the detailed balance equation relative to π.*

Proof. The aggregate kernel can be written as

$$\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \Pr(\mathbf{x}' \mid \mathbf{x}) \tag{3.75}$$

$$= \sum_k \Pr(\mathbf{x}', k \mid \mathbf{x}) \tag{3.76}$$

$$= \sum_k \Pr(\mathbf{x}' \mid k, \mathbf{x})\Pr(k \mid \mathbf{x}) \tag{3.77}$$

$$= \sum_k \mathcal{T}_k(\mathbf{x} \to \mathbf{x}')p_k \tag{3.78}$$

Using this, we can prove the detailed balance as follows

$$\pi(\mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \pi(\mathbf{x})\sum_k \mathcal{T}_k(\mathbf{x} \to \mathbf{x}')p_k \tag{3.79}$$

$$= \sum_k \pi(\mathbf{x})\mathcal{T}_k(\mathbf{x} \to \mathbf{x}')p_k \tag{3.80}$$

$$= \sum_k \pi(\mathbf{x}')\mathcal{T}_k(\mathbf{x}' \to \mathbf{x})p_k \tag{3.81}$$

$$= \pi(\mathbf{x}')\sum_k \mathcal{T}_k(\mathbf{x}' \to \mathbf{x})p_k \tag{3.82}$$

$$= \pi(\mathbf{x}')\mathcal{T}(\mathbf{x}' \to \mathbf{x}) \tag{3.83}$$

□

Proposition 3.7.3. *Let $\mathcal{T}_1, \ldots, \mathcal{T}_K$ be a set of kernels each of which satisfies detailed balance w.r.t. π. The aggregate MC, $\mathcal{T}$, where each step consists of a sequence of K steps, with step k being sampled from $\mathcal{T}_k$ has π as its stationary distribution.*

Proof. The aggregate kernel can be written as

$$\begin{aligned}
\mathcal{T}(\mathbf{x} \to \mathbf{x}') &= \Pr(\mathbf{x}' \mid \mathbf{x}) && (3.84)\\
&= \sum_{\mathbf{x}_{1:K-1}} \Pr(\mathbf{x}', \mathbf{x}_{K-1}, \ldots, \mathbf{x}_1 \mid \mathbf{x}) && (3.85)\\
&= \sum_{\mathbf{x}_{1:K-1}} \Pr(\mathbf{x}_K, \ldots, \mathbf{x}_1 \mid \mathbf{x}_0) && (3.86)\\
&= \sum_{\mathbf{x}_{1:K-1}} \Pr(\mathbf{x}_1 \mid \mathbf{x}_0) \cdots \Pr(\mathbf{x}_K \mid \mathbf{x}_{K-1}) && (3.87)\\
&= \sum_{\mathbf{x}_{1:K-1}} \mathcal{T}_1(\mathbf{x}_0 \to \mathbf{x}_1) \cdots \mathcal{T}_K(\mathbf{x}_{K-1} \to \mathbf{x}_K) && (3.88)
\end{aligned}$$

where we've used the substitution $\mathbf{x} = \mathbf{x}_0$ and $\mathbf{x}' = \mathbf{x}_K$. Using this, we can prove that π is the stationary distribution as follows

