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

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

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

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

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

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

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

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

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

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

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

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

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

Evidence

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

2.5.2 Derivations

3 Sampling algorithms

3.1 Introduction

Let p be a probability distribution with pdf $p(\mathbf{x})$, which we assume can be evaluated only up to a constant of proportionality (i.e. we can only evaluate $p^*(\mathbf{x}) = Z_p p(\mathbf{x})$, where $Z_p = \int p^*(\mathbf{x})\mathrm{d}\mathbf{x}$). We want to achieve the following:

Problem 1 Generate samples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(R)}\}$ (shorthand notation $\{\mathbf{x}^{(r)}\}$) from the probability distribution p.

Problem 2 Estimate the expectation of an arbitrary function $f(\mathbf{X})$ given that $\mathbf{X} \sim p$, $\mathrm{E}[f(\mathbf{X})]$.

3.2 Importance sampling

Assume that we can sample from a proposal distribution q with a pdf $q(\mathbf{x})$, which can be evaluated only up to a constant of proportionality (i.e. we can only evaluate $q^*(\mathbf{x}) = Z_q q(\mathbf{x})$). To solve problem 2, we follow

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 $\mathrm{E}[f(\mathbf{X})]$.

3.2.1 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 $\mathrm{E}[f(\mathbf{X})]$, as R increases. We consider the the expectations of the numerator and denominator separately:

$$\begin{aligned}
\mathrm{E}_q[\text{numer}] &= \mathrm{E}_q\left[\sum_r w_r f(\mathbf{x}^{(r)})\right] && (3.1)\\
&= \sum_r \mathrm{E}_q\left[w_r f(\mathbf{x}^{(r)})\right] && (3.2)\\
&= \sum_r \mathrm{E}_q\left[\frac{p^*(\mathbf{x}^{(r)})}{q^*(\mathbf{x}^{(r)})} f(\mathbf{x}^{(r)})\right] && (3.3)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_r \mathrm{E}_q\left[\frac{Z_p p(\mathbf{x}^{(r)})}{Z_q q(\mathbf{x}^{(r)})} f(\mathbf{x}^{(r)})\right] && (3.4)\\
&= \frac{Z_p}{Z_q}\sum_r \int_{\mathbf{x}^{(r)}} p(\mathbf{x}^{(r)}) f(\mathbf{x}^{(r)})\,\mathrm{d}\mathbf{x}^{(r)} && (3.5)\\
&= \frac{Z_p}{Z_q}\sum_r \mathrm{E}_p\left[f(\mathbf{x}^{(r)})\right] && (3.6)\\
&= \frac{Z_p}{Z_q} R\,\mathrm{E}_p\left[f(\mathbf{x})\right] && (3.7)\\
\mathrm{E}_q[\text{denom}] &= \mathrm{E}_q\left[\sum_r w_r\right] && (3.8)\\
&= \sum_r \mathrm{E}_q\left[\frac{p^*(\mathbf{x}^{(r)})}{q^*(\mathbf{x}^{(r)})}\right] && (3.9)\\
&= \sum_r \mathrm{E}_q\left[\frac{Z_p p(\mathbf{x}^{(r)})}{Z_q q(\mathbf{x}^{(r)})}\right] && (3.10)\\
&= \frac{Z_p}{Z_q}\sum_r \int_{\mathbf{x}^{(r)}} p(\mathbf{x}^{(r)})\,\mathrm{d}\mathbf{x}^{(r)} && (3.11)\\
&= \frac{Z_p}{Z_q} R && (3.12)
\end{aligned}
$$

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

3.2.2 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

$$
\begin{aligned}
\mathrm{var}_q\left[\frac{p(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x})\right] &= \mathrm{E}_q\left[\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})} f^2(\mathbf{x})\right] - \left(\mathrm{E}_q\left[\frac{p(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x})\right]\right)^2 && (3.13)\\
&= \mathrm{E}_q\left[\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})} f^2(\mathbf{x})\right] - \left(\mathrm{E}_p\left[f(\mathbf{x})\right]\right)^2 && (3.14)
\end{aligned}
$$

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:

$$
\mathrm{E}_q\left[\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})} f^2(\mathbf{x})\right] \geq \left(\mathrm{E}_q\left[\frac{p(\mathbf{x})}{q(\mathbf{x})}|f(\mathbf{x})|\right]\right)^2 = \left(\mathrm{E}_p[|f(\mathbf{x})|]\right)^2 \tag{3.15}
$$

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

$$
q_{\text{optimal}}(\mathbf{x}) = \frac{|f(\mathbf{x})|p(\mathbf{x})}{\int |f(\mathbf{x}')|p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'} \tag{3.16}
$$