Contents

Chapter 1

Introduction

1.1 Curve Fitting

Problem 1

This can be solved by substituting the definition of:

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

into the error function and then taking the derivative.

$$\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^{N} \left(y(x, \mathbf{w}) - t_n\right)^2 \\
&= \frac{1}{2} \sum_{n=1}^{N} \left(\sum_{j=0}^{M} w_j x^j - t_n \right)^2 && \text{Substitute} \\
\frac{dE(\mathbf{w})}{dw_i} &= \sum_{n=1}^{N} \left(\left(\sum_{j=0}^{M} w_j x^j - t_n \right) x^i \right) && \text{Take the derivative} \\
0 &= \sum_{n=1}^{N} \left(\left(\sum_{j=0}^{M} w_j x^j - t_n \right) x^i \right) && \text{Set derivative to 0} \\
0 &= \sum_{n=1}^{N} \left(\sum_{j=0}^{M} w_j x^j x^i - t_n x^i \right) && \text{Set derivative to 0} \\
\sum_{n=1}^{N} t_n x^i &= \sum_{n=1}^{N} \sum_{j=0}^{M} w_j x^j x^i \\
\sum_{n=1}^{N} t_n x^i &= \sum_{n=1}^{N} \sum_{j=0}^{M} w_j x^{i+j}
\end{aligned}$$

Problem 2

This is solved in almost the same way we just have one additional term for the regularization so:

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}(y(x,\mathbf{w})-t_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x^j - t_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \end{aligned}$$

$$\frac{dE(\mathbf{w})}{dw_i} = \sum_{n=1}^{N}\left(\left(\sum_{j=0}^{M} w_j x^j - t_n\right)x^i\right) + \lambda w_i$$

$$\begin{aligned} -\lambda w_i &= \sum_{n=1}^{N}\left(\left(\sum_{j=0}^{M} w_j x^j - t_n\right)x^i\right) + \lambda w_i \\ -\lambda w_i &= \sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x^j x^i - t_n x^i\right) \\ \sum_{n=1}^{N} t_n x^i - \lambda w_i &= \sum_{n=1}^{N}\sum_{j=0}^{M} w_j x^j x^i \\ \sum_{n=1}^{N} t_n x^i - \lambda w_i &= \sum_{n=1}^{N}\sum_{j=0}^{M} w_j x^{i+j} \end{aligned}$$

1.2 Probability Theory

Problem 3 The Probability of Selecting an Apple can be decomposed as:

$$\begin{aligned} \mathcal{P}(apple) &= \mathcal{P}(apple, red) + \mathcal{P}(apple, blue) + prob(apple, green) \\ &= \mathcal{P}(apple|red)\mathcal{P}(red) + \mathcal{P}(apple|blue)\mathcal{P}(blue) + \mathcal{P}(apple|green)\mathcal{P} \\ &= (.3)(.2) + (.5)(.2) + (.3)(.6) \\ &= .34 \end{aligned}$$

The probability that observing an orange came from the green box

can be solved using Bayes rule:

$$\begin{aligned}\mathcal{P}(green|orange) &= \frac{\mathcal{P}(orange|green)\mathcal{P}(green)}{\mathcal{P}(orange)} \\ &= \frac{(.3)(.6)}{.66} \\ &= .27\end{aligned}$$

Problem 5

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] && \text{Distributive Law} \\ &= \mathbb{E}\left[X^2\right] + \mathbb{E}\left[-2X\mathbb{E}[X]\right] + \mathbb{E}\left[\mathbb{E}[X]^2\right] && \text{Linearity of } \mathbb{E}[X] \\ &= \mathbb{E}\left[X^2\right] + -2\mathbb{E}\left[X\mathbb{E}[X]\right] + \mathbb{E}[X]^2 && \mathbb{E}[\alpha X] = \alpha\mathbb{E}[X] \\ &= \mathbb{E}\left[X^2\right] + -2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && \mathbb{E}[X] \text{ is just another constan} \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

Problem 6

$\mathrm{Cov}[X, Y] = \mathbb{E}[X, Y] - \mathbb{E}[X]\mathbb{E}[Y]$ But because
$X \perp Y \Rightarrow \mathbb{E}[X, Y] = \mathbb{E}[X]\mathbb{E}[Y] \Rightarrow \mathrm{Cov}[X, Y] = 0$

Problem 7

$$
\begin{aligned}
I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right) dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right) dxdy \\
\\
&= \int_{0}^{\infty}\int_{0}^{2\pi} \exp\left(-\frac{1}{2\sigma^2}\left[r^2\cos^2(\theta) + r^2\sin^2(\theta)\right]\right) rd\theta dr \\
&= \int_{0}^{\infty}\int_{0}^{2\pi} \exp\left(-\frac{r^2}{2\sigma^2}\right) rd\theta dr \\
&= 2\pi\int_{0}^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) rdr \\
&= 2\pi\int_{0}^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}du \\
&= -2\pi\sigma^2 \exp\left(-\frac{u}{2\sigma^2}\right)|_0^{\infty} \\
&= 2\pi\sigma^2
\end{aligned}
$$

Now we just need to show that this normalizes the Gaussian. Take $y = x - \mu$ then

$$
\begin{aligned}
\mathcal{N}(x|\mu,\sigma^2) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy
\end{aligned}
$$

Problem 8

$$\begin{aligned}\mathbb{E}[x] &= \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma)x dx \\ &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (y+\mu) \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy \qquad y = x-\mu\end{aligned}$$

Now split this into the sum of two integrals. The second integral has the form:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mu \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy$$

Which is just a normalized Gaussian times μ so this is just μ. Now since we are trying to prove that his equals μ I'm fairly confident that the left integral will go to zero. Let's try and prove this.

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy$$

Well if we visualize this function it looks odd which would imply that the integral from $[-\infty, \infty]$ is 0. Let's prove that it is odd quickly. We want to show that $f(x) = -f(-x)$ where $f(x) = y \exp\left(-\frac{1}{2\sigma^2}y^2\right)$

$$-(-y)\exp\left(-\frac{1}{2\sigma^2}(-y)^2\right) = y\exp\left(-\frac{1}{2\sigma^2}y^2\right)$$

Thus or integral in question is odd and evaluates to 0 yielding μ as the desired result.

Next we need to show that:

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma)x^2 dx = \mu^2 + \sigma^2$$

To do this differentiate both sides by σ^2 as follows:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx &= 1 \\
\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx &= \sqrt{2\pi\sigma^2} \\
\int_{-\infty}^{\infty} \frac{d}{d\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx &= \frac{d}{d\sigma^2}\sqrt{2\pi\sigma^2} \\
\frac{1}{2\sigma^4} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) (x-\mu)^2 dx &= \frac{\pi}{\sqrt{(2\pi\sigma^2)}} \\
\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) (x-\mu)^2 dx &= \frac{2\pi\sigma^4}{\sqrt{(2\pi\sigma^2)}} \\
\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) (x-\mu)^2 dx &= \sigma^2 \\
\mathrm{Var}\left[x\right] &= \sigma^2
\end{aligned}$$

The left hand side of this equation is just the definition of variance. Now to complete the proof of (1.50) just use the alternate formulation of variance:

$$\begin{aligned}
\mathbb{E}\left[(x-\mu)^2\right] &= \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 \\
&= \mathbb{E}\left[x^2\right] - \mu^2 && \text{By part 1} \\
\sigma^2 &= \mathbb{E}\left[x^2\right] - \mu^2 && \text{By above} \\
\mathbb{E}\left[x^2\right] &= \sigma^2 + \mu^2
\end{aligned}$$

Problem 9

$$\begin{aligned}
\mathcal{N}(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\
\frac{d\mathcal{N}(x|\mu,\sigma^2)}{dx} &= -\frac{1}{\sigma^2}\mathcal{N}(x|\mu,\sigma^2)(x-\mu) \\
0 &= \mathcal{N}(x|\mu,\sigma^2)(x-\mu)
\end{aligned}$$