$$\begin{aligned}
\sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) \mathcal{T}(\mathbf{x} \to \mathbf{x}') &= \sum_{\mathbf{x}_0} \pi(\mathbf{x}_0) \sum_{\mathbf{x}_{1:K-1}} \mathcal{T}_1(\mathbf{x}_0 \to \mathbf{x}_1) \cdots \mathcal{T}_K(\mathbf{x}_{K-1} \to \mathbf{x}_K) && (3.89)\\
&= \sum_{\mathbf{x}_{0:K-1}} \pi(\mathbf{x}_0) \mathcal{T}_1(\mathbf{x}_0 \to \mathbf{x}_1) \cdots \mathcal{T}_K(\mathbf{x}_{K-1} \to \mathbf{x}_K) && (3.90)\\
&= \sum_{\mathbf{x}_{0:K-1}} \mathcal{T}_1(\mathbf{x}_1 \to \mathbf{x}_0) \pi(\mathbf{x}_1) \cdots \mathcal{T}_K(\mathbf{x}_{K-1} \to \mathbf{x}_K) && (3.91)\\
&\cdots\\
&= \sum_{\mathbf{x}_{0:K-1}} \mathcal{T}_1(\mathbf{x}_1 \to \mathbf{x}_0) \cdots \mathcal{T}_K(\mathbf{x}_K \to \mathbf{x}_{K-1}) \pi(\mathbf{x}_K) && (3.92)\\
&= \pi(\mathbf{x}_K) \sum_{\mathbf{x}_{0:K-1}} \mathcal{T}_K(\mathbf{x}_K \to \mathbf{x}_{K-1}) \cdots \mathcal{T}_1(\mathbf{x}_1 \to \mathbf{x}_0) && (3.93)\\
&= \pi(\mathbf{x}_K) \sum_{\mathbf{x}_{0:K-1}} \Pr(\mathbf{x}_{0:K-1} \mid \mathbf{x}_K) && (3.94)\\
&= \pi(\mathbf{x}_K). && (3.95)
\end{aligned}$$

□

3.7.2 Metropolis Hastings algorithm

The Metropolis Hastings (MH) algorithm is a recipe to create a MCMC with a particular stationary distribution. Assume we can sample from a proposal distribution $q(\cdot \mid \mathbf{x}) \equiv q(\mathbf{x} \to \cdot)$. Let $p \equiv \pi$ be the required distribution (stationary distribution for this MCMC). Assume we can only evaluate q and π up to a multiplicative factor (i.e. we can only evaluate $q^*(\mathbf{x} \to \mathbf{x}') = Z_q q(\mathbf{x} \to \mathbf{x}')$ and $\pi^*(\mathbf{x}) = Z_p \pi(\mathbf{x})$). The MH algorithm is outlined in Algorithm 7.

Algorithm 7 Metropolis Hastings algorithm

1: Sample $\mathbf{x}^{(0)}$ from an arbitrary probability distribution over $\mathcal{X}$.
2: **for** $t = 1, \ldots, T$ **do**
3: **repeat**
4: Sample $\mathbf{x}^{(t)} \sim q(\mathbf{x}^{(t-1)} \to \cdot)$.
5: Accept $\mathbf{x}^{(t)}$ with the acceptance probability

$$\mathcal{A}(\mathbf{x}^{(t-1)} \to \mathbf{x}^{(t)}) = \min\left(1, \frac{\pi^*(\mathbf{x}^{(t)})q^*(\mathbf{x}^{(t)} \to \mathbf{x}^{(t-1)})}{\pi^*(\mathbf{x}^{(t-1)})q^*(\mathbf{x}^{(t-1)} \to \mathbf{x}^{(t)})}\right) \tag{3.96}$$

6: **until** $\mathbf{x}^{(t)}$ is accepted.

Why it works?

We need to prove that π is the unique stationary distribution of this MCMC.

We can express the aggregate transition model to be

$$\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \begin{cases} q(\mathbf{x} \to \mathbf{x}')\mathcal{A}(\mathbf{x} \to \mathbf{x}') & \text{if } \mathbf{x} \neq \mathbf{x}' \\ q(\mathbf{x} \to \mathbf{x}) + \sum_{\mathbf{x}', \mathbf{x}' \neq \mathbf{x}} q(\mathbf{x} \to \mathbf{x}')(1 - \mathcal{A}(\mathbf{x} \to \mathbf{x}')) & \text{if } \mathbf{x} = \mathbf{x}' \end{cases} \tag{3.97}$$

To prove that π is a stationary distribution of this MCMC, we make sure the DB equation holds.

