Chapter 2

1. (A twist on exercise 2.6). Consider 11 urns $u = 0, \ldots, 10$ each with 10 balls. Urn $u$ has $u$ black balls and $10 - u$ white balls. Select one urn at random, and draw $N$ times with replacement from that urn. Suppose that the outcome after $N = 10$ draws is that the number of black balls that have been drawn is even. What is the probability that urn $u$ was selected?

2. (Inference on the parameters of a Gaussian distribution). The Gaussian distribution is the probability density of a continuous variable $x \in \mathbb{R}$ and has the form

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$\mu$ is the peak of the Gaussian distribution and also the expected value of $x$: $\mu = \mathbb{E}x = \int dx p(x|\mu, \sigma)$, where $\mathbb{E}$ denotes expected value. $\sigma^2$ is the variance of $x$: $\sigma^2 = \mathbb{E}(x - \mu)^2 = \int dx (x - \mu)^2 p(x|\mu, \sigma)$.

Suppose that we are told that data are drawn from an oracle that is a Gaussian distribution with $\sigma = 1$ and we believe that $\mu$ can have any value with equal probability.

(a) The oracle produces one data point with value $x$. What can we infer about $\mu$?

(b) The oracle produces a set of $N$ data points with values $x_1, \ldots, x_N$. What can we infer about $\mu$? Hint: use the fact that the data points are produced independently. What is the probability that the data set is drawn from the oracle? Show that the result is Gaussian in $\mu$ with mean $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ and variance $\frac{1}{N}$.

3. (Maximum entropy distributions). Consider the $n$-dimensional multivariate Gaussian distribution

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

with $x, \mu \in \mathbb{R}^n$ and $\Sigma$ a positive definite $n \times n$ matrix.

(a) Show that this distribution is an exponential family distribution.

(b) The multivariate Gaussian distribution is the a maximum entropy solution. What are the constraints?

4. (Variational and reverse KL approximation). Consider a two-dimensional Gaussian distribution

$$p(z_1, z_2) \propto \exp\left(-\frac{1}{2} (az_1^2 + az_2^2 + 2bz_1z_2)\right)$$

$p$ has mean zero: $\mathbb{E}_p z_1 = \mathbb{E}_p z_2 = 0$ and covariance $\mathbb{E}_p z_1 z_j = (\Lambda^{-1})_{ij}$ with $\Lambda = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$. We wish to approximate $p$ by a simpler spherical Gaussian distribution

$$q(z_1, z_2) = q_1(z_1)q_2(z_2) \quad q_i(z_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2} z_i^2)$$

(Since $p$ has mean zero, we already know that we should take $q$ also mean zero). First, we compute the variational solution in the following steps.
(a) Show that \( \int dz_1 q_1(z_1) \log q_1(z_1) = - \log \sqrt{2\pi\sigma_1^2} - \frac{1}{2} \) and similar for \( q_2(z_2) \).

(b) Use this result to show that

\[
KL(q\|p) = -1 - \log \sqrt{2\pi\sigma_1^2} - \log \sqrt{2\pi\sigma_2^2} + \frac{1}{2}a(\sigma_1^2 + \sigma_2^2)
\]

(c) Show that the variational solution is given by \( q_i(z_i) \) with \( \sigma_i^2 = \frac{1}{a} \).

(d) Now we compute the reverse solution. Show that reverse variational solution is given by \( q_i(z_i) \) with \( \sigma_i^2 = \mathbb{E}_p z_i^2 \), \( i = 1, 2 \).

(e) Show that \( \mathbb{E}_p z_1^2 = \mathbb{E}_p z_2^2 = \frac{a}{a^2 - b^2} \). Thus the variance of the variational solution is less that the (correct) variance of the reverse variational solution:

\[
\frac{1}{a} < \frac{a}{a^2 - b^2}
\]

unless \( b = 0 \), in which case \( p \) itself is a factorized distribution.

**Chapter 3**

1. Consider the bent coin model comparison example of Mackay section 3.2 with \( N = 2 \).

   (a) Compute the posterior probability of the two models \( H_{0,1} \) for \( N_H = 0, 1, 2 \).

   (b) You will find that for \( N_H = 0, 2 \) model \( H_1 \) is more likely and for \( N_H = 1 \) model \( H_0 \) is more likely. Explain these results.

**Chapter 27**

1. (Adapted from Mackay 27.1). A photon counter is pointed at a remote star for one minute, in order to infer the rate of photons arriving at the counter per minute, \( \lambda \). Assuming the number of photons collected \( r \) has a Poisson distribution with mean \( \lambda \)

\[
p(r|\lambda) = \exp(-\lambda) \frac{\lambda^r}{r!}
\]

and assuming the improper prior \( p(\lambda) = \frac{1}{\lambda} \).

   (a) compute the Laplace approximations \( p_1(\lambda) \) to the posterior distribution

   (b) Consider the transformation to new coordinate \( y = \log \lambda \). show that the prior distribution over \( y \) transforms to \( p(y) = 1 \). Hint: When changing variables, the distribution changes as \( p(\lambda)d\lambda = p(y)dy \).

   (c) Transform the posterior distribution to the new variable \( y \) and compute the Laplace approximation \( p_2(y) \).