But since $\mathcal{N}(x|\mu,\sigma^2)(x-\mu) > 0$ the only term that matters is $(x-\mu)$, which goes to 0 at $x=\mu$

The proof for the multivariate case is almost identical and is omitted

Problem 10

$$\begin{aligned}
\log p(\mathbf{x}|\mu,\sigma^2) &= -\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi) \\
\frac{d\log p(\mathbf{x}|\mu,\sigma^2)}{d\mu} &= -\frac{1}{\sqrt{\pi\sigma^2}}\sum_{n=1}^{N}(x_n-\mu) \\
0 &= \sum_{n=1}^{N}(x_n-\mu) \\
&= \sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\mu \\
\mu &= \sum_{n=1}^{N} x_n
\end{aligned}$$

Problem 11

Just take the derivative with respect to μ of our function:

$$\ln(x|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\begin{aligned}
\frac{d}{d\mu}\ln(x|\mu,\sigma^2) &= \frac{1}{\sigma^2}\sum_{n=1}^{N}x_n - \mu \\
0 &= \sum_{n=1}^{N}x_n - \mu \\
\sum_{n=1}^{N}\mu &= \sum_{n=1}^{N}x_n \\
N\mu &= \sum_{n=1}^{N}x_n \\
\mu &= \frac{1}{N}\sum_{n=1}^{N}x_n
\end{aligned}$$

Now repeat the same for σ^2

$$\frac{d}{d\mu}\ln(x|\mu,\sigma^2) = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^2}$$
$$0 = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^2}$$
$$\frac{N}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2$$
$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

1.3 Curse of Dimensionality

Problem 17

The gamma function is defined as:

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

We use integration by parts with the following substitutions

$$u = u^{x-1} \qquad v = -e^{-u}$$
$$du = (x-1)u^{x-2}du \qquad dv = e^{-u}du$$

By integration by parts we have:

$$\begin{aligned}
\Gamma(x) &= -u^{x-1}e^{-u}|_0^\infty + (x-1)\int_0^\infty u^{x-2}e^{-u}du \\
&= (x-1)\int_0^\infty u^{x-2}e^{-u}du \\
&= (x-1)\Gamma(x-1) \\
&= z\Gamma(z) \qquad\qquad z = x+1
\end{aligned}$$

Now see that $\Gamma(1) = 1$

$$\begin{aligned}
\Gamma(1) &= \int_0^\infty u^0 e^{-u}du \\
&= \int_0^\infty e^{-u}du \\
&= -e^{-u}|_0^\infty \\
&= 1
\end{aligned}$$

To show that for all $x \in \mathbb{Z}^+$ $\Gamma(x+1) = x!$ we use an inductive argument. Our base case is $\Gamma(1) = 1$ which we have already done. Now let's assume that our inductive hypothesis is true and $\Gamma(x+1) = x!$. Then $\Gamma(x+2) = (x+1)\Gamma(x+1) = (x+1)x!$ and we are done.

Problem 18

We want to solve for S_D the surface area of a D dimensional sphere.

$$\prod_{i=1}^{D}\int_{-\infty}^{\infty} e^{-x_i^2} = S_D\int_0^\infty e^{-r^2}r^{D-1}dr$$

Let's solve the left hand side first. From problem 7 above we know that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}x^2} = \sqrt{2\pi\sigma^2}$$

By setting $\sigma^2 = \frac{1}{2}$ we get that

$$\begin{aligned}\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x^2} &= \prod_{i=1}^{D} \sqrt{\pi} \\ &= \pi^{\frac{D}{2}}\end{aligned}$$

Now let's look at the right hand side of the equation. We can tease this function into the Gamma function by making the substitution $u = r^2$ which implies that $dr = \frac{1}{2}u^{-1/2}$. With this insight we see that:

$$\begin{aligned}\int_0^{\infty} e^{-r^2} r^{D-1} dr &= \frac{1}{2} \int_0^{\infty} e^{-u} u^{\frac{D}{2}-1} du \\ &= \frac{1}{2}\Gamma(D/2)\end{aligned}$$

Now just use simple algebra:

$$\begin{aligned}\pi^{\frac{D}{2}} &= S_D \frac{1}{2}\Gamma(D/2) \\ S_D &= \frac{2\pi^{D/2}}{\Gamma(D/2)}\end{aligned}$$

.

Next integrate the surface area with respect to the radius in order to get the equation for volume:

$$\begin{aligned} V_D &= S_D \int_0^1 r^{D-1} dr \\ &= S_D \left. \frac{r}{D} \right|_0^1 \\ &= \frac{S_D}{D} \end{aligned}$$

Now to derive some common known volumes. For D =2:

$$\begin{aligned} \frac{2\pi^{D/2}}{D\Gamma(D/2)} &= \frac{2\pi}{2\Gamma(1)} \\ &= \pi \end{aligned}$$

For D=3:

$$\begin{aligned} \frac{2\pi^{D/2}}{D\Gamma(D/2)} &= \frac{2\pi^{3/2}}{3\Gamma(3/2)} \\ &= \frac{2\pi^{3/2}}{3\frac{\sqrt{\pi}}{2}} \\ &= \frac{4\pi}{3} \end{aligned}$$

Problem 19

From problem 18 we know what the volume of a hypersphere is. The volume of a D-dimensional cube is simply l^D where l is the length of a side. In our case this is $2a$ so the ratio is:

$$\begin{aligned}\frac{\text{Volume Sphere}}{\text{Volume Cube}} &= \frac{2\pi^{D/2}a^D}{(2a)^D D\Gamma(D/2)} \\ &= \frac{\pi^{D/2}}{2^{D-1}D\Gamma(D/2)}\end{aligned}$$

To show that the ratio goes to zero I'm going to eschew Stirling's formula because there is a much easier way to show this.... Begin by noticing that the above equation is always positive. Therefore we know that $0 \leq \frac{\pi^{D/2}}{2^{D-1}D\Gamma(D/2)}\ \forall D \in \mathbb{Z}$. Now let's replace the gamma function by the factorial function because they are identical for integer operands. I will also drop the D and 2^{D-1} from the bottom which increases the total value of the expression giving me:

$$\begin{aligned}0 &\leq \frac{\pi^{D/2}}{2^{D-1}D\Gamma(D/2)} \leq \frac{\pi^{D/2}}{(D/2)!} \\ 0 &\leq \frac{\pi^{x}}{x2^{2x}\Gamma(x)} \leq \frac{\pi^{x}}{x!}\end{aligned}$$

Now we just need to show that $\lim_{x\to\infty} \frac{\pi^x}{x!} = 0$. Which is easy! Notice that:

$$\begin{aligned}0 &< \frac{\pi^x}{x!} \\ \frac{\pi^x}{x!} &= \frac{\pi}{1}\cdot\frac{\pi}{2}\cdot\frac{\pi}{3}\cdot\ldots\cdot\frac{\pi}{x} \\ &< \frac{\pi}{1}\cdot\frac{\pi}{2}\cdot\frac{\pi}{3}\cdot\frac{\pi}{4}\cdot\frac{\pi}{4}\cdot\ldots\cdot\frac{\pi}{4} \\ &= \frac{\pi}{6}\left(\frac{\pi}{4}\right)^{n-3}\end{aligned}$$

Now we know that:

$$\lim_{x\to\infty}\left(\frac{\pi}{4}\right)^{n-3}=0$$

Therefore by the squeeze theorem so does $\frac{\pi^{D/2}}{2^{D-1}D\Gamma(D/2)}$ and we are done.