For $\mathbf{x} \neq \mathbf{x}'$, we have

$$\pi(\mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \pi(\mathbf{x})q(\mathbf{x} \to \mathbf{x}')\min\left(1, \frac{\pi(\mathbf{x}')q(\mathbf{x}' \to \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x} \to \mathbf{x}')}\right) \tag{3.98}$$

$$= \min\left(\pi(\mathbf{x})q(\mathbf{x} \to \mathbf{x}'), \pi(\mathbf{x}')q(\mathbf{x}' \to \mathbf{x})\right) \tag{3.99}$$

$$= \pi(\mathbf{x}')q(\mathbf{x}' \to \mathbf{x})\min\left(1, \frac{\pi(\mathbf{x})q(\mathbf{x} \to \mathbf{x}')}{\pi(\mathbf{x}')q(\mathbf{x}' \to \mathbf{x})}\right) \tag{3.100}$$

$$= \pi(\mathbf{x}')\mathcal{T}(\mathbf{x}' \to \mathbf{x}) \tag{3.101}$$

For $\mathbf{x} = \mathbf{x}'$, the DB equation $\pi(\mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x}') = \pi(\mathbf{x}')\mathcal{T}(\mathbf{x}' \to \mathbf{x})$ obviously holds.

Hence π is a stationary distribution of the MCMC described via $\mathcal{T}$. Unfortunately, regularity doesn't hold in general. We need to make sure our created MCMC is regular before we can claim that π is the unique stationary distribution of this MCMC.

3.7.3 Gibbs sampling

Assume we want to sample from $p(\mathbf{x}) = p(x_1, \ldots, x_D)$. We can only sample from the conditionals $p(x_i \mid \mathbf{x}_{-i})$ where $\mathbf{x}_{-i}$ denotes $\mathbf{x}$ with the i^{th} component ommited. The Gibbs sampling algorithm (8) is given below.

Algorithm 8 Gibbs sampling algorithm

1: Sample $\mathbf{x}^{(0)}$ from an arbitrary probability distribution over $\mathcal{X}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $x_1^{(t)} \sim p\left(\cdot \mid x_2^{(t-1)}, x_3^{(t-1)}, \ldots, x_D^{(t-1)}\right)$
4: Sample $x_2^{(t)} \sim p\left(\cdot \mid x_1^{(t)}, x_3^{(t-1)}, \ldots, x_D^{(t-1)}\right)$
5: $\vdots$
6: Sample $x_D^{(t)} \sim p\left(\cdot \mid x_1^{(t)}, x_2^{(t)}, \ldots, x_{D-1}^{(t)}\right)$

Why it works?

Each of the sampling steps can be viewed to be governed by a different kernel with the whole process being governed by the aggregate kernel. We prove that the single kernels follow the DB equation with respect to p:

$$\begin{aligned} p(\mathbf{x})\mathcal{T}_i(\mathbf{x} \to \mathbf{x}') &= p(\mathbf{x})p(\mathbf{x}_{-i}, x_i' \mid \mathbf{x}) && (3.102) \\ &= p(\mathbf{x}_{-i}, x_i', \mathbf{x}) && (3.103) \\ &= p(\mathbf{x}, x_i', \mathbf{x}_{-i}) && (3.104) \\ &= p(\mathbf{x}')p(\mathbf{x} \mid x_i', \mathbf{x}_{-i}) && (3.105) \\ &= p(\mathbf{x}')\mathcal{T}_i(\mathbf{x}' \to \mathbf{x}) && (3.106) \end{aligned}$$

This is the premise of Proposition 3.7.3, hence the aggregate kernel $\mathcal{T}$ has p as its stationary distribution.

We can also view Gibbs sampling as an instance of the MH algorithm. If the proposal of MH $q_i(\mathbf{x} \to \mathbf{x}')$ is set to be $p(\mathbf{x}' \mid \mathbf{x}) = p(x_i' \mid \mathbf{x})$ the acceptance probability is one (shown below) and so it is equivalent to one sampling step in Gibbs sampling.

$$\begin{aligned} \mathcal{A}(\mathbf{x} \to \mathbf{x}') &= \min\left(1, \frac{p(\mathbf{x}')p(\mathbf{x} \mid \mathbf{x}')}{p(\mathbf{x})p(\mathbf{x}' \mid \mathbf{x})}\right) && (3.107) \\ &= \min\left(1, \frac{p(\mathbf{x}', \mathbf{x})}{p(\mathbf{x}', \mathbf{x})}\right) && (3.108) \\ &= 1 && (3.109) \end{aligned}$$

Bibliography

[1] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.