   (d) Transform \( p_2(y) \) back to \( p_2(\lambda) \). Plot the posterior \( p(\lambda|r) \) and its two approximations \( p_1(\lambda), p_2(\lambda) \) for \( r = 2 \) and \( r = 10 \) versus \( \lambda \). Hint: Devide all distributions by their sum to get comparable magnitudes not not have to worry about explicit normalization. It should look like this
Chapter 28

1. (Variant on Mackay Ex. 28.2) Data points \( D = (x_i, t_i), i = 1, \ldots, N \) are believed to come from a straight line with noise. The experimenter chooses \( x_i \) and assumes that \( t_i \) is Gaussian distributed about \( w_0 + w_1 x_i \) with variance \( \sigma^2 \). According to model \( H_1 \), the straight line is horizontal, so \( w_1 = 0 \). According to model \( H_2 \), \( w_1 \) is a parameter with prior distribution a Normal Gaussian distribution \( N(w_1 | 0, 1) \). Both models assign a prior distribution \( N(w_0 | 0, 1) \) to \( w_0 \). Given the data, what is the evidence for each model? You may assume that the input data \( x_i \) are 'centered' such that \( \sum_{i=1}^{N} x_i = 0 \). Proceed in the following way.

(a) Write an expression for \( p(D|w_0, w_1) \). Note, that the result is a product of a function of \( w_0 \) and a function of \( w_1 \).

(b) You can now compute the odds ratio \( p(D|H_2)/p(D|H_1) \). Show the result as a function of the input variance \( \langle x^2 \rangle = \frac{1}{N} \sum_i x_i^2 \), the input-output correlation \( \langle xt \rangle = \frac{1}{N} \sum_i x_i t_i \) and \( N \). Hint: make use of the fact that the integrals over \( w_0, w_1 \) factorize. As a result many terms cancel in the odds ratio \( p(D|H_2)/p(D|H_1) \) and you dont have to compute some of the integrals.

(c) Show that in the limit of large \( N \) the simpler model \( H_1 \) is preferred when the correlations \( \langle xt \rangle^2 \lesssim \frac{\log N}{N} \) and the complex model is preferred otherwise. Use \( \sigma^2 = \langle x \rangle^2 = 1 \).

2. (Variant on Mackay Ex. 28.3) A \( k \) sided die is rolled \( n \) times and the outcomes are \( D = (n_1, \ldots, n_k) \) where \( n_i \) is the number of times the die lands with side \( i \) up, \( i = 1, \ldots, k \) and \( n = \sum_{i=1}^{k} n_i \).

(a) Compute the probability of the observed data assuming that the die is perfectly fair (\( H_0 \)). Hint: the data \( D = (n_1, \ldots, n_k) \) specifies that we get \( n_1 \) times the outcome 1, etc. The data does not specify the order, just the frequencies. So the probability of the data is a combinatorial factor times the probability of a specific order. The combinatorial factor for a partition of \( n_1, \ldots, n_k \) of \( n \) with \( \sum_{i=1}^{k} n_i = n \) is \( \frac{n!}{n_1! \ldots n_k!} \).
(b) Compute the probability of the data under the alternative hypothesis \( \mathcal{H}_1 \) that the die has an unknown probability \( p_i \) for outcome \( i \) and that \( p_i \) is uniformly distributed. Hint: You will encounter the following integral 

\[
I = \int_0^1 dp_1 \cdots \int_0^1 p_k \prod_{i=1}^k f_i(p_i) \delta \left( \sum_{i=1}^k p_i - 1 \right).
\]

You will need to use that 

\[
I = (k-1)!I_1
\]

with 

\[
I_1 = \int_0^1 dp_1 \int_0^{p_1} dp_2 \cdots \int_0^{p_{k-1}} dp_k \prod_{i=1}^k f_i(p_i) \delta \left( \sum_{i=1}^k p_i - 1 \right) \equiv \int_{simplex} d\vec{p} \prod_{i=1}^k f_i(p_i) \delta \left( \sum_{i=1}^k p_i - 1 \right)
\]

the integral over the probability simplex. This latter form is used in the normalization MacKay Eq. 23.31.

(c) Consider \( k = 6, n = 30 \) and \((n_1, n_2, n_3, n_4, n_5, n_6) = (3, 3, 2, 2, 9, 11)\). Compute the posterior probability of the models \( \mathcal{H}_0, \mathcal{H}_1 \) assuming equal priors. Show that \( \mathcal{H}_1 \) is more likely. Instead, when \( n_1, n_2, n_3, n_4, n_5, n_6) = (5, 5, 5, 5, 5, 5) \) show that \( \mathcal{H}_0 \) is more likely.

Chapter 29

1. Show that \( \hat{\Phi} \) as given in eq. 29.6 is an unbiased estimator of \( \Phi \) as given by eq. 29.3 and that the variance in \( \hat{\Phi} \) decreases with \( R \).

Chapter 31

1. Show, that the Ising model Eq. 31.1 on the square lattice with ferromagnetic interactions (all \( J = 1 \) and \( H = 0 \)) has energy per spin -2.

2. Show analytically by dividing the spins into two groups, that the infinite lattice with anti-ferromagnetic interactions (all \( J = -1 \) and \( H = 0 \)) can be transformed into an equivalent ferromagnetic model.

3. Use the result of MacKay exercise 31.1c to explain the slope of \( F \) vs. \( T \) in Fig. 31.15.

Chapter 34

1. Derive Eq. 34.11.

2. Derive Eq. 34.12

3. Derive the ICA algorithm from the likelihood 34.9.