To show that the ratio of the distance from the center to a corner to the distance from the center to a side is $\sqrt{D}$ start by centering the cube at the origin. Then we know that one corner will be at $(a, a, ..., a)$ and the distance to that corner will be $a\sqrt{D}$ in euclidean space. Now look at the distance to a side. A side will be at the point $(a, 0, 0, ...)$ which will be distance a from the origin in Euclidean space so the ratio is $\sqrt{D}$

Problem 20

1.4 Decision Theory

Problem 21

First we prove that given two non negative numbers a, b and $a \leq b$ $\Rightarrow a \leq \sqrt{ab}$:

$$\begin{aligned} a &\leq ab \\ a^2 &\leq ab \\ |a| &\leq \sqrt{ab} \\ a &\leq \sqrt{ab} \end{aligned}$$

the probability of a mistake is the probability of making an error over all classification sub regions.

$$p(mistake) = \int_{R_1} p(x, C_1)dx + \int_{R_2} p(x, C_2)dx$$

We know that over R_1 $p(C_1|x) \geq p(C_2|x)$. By our first proof we know that:

$$\begin{aligned} p(C_2|x) &\leq \sqrt{p(C_1|x)p(C_2|x)} \\ p(C_2|x)p(x) &\leq p(x)\sqrt{p(C_1|x)p(C_2|x)} \\ p(x, C_2) &\leq \sqrt{p(C_1|x)p(C_2|x)p(x)^2} \\ &\leq \sqrt{p(x, C_1)p(x, C_2)} \end{aligned}$$

By an identical argument over R_2 $p(C_1|x) \leq p(C_2|x)$ and

$$p(x, C_1) \leq \sqrt{p(C_1|x)p(C_2|x)p(x)^2}$$

substituting these into the definition of a mistake probability:

$$\begin{aligned} p(mistake) &\leq \int_{R_1} \sqrt{p(x, C_1)p(x, C_2)}dx + \int_{R_2} \sqrt{p(x, C_1)p(x, C_2)}dx \\ &\leq \int \sqrt{p(x, C_1)p(x, C_2)}dx \end{aligned}$$

Problem 22

The loss matrix $L_{kj} = 1 - I_{kj}$ can be visualized as:

$$L = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 1 & \cdots & \cdots & 1 & 0 \end{bmatrix}$$

This is interpreted as there is no penalty for choosing the correct class, however any type of misclassification is penalized the same. This means that we just want to choose the class with the largest posterior probability because it will be the most likely to be correct, if it is incorrect it is no more incorrect than any other non correct guess. To see this formally we use the delta function which is defined as:

$$\delta_{kj} = \begin{cases} 0, & \text{for } k = j \\ 1, & \text{for else} \end{cases}$$

Using this notation we can write $L_{kj} = 1 - \delta_{kj}$ This makes the function we want to minimize:

$$\begin{aligned}\sum_k L_{kj} p(C_k|x) &= \sum_k (1 - \delta_{kj}) p(C_k|x) \\ &= \sum_k p(C_k|x) - \delta_{kj} p(C_k|x) \\ &= \sum_k p(C_k|x) - \sum_k \delta_{kj} p(C_k|x) \\ &= 1 - \sum_k \delta_{kj} p(C_k|x)\end{aligned}$$

In order to minimize this function we want $p(C_k|x)$ to be as large as possible so we choose the class which has the largest posterior probability.

Problem 24

Problem 25

$$
\begin{aligned}
\mathbb{E}\left[L(t, y(x))\right] &= \int\int \|y(x) - t\|_2^2 p(x,t) dx dt \\
\frac{d\mathbb{E}\left[L(t, y(x))\right]}{dy(x)} &= 2\int \|y(x) - t\|_2 \frac{y(x) - t}{\|y(x) - t\|_2} p(x,t) dt \\
0 &= \int y(x) p(x,t) dt - \int t p(x,t) dt \\
y(x) &= \frac{\int t p(x,t) dt}{\int p(x,t) dt} \\
&= \frac{\int t p(x,t) dt}{p(x)} \\
&= \frac{\int t p(t|x) p(x) dt}{p(x)} = \int t p(t|x) dt
\end{aligned}
$$

Problem 26

To start. If we ask the question what is the information gain of $p(x)^2$ this is the same as how much do we learn from observing that event twice. This should intuitively be double. To show this. Take $p(x)^2 = p(x)p(x) = p(x,x)$. This means that events are independent and we have already show what this is in terms of information gain $h(x,x) = h(x) + h(x) = 2h(x)$. Now let's assume that this is true for arbitrary integers i.e. $p(x)^n \Rightarrow nh(x)$. Call $p(y) = p(x)^n$ then $p(x)p(y) = p(x,y)$ and $h(x,y) = h(x) + h(y)$ by the induction hypothesis $h(y) = nh(x)$ so we have $h(x) + nh(x) = (n+1)h(x)$ and we are done.

Problem 29

We want to show that $H[x] \leq \ln M$ using Jensen's inequality. First note that for discrete random variables Jensen's inequality is:

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

Notice that the definition of entropy is:

$$\begin{aligned} H[x] &= -\sum_{i=1}^{M} p(x_i) \ln p(x_i) \\ &= \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)} \end{aligned}$$

Set $f(x) = \ln(x)$ then we have that:

$$H[x] = \mathbb{E}[f(x)]$$

and

$$\begin{aligned} f(\mathbb{E}[x]) &= \ln \left(\sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)} \right) \\ &= \ln M \end{aligned}$$

Because ln is concave, for a handwavy visual proof of this observe that the epigraph of $-\ln$ is convex then we can reverse Jensen's inequality to get the desired results.

Problem 30

We want to determine what the KL divergence between two Gaussians is. The KL Divergence is defined as:

$$KL(p\|q) = -\int p(x)\ln q(x) + \int p(x)\ln p(x)$$

Our Gaussians in question are:

$$p(x) = \mathcal{N}(x;\mu,\sigma^2) \qquad\qquad q(x) = \mathcal{N}(x;m,s^2)$$

Let's examine the left hand integral first because it is the most troublesome.

$$\begin{aligned}
L &= \int (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \ln\left[(2\pi s^2)^{-\frac{1}{2}} e^{-\frac{1}{2s^2}(x-m)^2}\right] \\
&= \int (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \left[-\frac{1}{2}\ln(2\pi s^2) - \frac{1}{2s^2}(x-m)^2\right] \\
&= -\frac{1}{2}\ln(2\pi s^2)\int (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} - \frac{1}{2s^2}\int (x-m)^2 (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\
&= -\frac{1}{2}\ln(2\pi s^2)\int \mathcal{N}(x;\mu,\sigma^2) - \frac{1}{2s^2}\int (x-m)^2\mathcal{N}(x;\mu,\sigma^2) \\
&= -\frac{1}{2}\ln(2\pi s^2) - \frac{1}{2s^2}\int (x-m)^2\mathcal{N}(x;\mu,\sigma^2)
\end{aligned}$$

Now let's expand this last integral:

$$\begin{aligned}
I &= -\frac{1}{2s^2}\left[\int x^2\mathcal{N}(x;\mu,\sigma^2) - \int 2xm\mathcal{N}(x;\mu,\sigma^2) + \int m^2\mathcal{N}(x;\mu,\sigma^2)\right] \\
&= -\frac{1}{2s^2}\left[(\sigma^2+\mu^2) - 2m\mu + m^2\right]
\end{aligned}$$

This last line comes by recognizing that the first integral is the $\mathbb{E}\left[(x-\mu)^2\right]$ and the second integral is just the expected value of the

Gaussian. Putting it all together for the left hand integral in our original equation we get:

$$\int p(x) \ln q(x) = -\frac{1}{2} \ln(2\pi s^2) - \frac{1}{2s^2} \left[(\sigma^2 + \mu^2) - 2m\mu + m^2 \right]$$

We can get the right hand integral value by substituting in $s^2 = \sigma^2$ and $m = \mu$ to yield:

$$\int p(x) \ln p(x) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2}$$

Now it's just simple algebra on these tow equations:

$$\begin{aligned} KL(p\|q) &= \frac{1}{2} \ln(2\pi s^2) + \frac{1}{2s^2} \left[\sigma^2 + \mu^2 - 2m\mu + m^2 \right] - \frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \\ &= \frac{1}{2} \ln(2\pi s^2) + \frac{\sigma^2 + (\mu - m)^2}{2s^2} - \frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \end{aligned}$$

We can check this by setting $p(x) = q(x)$ then this equation goes to 0 which is what we would expect and we are done.

Problem 31

To show that $H[x, y] \leq H[x] + H[y]$ with equality iff $x \perp y$ we just need to look at the mutual information. Which is:

$$I[x, y] = H[y] - H[y|x]$$

We know that $I[x, y] \geq 0$ with eqality iff $x \perp y$. Therefore if x and y are not orthogonal then $H[y] - H[y|x] > 0$ because $I[x, y] \geq 0$. This implies that $H[y|x] < H[y]$. Now by the definition of conditional

entropy we have that $H[x, y] = H[y|x] + H[x]$. Combining this and the inequality derived above we get that $H[x, y] \leq H[x] + H[y]$. Intuitively this is saying that the conditional entropy is greatest when the events are independent. This makes sense because if they are independent then there is no sharing of information.

Problem 32

Problem 36

The definition of convexity is:

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$$

we start by arranging the terms to have 0 on one side.

$$\begin{aligned}
0 &\leq \lambda f(a) + f(b) - \lambda f(b) - f(\lambda a + b - \lambda b) \\
0 &\leq \lambda f(x + h) + f(b) - \lambda f(b) - f(\lambda(x + h) + b - \lambda b) \\
0 &\leq \lambda f(x + h) + f(x - h) - \lambda f(x - h) - f(\lambda(x + h) + (x - h) - \lambda(x - h)) \\
0 &\leq \lambda f(x + h) + f(x - h) - \lambda f(x - h) - f(x - h + 2\lambda h) \\
0 &\leq \frac{1}{2} f(x + h) + f(x - h) - \frac{1}{2} f(x - h) - f(x) \\
0 &\leq f(x + h) + 2f(x - h) - f(x - h) - 2f(x) \\
0 &\leq f(x + h) + f(x - h) - 2f(x) \\
0 &\leq \frac{f(x + h) + f(x - h) - 2f(x)}{h^2} \\
0 &\leq f''(x)
\end{aligned}$$

Problem 37

$$
\begin{aligned}
H[y|x] &= -\int\int p(y,x)\ln p(y|x)dxdy \\
&= -\int\int p(y,x)\ln\left(\frac{p(x,y)}{p(x)}\right)dxdy \\
&= -\int\int p(y,x)\ln p(x,y)dxdy + \int\int p(y,x)\ln p(x)dxdy \\
&= H[x,y] + \int\int p(y,x)\ln p(x)dxdy \\
&= H[x,y] + \int p(x)\ln p(x)dx && \text{sum rule} \\
&= H[x,y] - H[x]
\end{aligned}
$$

Problem 39

(a) H[x]
Marginalize out y to get $x = [2/3, 1/3]$ then:

$$H[x] = -2/3\log[2/3] - 1/3\log[1/3]$$

(b) H[y]
Marginalize out x to get $y = [1/3, 2/3]$ then:

$$H[x] = -1/3\log[1/3] - 2/3\log[2/3]$$

(c) H[y|x]

$$
\begin{aligned}
p(y=0|x=0) &= 1/2 \\
p(y=0|x=1) &= 0 \\
p(y=1|x=0) &= 1/2 \\
p(y=1|x=1) &= 1
\end{aligned}
$$

$$H[y|x] = -1/2\log[1/2] - 1/2\log[1/2]$$

(d) H[x|y]

$$p(x=0|y=0)=1$$
$$p(x=0|y=1)=1/2$$
$$p(x=1|y=0)=0$$
$$p(x=1|y=1)=1/2$$

$$H[x|y]=-1/2\log[1/2]-1/2\log[1/2]$$

(e) H[x, y]
This is just the elements in the table.

$$H[x,y]=-1/3\log[1/3]-1/3\log[1/3]-1/3\log[1/3]=-\log[1/3]$$

(f) I[x, y]
$I[x,y]=H[x]-H[x|y]=2/3\log[2/3]-1/3\log[1/3]-1/2\log[1/2]-1/2\log[1/2]$

Problem 40

Start by recognizing that $\ln(x)$ is a concave function. (You could take the second derivative to see this.) This means that we should reverse

Jensen's inequality to get:

$$\sum_{i=1}^{N} \lambda_i f(x_i) \leq f\left(\sum_{i=1}^{N} \lambda_i x_i\right)$$

$$\sum_{i=1}^{N} \lambda_i \ln(x_i) \leq \ln\left(\sum_{i=1}^{N} \lambda_i x_i\right) \qquad f(x) = \ln(x)$$

$$\lambda_i \ln\left(\prod_{i=1}^{N} x_i\right) \leq \ln\left(\sum_{i=1}^{N} \lambda_i x_i\right) \qquad \ln(xy) = \ln(x) + \ln(y)$$

$$\frac{1}{N} \ln\left(\prod_{i=1}^{N} x_i\right) \leq \ln\left(\sum_{i=1}^{N} \frac{1}{N} x_i\right) \qquad \lambda_i = \frac{1}{N}$$

$$\ln\left(\sqrt[N]{\prod_{i=1}^{N} x_i}\right) \leq \ln\left(\sum_{i=1}^{N} \frac{1}{N} x_i\right)$$

$$\sqrt[N]{\prod_{i=1}^{N} x_i} \leq \sum_{i=1}^{N} \frac{1}{N} x_i$$

Problem 41 To prove that $I[x, y] = H[y] - H[y|x]$ we just need to use the definition of mutual information:

$$
\begin{aligned}
I[x,y] &= -\int\int p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)dxdy \\
&= -\int\int p(x,y)\ln(p(x)p(y))dxdy + \int\int p(x,y)\ln p(x,y)dxdy \\
&= -\int\int p(x,y)\ln(p(x)p(y))dxdy - H[x,y] \\
&= -\int\int p(x,y)\ln p(y)dxdy - \int\int p(x,y)\ln p(x)dxdy - H[x,y] \\
&= -\int p(y)\ln p(y)dy - \int p(x)\ln p(x)dx - H[x,y] \\
&= H[x] + H[y] - H[x,y] \\
&= H[x] + H[y] - H[x] - H[y|x] \\
&= H[x] - H[y|x]
\end{aligned}
$$

Chapter 2

Probability Distributions

2.1 Binary Variables

Problem 1

$$\begin{aligned}\sum_{x=0}^{1} p(x|\mu) &= \sum_{x=0}^{1} \mu^x (1-\mu)^{1-x} \\ &= \mu^0(1-\mu) + \mu(1-\mu)^0 \\ &= 1 - \mu + \mu \\ &= 1\end{aligned}$$

$$\begin{aligned}\mathbb{E}[x] &= \sum_{x=0}^{1} x\mu^x (1-\mu)^{1-x} \\ &= 0 + \mu(1-\mu)^0 \\ &= \mu\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\,[x] &= \sum_{x=0}^{1}(x-\mu)^2\mu^x(1-\mu)^{1-x} \\
&= \mu^2(1-\mu) + (1-\mu)^2\mu \\
&= \mu^2 - \mu^3 + (1-2\mu+\mu^2)\mu \\
&= \mu^2 - \mu^3 + \mu - 2\mu^2 + \mu^3 \\
&= \mu - \mu^2 \\
&= \mu(1-\mu)
\end{aligned}$$

$$\begin{aligned}
H[x] &= -\sum_{x=0}^{1} p(x)\ln p(x) \\
&= -\sum_{x=0}^{1} \mu^x(1-\mu)^{1-x}\ln \mu^x(1-\mu)^{1-x} \\
&= -(1-\mu)\ln(1-\mu) - \mu\ln\mu
\end{aligned}$$

Problem 2

$$\begin{aligned}
\sum_{x\in\{-1,1\}} p(x|\mu) &= \sum_{x\in\{-1,1\}} \left(\frac{1-\mu}{2}\right)^{(1-x)/2}\left(\frac{1+\mu}{2}\right)^{(1+x)/2} \\
&= \left(\frac{1-\mu}{2}\right) + \left(\frac{1+\mu}{2}\right) \\
&= 1
\end{aligned}$$

$$\begin{aligned}\mathbb{E}[x] &= \sum_{x\in\{-1,1\}} x \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2} \\ &= -\left(\frac{1-\mu}{2}\right) + \left(\frac{1+\mu}{2}\right) \\ &= \mu\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[x] &= \sum_{x\in\{-1,1\}} (x-\mu)^2 \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2} \\ &= (-1-\mu)^2 \left(\frac{1-\mu}{2}\right) + (1-\mu)^2 \left(\frac{1+\mu}{2}\right) \\ &= \frac{(1+2\mu+\mu^2)(1-\mu) + (1-2\mu+\mu^2)(1+\mu)}{2} \\ &= 1-\mu^2\end{aligned}$$

$$\begin{aligned}H[x] &= -\sum_{x\in\{-1,1\}} p(x)\ln p(x) \\ &= -\sum_{x\in\{-1,1\}} \left(\frac{1-\mu}{2}\right)^{\frac{1-x}{2}} \left(\frac{1+\mu}{2}\right)^{\frac{1+x}{2}} \ln\left[\left(\frac{1-\mu}{2}\right)^{\frac{1-x}{2}} \left(\frac{1+\mu}{2}\right)^{\frac{1+x}{2}}\right] \\ &= -\left(\frac{1-\mu}{2}\right)\ln\left(\frac{1-\mu}{2}\right) - \left(\frac{1+\mu}{2}\right)\ln\left(\frac{1+\mu}{2}\right) \\ &= \frac{-(1-\mu)\ln(1-\mu) + (1-\mu)\ln(2) - (1+\mu)\ln(1+\mu) + (1+\mu)\ln(2)}{2}\end{aligned}$$

Problem 3

First we prove that

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$$

To show this imagine that we have a box filled with N blue balls and 1 red ball. Then there are $\binom{N+1}{m}$ different ways to draw balls from this box where order doesn't matter. We can break this up into two separate cases where we examine the number of ways to draw only blue balls from the box, and the number of ways to draw blue balls with a red ball. There are $\binom{N}{m}$ ways to draw only blue balls from the box because we remove the red ball from our selection so there are N balls and we draw m of them. Then there are $\binom{N}{m-1}$ ways to draw balls from a box when we are gauranteed to have a red one in our selection. This is because we select the red ball so we only have N balls to choose the remaining blue from, and we only need to draw $m-1$ balls because we've already drawn the red.

Now to prove the binomial theorem we start with the base case where $N = 0$.

$$(x+1)^m = \sum_{m=0}^{N} \binom{N}{m} x^m$$

$$(x+1)^0 = \sum_{m=0}^{0} \binom{0}{0} x^0$$

$$1 = 1$$

Now we use the induction hypothesis to show that it is true for $N+1$

$$
\begin{aligned}
(x+1)^m &= \sum_{m=0}^{N} \binom{N}{m} x^m \\
(x+1)^m(x+1) &= (x+1) \sum_{m=0}^{N} \binom{N}{m} x^m \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + x \sum_{m=0}^{N} \binom{N}{m} x^m \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=0}^{N} \binom{N}{m} x^{m+1} \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=1}^{N+1} \binom{N}{m-1} x^{m+1-1} \\
&= \binom{N}{0} x^0 + \sum_{m=1}^{N} \left[\binom{N}{m} + \binom{N}{m-1} \right] x^m + \binom{N}{N} x^{N+1} \\
&= \binom{N}{0} x^0 + \sum_{m=1}^{N} \binom{N+1}{m} x^m + \binom{N}{N} x^{N+1} \\
&= \sum_{m=0}^{N} \binom{N+1}{m} x^m + \binom{N}{N} x^{N+1} \\
&= \sum_{m=0}^{N+1} \binom{N+1}{m} x^m
\end{aligned}
$$

Thus we have shown by induction that if our hypothesis is true for N this implies that it is true for $N+1$ as well.

The last thing to show is that the binomial distribution is normalized.

$$\begin{aligned}
\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} &= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{-m} \\
&= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \left(\frac{\mu}{1-\mu}\right)^m \\
&= (1-\mu)^N \left(1 + \frac{\mu}{1-\mu}\right)^N \\
&= \left(1 + \frac{\mu}{1-\mu} - \mu - \frac{\mu^2}{1-\mu}\right)^N \\
&= \left(1 + \frac{\mu}{1-\mu} - \frac{\mu(1-\mu)}{1-\mu} - \frac{\mu^2}{1-\mu}\right)^N \\
&= \left(1 + \frac{\mu - \mu + \mu^2 - \mu^2}{1-\mu}\right)^N \\
&= 1^N \\
&= 1
\end{aligned}$$

Problem 4

To prove that the expected value of the binomial distribution is $N\mu$ we begin by differentiating both sides of the normalized distribution with respect to μ and then using basic algebra.

$$\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1$$

$$\frac{d}{d\mu} \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 0$$

$$\sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1} (1-\mu)^{N-m} = \sum_{m=0}^{N} (N-m) \binom{N}{m} \mu^m (1-\mu)^{N-m-1}$$

$$(1-\mu) \sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1} (1-\mu)^{N-m} = \sum_{m=0}^{N} (N-m) \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$(1-\mu) \sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1} (1-\mu)^{N-m} = N - \mathbb{E}[m]$$

$$\sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1} (1-\mu)^{N-m} = N - \mathbb{E}[m] + \sum_{m=0}^{N} \binom{N}{m} m\mu^m (1-\mu)^{N-}$$

$$\sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1} (1-\mu)^{N-m} = N - \mathbb{E}[m] + \mathbb{E}[m]$$

$$\mu \sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1} (1-\mu)^{N-m} = N\mu - \mathbb{E}[m]$$

$$\sum_{m=0}^{N} \binom{N}{m} m\mu^m (1-\mu)^{N-m} = N\mu - \mathbb{E}[m]$$

$$\mathbb{E}[m] = N\mu$$

Problem 6

$$\begin{aligned}
\mathbb{E}[x] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x x^{\alpha-1}(1-x)^{\beta-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{\alpha}(1-x)^{\beta-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+1)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{a\Gamma(a)}{(a+b)\Gamma(a+b)} \\
&= \frac{a}{a+b}
\end{aligned}$$

To calculate the Variance I will use the fact that
$\mathrm{Var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$ First let's calculate $\mathbb{E}\left[x^2\right]$

$$\begin{aligned}
\mathbb{E}\left[x^2\right] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^2 x^{\alpha-1}(1-x)^{\beta-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+2)}{\Gamma(a+b+2)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{(a+1)a\Gamma(a)}{(a+b+1)(a+b)\Gamma(a+b)} \\
&= \frac{(a+1)a}{(a+b+1)(a+b)}
\end{aligned}$$

Now we know that $\mathbb{E}\left[x\right]^2 = \left(\frac{a}{a+b}\right)^2$ so the Variance is:

$$\begin{aligned}
\mathrm{Var}\left[x\right] &= \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 \\
&= \frac{(a+1)a}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2} \\
&= \frac{(a+1)a(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} \\
&= \frac{(a^2+a)(a+b) - (a^3+a^2b+a^2)}{(a+b+1)(a+b)^2} \\
&= \frac{a^3+a^2b+a^2+ab-a^3-a^2b-a^2}{(a+b+1)(a+b)^2} \\
&= \frac{ab}{(a+b+1)(a+b)^2}
\end{aligned}$$

Lastly the mode. This can be calculated simply by taking the derivative of the Beta distribution and setting it equal to zero

$$\begin{aligned}
\text{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
0 &= \frac{d}{d\mu}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{d}{d\mu}\mu^{a-1}(1-\mu)^{b-1} \\
&= (a-1)\mu^{a-2}(1-\mu)^{b-1} - (b-1)\mu^{a-1}(1-\mu)^{b-2} \\
&= \mu^{a-2}(1-\mu)^{b-2}\left((a-1)(1-\mu) - (b-1)\mu\right) \\
&= (a-1)(1-\mu) - (b-1)\mu \\
&= a - a\mu - 1 + \mu - b\mu + \mu \\
&= a - a\mu - 1 + \mu - b\mu + \mu \\
&= (a-1) - \mu(a+b-2) \\
(1-a) &= -\mu(a+b-2) \\
\mu &= \frac{a-1}{a+b-2}
\end{aligned}$$

Problem 7

To begin we notice that the posterior distribution is:

$$p(\mu|x) = p(x|\mu)p(\mu)$$

So we can write this in terms of the provided distributions as:

$$\begin{aligned} p(\mu|x) &= \text{Bin}(m|N,\mu)\text{Beta}(a,b) \\ &= \binom{m+l}{m}\mu^m(1-\mu)^l\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ &\propto \mu^m(1-\mu)^l\mu^{a-1}(1-\mu)^{b-1} \\ &= \mu^{a+m-1}(1-\mu)^{b+l-1} \\ &= \text{Beta}(a+m,b+l) \end{aligned}$$

Therefore the posterior expectation is just the expected value of the beta distribution above:

$$\begin{aligned} \mathbb{E}\left[p(\mu|m)\right] &= \frac{a+m}{a+b+m+l} \\ &= \frac{a}{a+b+m+l} + \frac{m}{a+b+m+l} \end{aligned}$$

Now simply set $\lambda = \frac{a+b+m+l}{a+b}$ which implies that $1-\lambda = \frac{m+l}{a+b+m+l}$. If we substitute these into our equation for the expected value of the posterior then we get:

$$\mathbb{E}\left[p(\mu|m)\right] = \frac{a}{a+b}\lambda + (1-\lambda)\frac{m}{m+l}$$

And we are done.

Problem 8

Here we derive the law of total expectation and the law of total variance. To derive the law of total expectation:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] &= \mathbb{E}\left[\sum_x xp(X=x|Y)\right] \\
&= \sum_y \sum_x xp(X=x|Y=y)p(Y=y) \\
&= \sum_x x \sum_y p(X=x|Y=y)p(Y=y) \\
&= \sum_x x \sum_y p(X=x, Y=y) \\
&= \sum_x xp(x) \\
&= \mathbb{E}\left[X\right]
\end{aligned}$$

Now to derive the law of total variance we begin with the definition of variance:

$$\begin{aligned}
\mathrm{Var}\left[X\right] &= \mathbb{E}\left[X^2\right]_x - \mathbb{E}\left[X\right]_x^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[X^2|Y\right]_x\right]_y - \mathbb{E}\left[\mathbb{E}\left[X|Y\right]_x\right]_y^2 \\
&= \mathbb{E}\left[\mathrm{Var}\left[X|Y\right]_x + \mathbb{E}\left[X|Y\right]_x^2\right]_y - \mathbb{E}\left[\mathbb{E}\left[X|Y\right]_x\right]_y^2 \\
&= \mathbb{E}\left[\mathrm{Var}\left[X|Y\right]_x\right]_y + \mathbb{E}\left[\mathbb{E}\left[X|Y\right]_x^2\right]_y + \mathrm{Var}\left[\mathbb{E}\left[X|Y\right]_x\right]_y - \mathbb{E}\left[\mathbb{E}\left[X|Y\right]^2\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[X|Y\right]_x\right]_y + \mathrm{Var}\left[\mathbb{E}\left[X|Y\right]_x\right]_y
\end{aligned}$$

Step two above we use the law of total expectation. In step three we use the fact that $\mathrm{Var}\left[X\right] + \mathbb{E}\left[X\right]^2 = \mathbb{E}\left[X^2\right]$. In step four we use the linearity of expectation. In step five we use the fact that $\mathrm{Var}\left[X\right] - \mathbb{E}\left[X^2\right] = -\mathbb{E}\left[X\right]^2$

Problem 10

Here we derive properties of the Dirichlet distribution. We start with the expected value.

Without loss of generality I will prove the case for $\mathbb{E}\left[[]\,\mu_1\right]$ it makes

some of the notation a little clearer this way.

$$
\begin{aligned}
\mathbb{E}\left[\mu_1\right] &= \frac{1}{B(\alpha)} \int \mu_1 \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} \\
&= \frac{1}{B(\alpha)} \int \mu_1^{\alpha_1} \prod_{i=2}^{K} \mu_i^{\alpha_i - 1} \\
&= \frac{1}{B(\alpha)} \int \mu_1^{\beta_1 - 1} \prod_{i=2}^{K} \mu_i^{\alpha_i - 1} && \beta_1 = \alpha_1 + 1 \\
&= \frac{1}{B(\alpha)} \int \mu_1^{\beta_1 - 1} \prod_{i=2}^{K} \mu_i^{\beta_i - 1} && \beta_i = \alpha_i \; \forall i \neq 1 \\
&= \frac{1}{B(\alpha)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_0 \Gamma(\alpha_0)}{\alpha_0 \Gamma(\alpha_1) \prod_{i=2}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_1 \Gamma(\alpha_0 + 1)}{\alpha_0 \alpha_1 \Gamma(\alpha_1) \prod_{i=2}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_1 \Gamma(\alpha_0 + 1)}{\alpha_0 \Gamma(\alpha_1 + 1) \prod_{i=2}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_1 \Gamma(\beta_0)}{\alpha_0 \Gamma(\beta_1) \prod_{i=2}^{K} \Gamma(\beta_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_1}{\alpha_0} \int \frac{\Gamma(\beta_0)}{\prod_{i=1}^{K} \Gamma(\beta_i)} \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_1}{\alpha_0} \int \mathrm{Dir}(\beta) \\
&= \frac{\alpha_1}{\alpha_0}
\end{aligned}
$$

For the Variance we will again use the fac that $varx = epxx^2 - expx^2$

$$
\begin{aligned}
\mathbb{E}\left[\mu_1^2\right] &= \frac{1}{B(\alpha)} \int \mu_1^2 \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} \\
&= \frac{1}{B(\alpha)} \int \mu_1^{\alpha_1 + 1} \prod_{i=2}^{K} \mu_i^{\alpha_i - 1} \\
&= \frac{1}{B(\alpha)} \int \mu_1^{\beta_1 - 1} \prod_{i=2}^{K} \mu_i^{\alpha_i - 1} && \beta_1 = \alpha_1 + 2 \\
&= \frac{1}{B(\alpha)} \int \mu_1^{\beta_1 - 1} \prod_{i=2}^{K} \mu_i^{\beta_i - 1} && \beta_i = \alpha_i \ \forall i \neq 1 \\
&= \frac{(\alpha_0 + 1)\alpha_0 \Gamma(\alpha_0)}{(\alpha_0 + 1)\alpha_0 \Gamma(\alpha_1) \prod_{i=2}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\Gamma(\beta_0)}{(\alpha_0 + 1)\alpha_0 \Gamma(\alpha_1) \prod_{i=2}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{(\alpha_1 + 1)\alpha_1 \Gamma(\beta_0)}{(\alpha_0 + 1)(\alpha_1 + 1)\alpha_1 \alpha_0 \Gamma(\alpha_1) \prod_{i=2}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{(\alpha_1 + 1)\alpha_1 \Gamma(\beta_0)}{(\alpha_0 + 1)(\alpha_1 + 1) \prod_{i=1}^{K} \Gamma(\beta_i)} \int \prod_{i=1}^{K} \mu_i^{\beta - 1} \\
&= \frac{\alpha_1(\alpha_1 + 1)}{\alpha_0(\alpha_0 + 1)}
\end{aligned}
$$

Now we can just plug in this into the equation for variance and solve:

$$\begin{aligned}
\text{Var}\left[\mu_1\right] &= \mathbb{E}\left[\mu_1^2\right] - \mathbb{E}\left[\mu_1\right]^2 \\
&= \frac{\alpha_1(\alpha_1+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha_1^2}{\alpha_0^2} \\
&= \frac{\alpha_1\alpha_0(\alpha_1+1) - \alpha_1^2(\alpha_0+1)}{\alpha_0^2(\alpha_0+1)} \\
&= \frac{\alpha_0\alpha_1^2 + \alpha_0\alpha_1 - \alpha_0\alpha_1^2 - \alpha_1^2}{\alpha_0^2(\alpha_0+1)} \\
&= \frac{\alpha_1(\alpha_0-\alpha_1)}{\alpha_0^2(\alpha_0+1)}
\end{aligned}$$

Lastly we prove the covariance of the Dirichlet distribution. Which is defined as $\text{Cov}(\mu_i, \mu_j) = \mathbb{E}\left[\mu_i\mu_j\right] - \mathbb{E}\left[\mu_i\right]\mathbb{E}\left[\mu_j\right]$. The proof proceeds much as the other two have the only difference is now we just look at $\mathbb{E}\left[\mu_1\mu_2\right]$ Now we need to substitute $\beta_1 = \alpha_1 + 1$, $\beta_2 = \alpha_2 + 2$ and $\beta_i = \alpha_i$. If you do this you will eventually get:

$$\mathbb{E}\left[\mu_i\mu_j\right] = \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}$$

Now we just combine them all:

$$\begin{aligned}
\operatorname{Cov}(\mu_i, \mu_j) &= \mathbb{E}\left[\mu_i \mu_j\right] - \mathbb{E}\left[\mu_i\right] \mathbb{E}\left[\mu_j\right] \\
&= \frac{\alpha_i \alpha_j}{\alpha_0(\alpha_0+1)} - \frac{\alpha_i}{\alpha_0}\frac{\alpha_j}{\alpha_0} \\
&= \frac{\alpha_0 \alpha_i \alpha_j - \alpha_i \alpha_j (\alpha_0+1)}{\alpha_0^2(\alpha_0+1)} \\
&= \frac{\alpha_0 \alpha_i \alpha_j - \alpha_i \alpha_j \alpha_0 - \alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)} \\
&= \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}
\end{aligned}$$

Problem 11

As the hint suggests start by differentiating $\prod_{i=1}^{k} \mu_i^{\alpha_i - 1}$ with respect to α_j without loss of generality we will set $j = 1$ for simple notation.

$$\begin{aligned}
\frac{d}{d\alpha_1} \prod_{i=1}^{k} \mu_i^{\alpha_i - 1} &= \frac{d}{d\alpha_1} \prod_{i=1}^{k} \exp((\alpha_i - 1)\ln \mu_i) \\
&= \frac{d}{d\alpha_1} \exp((\alpha_1 - 1)\ln \mu_1) \prod_{i=2}^{k} \exp((\alpha_i - 1)\ln \mu_i) \\
&= \exp((\alpha_1 - 1)\ln \mu_1) \ln \mu_1 \prod_{i=2}^{k} \exp((\alpha_i - 1)\ln \mu_i) \\
&= \ln \mu_1 \prod_{i=1}^{k} \exp((\alpha_i - 1)\ln \mu_i) \\
&= \ln \mu_1 \prod_{i=1}^{k} \mu_i^{\alpha_i - 1}
\end{aligned}$$

Now we can use this fact to calculate the expectation:

$$\begin{aligned}
\mathbb{E}\left[\ln \mu_1\right] &= \beta(\alpha) \int \ln \mu_1 \prod_{i=1}^{k} \mu_i^{\alpha_i - 1} d\mu \\
&= \beta(\alpha) \int \frac{d}{d\alpha_1} \prod_{i=1}^{k} \mu_i^{\alpha_i - 1} d\mu \\
&= \beta(\alpha) \frac{d}{d\alpha_1} \int \prod_{i=1}^{k} \mu_i^{\alpha_i - 1} d\mu \\
&= \beta(\alpha) \frac{d}{d\alpha_1} \frac{1}{\beta(\alpha)} \\
&= -\frac{d}{d\alpha_1} \ln \beta(\alpha)
\end{aligned}$$

Now the last step is just solving for $\ln \beta(\alpha)$:

$$\begin{aligned}
\ln \beta(\alpha) &= \ln \frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_i)} \\
&= \ln \Gamma(\alpha_0) - \ln \prod \Gamma(\alpha_i) \\
&= \ln \Gamma(\alpha_0) - \sum_i \ln \Gamma(\alpha_i)
\end{aligned}$$

Now you should be able to see that by applying the negative derivative we end up with our desired result of:

$$\mathbb{E}\left[\ln \mu_j\right] = \psi(\alpha_j) - \psi(\alpha_0)$$

Problem 12

Here we just prove aspects of the uniform distribution. Let's start with normalization.

$$\begin{aligned}\int_{-\infty}^{\infty} \text{Unif}(x|a,b) &= \int_a^b \frac{1}{b-a} dx \\ &= \frac{x}{b-1}\Big|_a^b \\ &= \frac{b}{b-a} - \frac{a}{b-a} \\ &= 1\end{aligned}$$

Now the expected value:

$$\begin{aligned}\mathbb{E}[x] &= \int_a^b \frac{x}{b-a} dx \\ &= \frac{x^2}{2(b-a)}\Big|_a^b \\ &= \frac{b^2}{2(b-a)} - \frac{a^2}{2(b-a)} \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} \\ &= \frac{1}{2}(b+a)\end{aligned}$$

Lastly we do the variance. It is very similar to the expectation.

$$\begin{aligned}
\mathbb{E}\left[x^2\right] &= \int_a^b \frac{x^2}{b-a} dx \\
&= \left.\frac{x^3}{3(b-a)}\right|_a^b \\
&= \frac{b^3}{3(b-a)} - \frac{a^3}{3(b-a)} \\
&= \frac{b^3 - a^3}{3(b-a)} \\
&= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} \\
&= \frac{1}{3}(b^2+ab+a^2)
\end{aligned}$$

$$\begin{aligned}
\text{Var}\left[x\right] &= \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 \\
&= \frac{1}{3}(b^2+ab+a^2) - \frac{1}{4}(b+a)^2 \\
&= \frac{4b^2+4ab+4a^2-3b^2-6ab-3a^2}{12} \\
&= \frac{1}{12}(b^2-2ab+a^2) \\
&= \frac{1}{12}(b-a)^2
\end{aligned}$$

Problem 13

Now we need to derive the KL Divergence for two multivariate Gaussian. We define our two Gaussians as:

$$p(x) = \mathcal{N}(x|, \mu, \Sigma) \qquad\qquad q(x) = \mathcal{N}(x|, m, L)$$

Now we can start with the definition of the KL Divergence:

$$
\begin{aligned}
KL &= \int p(x) \ln \frac{q(x)}{p(x)} dx \\
&= \int p(x) \left[\ln q(x) - \ln p(x)\right] dx \\
&= \int p(x) \frac{1}{2} \left[\ln \frac{|L|}{|\Sigma|} + (x-m)^T L^{-1}(x-m) - (x-\mu)^T \Sigma^{-1}(x-\mu)\right] dx \\
&= \frac{1}{2} \left[\ln \frac{|L|}{|\Sigma|} + \int (x-m)^T L^{-1}(x-m) p(x) - \int (x-\mu)^T \Sigma^{-1}(x-\mu) p(x)\right] \\
&= \frac{1}{2} \left[\ln \frac{|L|}{|\Sigma|} + \mathbb{E}\left[(x-m)^T L^{-1}(x-m)\right] - \mathbb{E}\left[(x-\mu)^T \Sigma^{-1}(x-\mu)\right]\right] \\
&= \frac{1}{2} \left[\ln \frac{|L|}{|\Sigma|} + tr(Ł^{-1}\Sigma) + (\mu-m)^T L^{-1}(\mu-m) - tr(\Sigma^{-1}\Sigma) - (\mu-\mu)^T \Sigma\right. \\
&= \frac{1}{2} \left[\ln \frac{|L|}{|\Sigma|} + tr(Ł^{-1}\Sigma) + (\mu-m)^T L^{-1}(\mu-m) - tr(I_d)\right] \\
&= \frac{1}{2} \left[\ln \frac{|L|}{|\Sigma|} + tr(Ł^{-1}\Sigma) + (\mu-m)^T L^{-1}(\mu-m) - D\right]
\end{aligned}
$$

Now from step 5 to step 6 remember that the expectation is with respect to the distribution $p(x)$ which has mean μ and variance Σ. We use the fact that $\mathbb{E}\left[x^T A x\right] = c^T A c + tr(A\Sigma)$ where c is the mean of the distribution and Σ is the variance.

Problem 14

Problem 15 Entropy is defined as:

$$
\begin{aligned}
H[x] &= -\int p(x)\ln p(x)dx \\
&= -\int p(x)\ln\left[\frac{1}{((2\pi)^D|\Sigma|)^2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right]dx \\
&= -\int p(x)\left[\ln\frac{1}{((2\pi)^D|\Sigma|)^2}-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]dx \\
&= \frac{1}{2}\int p(x)\left[\ln(2\pi)^D|\Sigma|+(x-\mu)^T\Sigma^{-1}(x-\mu)\right]dx \\
&= \frac{1}{2}\ln(2\pi)^D|\Sigma|\int p(x)+\frac{1}{2}\int(x-\mu)^T\Sigma^{-1}(x-\mu)p(x)dx \\
&= \frac{D}{2}\ln 2\pi+\frac{1}{2}\ln|\Sigma|+\frac{1}{2}\mathbb{E}\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \\
&= \frac{1}{2}\ln|\Sigma|+\frac{D}{2}\ln 2\pi+\frac{1}{2}(\mu-\mu)^T\Sigma^{-1}(\mu-\mu)+\frac{1}{2}tr(\Sigma^{-1}\Sigma) \\
&= \frac{1}{2}\ln|\Sigma|+\frac{D}{2}\ln 2\pi+\frac{1}{2}tr(I_D) \\
&= \frac{1}{2}\ln|\Sigma|+\frac{D}{2}\ln 2\pi+\frac{D}{2} \\
&= \frac{1}{2}\ln|\Sigma|+\frac{D}{2}(1+\ln 2\pi)
\end{aligned}
$$

Problem 17

We know that any square matrix can be written in the form of a summation between a symmetric matrix and an antisymmetric matrix. To see this notice that for a square matrix A we can decompose it as:

$$A=\frac{1}{2}(A-A^T)+\frac{1}{2}(A+A^T)$$

Also notice that $(A - A^T)^T = A^T - A = -(A - A^T)$ and $(A + A^T)^T = (A^T + A) = (A + A^T)$.

With this in mind all we have to do is write the precision matrix Λ in terms of these two components.

$$\Lambda = \frac{1}{2}(\Lambda - \Lambda^T) + \frac{1}{2}(\Lambda + \Lambda^T)$$
$$\Lambda_{ij} = \frac{1}{2}(\Lambda_{ij} - \Lambda_{ji}) + \frac{1}{2}(\Lambda_{ij} + \Lambda_{ji})$$

Now let's examine the exponent of the Gaussian in summation form:

$$= \frac{1}{2}\sum_i \sum_j (x_i - \mu_i)(\Lambda^S_{ij} + \Lambda^A_{ij})(x_j - \mu_j)$$

If we look at all the terms involving the antisymmetric portion Λ^A_{ij} we see that they are:

$$x_i x_j \Lambda^A_{ij} - x_j \mu_i \Lambda^A_{ij} - x_i \mu_j \Lambda^A_{ij} + \mu_i \mu_j \Lambda^A_{ij}$$

But as we sum over i and j we also get:

$$x_j x_i \Lambda^A_{ji} - x_i \mu_j \Lambda^A_{ji} - x_j \mu_i \Lambda^A_{ji} + \mu_j \mu_i \Lambda^A_{ji}$$

But since $\Lambda^A_{ij} = -\Lambda^A_{ji}$ The first equation can be written as:

$$-x_i x_j \Lambda^A_{ji} + x_j \mu_i \Lambda^A_{ji} + x_i \mu_j \Lambda^A_{ji} - \mu_i \mu_j \Lambda^A_{ji}$$

Which cancels the second equation. Therefore the sum of the antisymmetric form disappears in the summation.

Problem 18

To prove that the eigenvalues of a real symmetric matrix are real begin with the definition of eigenvalue eigenvector pair:

$$\begin{aligned} Ax &= \lambda x \\ \overline{Ax} &= \overline{\lambda x} \\ A\overline{x} &= \overline{\lambda}\overline{x} && A \in \mathbb{R} \\ \overline{x}^T A^T &= \overline{x}^T \overline{\lambda} \\ \overline{x}^T A &= \overline{x}^T \overline{\lambda} && A = A^T \end{aligned}$$

Now just multiply the last equation on the right by x and the first equation on the left by $\overline{x}^T$ Then we get two equations:

$$\begin{aligned} \overline{x}^T Ax &= \overline{x}^T \lambda x \\ \overline{x}^T Ax &= \overline{x}^T \overline{\lambda} x \end{aligned}$$

The left hand side of both equations is identical so we can set them equal and we see:

$$\overline{x}^T \lambda x = \overline{x}^T \overline{\lambda} x$$

Which implies that $\lambda = \overline{\lambda}$. This means that λ must be real and we are done.

Now we need to show that the eigenvectors are orthogonal when $\lambda_1 \neq \lambda_2$. To do this we just use the eigenvalue equation for these two difference eigenvalues.

$$\lambda_1 x^T y = (Ax)^T y = x^T A^T y = x^T A y = \lambda_2 x^T y$$

This means that $\lambda_1 x^T y = \lambda_2 x^T y$ but $\lambda_1 \neq \lambda_2$ therefore $x^T y = 0$ which means that they are orthogonal.

The last piece is to show that this is in general true even when the two eigenvalues are the same. In the case of repeated eigenvalues I have 1 of two cases. Either they are repeated but non zero, or there are zero eigenvalues. If there are zero eigenvalues then the matrix is singular and $Ax = \lambda x = 0x = 0$ is true. This means that x is in the null space of A and is thus perpendicular to the column space. All of the eigenvectors with non zero eigenvalues will be in the column space of A so this new x is perpendicular to them . In the other case when there are repeated eigenvalues we know that the eigenvalues span the column space of the matrix. Let's just assume that the matrix is full rank. This means that the eigenvectors need to form a basis for the column space of the matrix A. With d repeated eigenvalues we have a d dimensional plane through which we can draw our eigenvectors choose them to be orthogonal.

Problem 19

Problem 21

The matrix is symmetric so we only need to count all elements above and on the diagonal of the matrix. Notice that each off diagonal of the matrix has one less element than the preceding diagonal. So we can create a series of the form:

$$\sum_{n=1}^{D} n = \frac{D(D+1)}{2}$$

Problem 26

This is just simple algebra:

$$I =$$

Problem 22

Here we prove that the inverse of a symmetric matrix is itself symmetric. Take B to be the inverse of A therefore $AB = BA = I$, and since A is symmetric we know that $A^T = A$.

$$\begin{aligned} AB &= I \\ B^T A^T &= I^T \\ B^T A^T &= BA \\ B^T A^T B &= BAB \\ B^T AB &= BAB \\ B^T I &= BI \\ B^T &= B \end{aligned}$$